Proceedings of

# The 2012 UKACC International Conference on Control

Cardiff, UK

3-5 September 2012

**The 2012 UKACC International Conference on Control**

**IEEE Catalog Number: CFP1270S-ART**
**ISBN: 978-1-4673-1560-9**

Additional copies of this publication are available from:
Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: (845)758-0400
Fax: (845)758-2633
E-mail: curran@proceedings.com

# The 2012 UKACC International Conference on Control (UKACC Control 2012)

Cardiff, UK, 3-5 September 2012

## GREETINGS FROM THE CONFERENCE CHAIRS

On behalf of the general conference committee, we cordially welcome you to attend the 2012 UKACC (United Kingdom Automatic Control Council) International Conference on Control (UKACC Control 2012). We hope you will enjoy your time in Cardiff in the UK and find them stimulating and fruitful.

UKACC Control 2012 is the ninth in this now well established series of prestigious biennial events. It will take place in Cardiff, UK during 3-5 September 2012 and is organised by the Advanced Control and Network Technology Research Unit in the University of Glamorgan. UKACC Control 2012 aims to provide a broad international forum for scientists, engineers, and educators working in the areas of control systems to discuss state-of-the-art science and technologies in the general area of control systems with particular emphasis being placed on applications.

The technologies in automatic control have been playing important roles in our modern civilization and expected to stay as main characters in the future. The conference intends to provide a platform for researchers, engineers, academicians as well as industrial professionals from all over the world to present their research results and development activities in automatic control. The conference will promote the development of automatic control, strengthen the international academic cooperation and communications, and provide an opportunity of exchanging research ideas.

With the tremendous work of the members of the International Program Committee, the final program has a unique set of technical sessions covering latest research and development results in automatic control. It contains a number of high level theoretical papers as well as many interesting application papers. A total of 254 papers were submitted to UKACC Control 2012 from different countries and regions. After being reviewed, 191 papers have been accepted in the final program. There are 3 plenary talks, 36 lecture sessions and 3 poster sessions. The papers were assigned with the purpose of forming coherent sessions.

We have invited an excellent set of plenary speakers. On the first day, after the official opening, there will be a plenary talk on "Control for Performance Optimization" presented by Professor Sebastian Engell in Technische Universität Dortmund, Germany. On the second day, we will have a plenary talk on "Control Systems Engineering: a Business-Driven Approach" jointly given by Professor Peter Fleming in the University of Sheffield, UK and Mr Stephen Hill in Rolls Royce, UK. The third day will start with the plenary talk on "Aircraft Tonal Active Noise Control Techniques and Their Re-use delivered by Mr Ian Stothers in Ultra Electronics Controls, UK.

In addition to the high quality of the academic contributions in the program, UKACC Control 2012 has received various supports from UK industry with sponsorship from commercial organizations and an involvement from the four professional organizations, namely, the Institution of Engineering and Technology (IET), the Institution of Mechanical Engineers (IMechE), the Institute of Measurement and Control (InstMC), and the Institute of Electrical and Electronics Engineers (IEEE). The strong industrial involvement is evident in the program for the three Mini-Symposia.

We would like to take this opportunity to thank the members of the organising committee and the international program committee for their fantastic work. Our sincere thanks also go to many volunteers and reviewers for their diligence and dedication in the preparation of this conference.

Guoping Liu (General Chair)

Duc pham (IPC Chair)

David Rees (Organising Chair)

# The 2012 UKACC International Conference on Control (UKACC Control 2012)

Cardiff, UK, 3-5 September 2012

## ORGANISING INSTITUTIONS

### Organiser and Sponsor

University of Glamorgan, UK

### Technical Co-sponsors

The United Kingdom Automatic Control Council, UK<br>
The Institution of Engineering and Technology, UK<br>
The Institution of Mechanical Engineers, UK<br>
The Institute of Measurement and Control, UK<br>
The Institute of Electrical and Electronics Engineers, USA

# The 2012 UKACC International Conference on Control (UKACC Control 2012)

Cardiff, UK, 3-5 September 2012

## GENERAL CONFERENCE COMMITTEE

**General Chair**

| | |
|---|---|
| Guoping Liu | University of Glamorgan |

**Program Chair**

| | |
|---|---|
| Duc Truong Pham | University of Birmingham |

**Program Co-chairs**

| | |
|---|---|
| Shuanghua Yang | Loughborough University |
| Hongnian Yu | Staffordshire University |

**Organising Chair**

| | |
|---|---|
| David Rees | University of Glamorgan |

**Invited session Chair**

| | |
|---|---|
| James Wildborne | Cranfield University |

**Publication Chair**

| | |
|---|---|
| Zidong Wang | Brunel University |

**Exhibition Chair**

| | |
|---|---|
| Roger Dixon | Loughborough University |

**Finance Chair**

| | |
|---|---|
| Hongji Yang | De Montfort University |

**Industrial Advisory Panel Chair**

| | |
|---|---|
| Stephen Daley | University of Southampton |

**Pre-Conference Workshop Chair**

| | |
|---|---|
| Jihong Wang | University of Warwick |

**UKACC Chair**

| | |
|---|---|
| Andrew Plummer | University of Bath |

# The 2012 UKACC International Conference on Control (UKACC Control 2012)

Cardiff, UK, 3-5 September 2012

## ORGANISING COMMITTEE

Jason Guo, University of Glamorgan
David Rees, University of Glamorgan (chair)
Kary Thanapalan, University of Glamorgan
Jonathan Williams, University of Glamorgan
Fan Zhang, University of Glamorgan

## INDUSTRIAL ADVISORY PANEL

Graham Andrews, Tata Steel
Steve Daley, University of Southampton (chair)
Ian Griffin, Rolls Royce
Ben Jeppesen, Instron
John Pearson, BAE Systems
Peter Scotson, TRW Conekt
Richard Stephens, GE Energy Power Conversion
Ian Stothers, Ultra Controls
Dave Vorley, MBDA-Systems

# The 2012 UKACC International Conference on Control (UKACC Control 2012)

Cardiff, UK, 3-5 September 2012

## INTERNATIONAL PROGRAM COMMITTEE

| | |
|---|---|
| Alessandro Astolfi | Imperial College London, UK |
| Matjaz Gams, | Jozef Stefan Institute, Slovenia |
| Christian Bohn | Clausthal University of Technology, DE |
| Keith Burnham | Coventry University, UK |
| Yi Cao | Cranfield University, UK |
| Fabio Celani | Sapienza University of Rome, Italy |
| Sheng Chen, | University of Southampton, UK |
| Wen-Hua Chen | Loughborough University, UK |
| Kai Cheng | Brunel University, UK |
| Jyh-Horng Chou | National Kaohsiung University of Applied Sciences, TW |
| Tim Clarke | University of York, UK |
| Steve Daley | University of Southampton, UK |
| Stephen Ding | University of Duisburg-Essen, DE |
| Zhengtao Ding | University of Manchester, UK |
| Roger Dixon | Loughborough University, UK |
| Roger Goodall | Loughborough University, UK |
| David Goodall | Coventry University, UK |
| John Gray | University of Manchester, UK |
| Dongbing Gu | University of Essex, UK |
| Lei Guo | Beihang University, PRC |
| Qing-Long Han | Central Queensland University, Australia |
| Chris Hann | University of Canterbury, NZ |
| Gerald Hearns | Converteam, US |
| Daniel Ho | City University of Hong Kong, HK |
| Davor Hrovat | Ford, US |
| Rolf Isermann | Technische Universität Darmstadt, DE |
| Wang Jin | Liverpool John Moores University, UK |
| Rolf Johansson | Lunds Univeristy, SE |
| Peter Jones | University of Warwick, UK |

| | |
|---|---|
| Dewi Jones | Gwefr Cyf, UK |
| Chandra Kambhampati | University of Hull, UK |
| Nick Karcanias | City University London, UK |
| Laszlo Keviczky | Hungarian Academy of Sciences, HU |
| Jozef Korbicz | Institute of Control & Computation engineering, PL |
| Ali Koshkouei | Conventry University, UK |
| James Lam | University of Hong Kong, HK |
| Ziqiang Lang | University of Sheffield, UK |
| Nicolas Langlois | ESIGELEC, FR |
| Paul Lewin | University of Southampton, UK |
| Guoping Liu | University of Glamorgan, UK |
| Lu Liu | University of Nottingham, UK |
| Yang Liu | University of Aberdeen, UK |
| Paul McKenna | Glasgow Caledonian University, UK |
| Xiangdong Ma | Lancaster University, UK |
| Tian Xiang Mei | University of Salford, UK |
| Geyong Min | University of Bradford, UK |
| Nazim Mir-Nasiri | Swinburne University of Technology, Malaysia |
| Brett Ninness | University of Newcastle, AU |
| Dmitry Novikov | Institute of Control Sciences, RU |
| Liatsis Panos | City University London, UK |
| Ron Patton | University of Hull, UK |
| Dobrila Petrovic | Conventry University, UK |
| Duc Truong Pham | University of Birmingham, UK |
| Andrew Plummer | University of Bath, UK |
| Radu-Emil Precup | |
| Giuliano C Premier | University of Glamorgan, UK |
| David Rees | University of Glamorgan, UK |
| Eric Rogers | University of Southampton, UK |
| Anthony Rossiter | University of Sheffield, UK |
| Johan Schoukens | Vrije universiteit Brussel, BE |
| Henry Selvaraj | University of Nevada. US |
| Jinhua She | Tokyo University of Technology, Japan |
| Silvio Simani | |
| Sarah Spurgeon | University of Kent, UK |
| Richard Stephens | Converteam, US |
| Allen Stubberud | University of California, US |
| Chun-Yi Su | Concordia University, CA |

| | |
|---|---|
| Bob Sutton | Plymouth University, UK |
| Ai Hui Tan | Multimedia University, MY |
| Min Tan | Chinese Academy of Sciences, PRC |
| James Taylor | Lancaster Universtiy, UK |
| Nina Thornhill | Imperial College London, UK |
| Ajit Verma | Indian Institute of Technology, Bombay, India |
| Hong Wang | University of Manchester, UK |
| Jihong Wang | University of Warwick, UK |
| Zidong Wang | Brunel University, UK |
| Kevin Warwick | University of Reading, UK |
| James Whidborne | Cranfield University, UK |
| Bob Williams | |
| Min Wu | Central South University, PRC |
| Yugeng Xi | Shanghai Jiao Tong University, PRC |
| Xun Xu | University of Auckland, NZ |
| Ling Xu | University of Manchester, UK |
| Taicheng Yang | University of Sussex, UK |
| Shuang-Hua Yang | Loughborough University, UK |
| James Yates | AstraZeneca Global |
| Peter Young | Lancaster Universtiy, UK |
| Dingli Yu | Liverpool John Moores University, UK |
| Hongnian Yu | Staffordshire University, UK |
| Yi Yue | Shanghai Jiao Tong University, PRC |
| Jie Zhang | Newcastle University, UK |
| Jun Zhao | Northeastern University, PRC |
| Donghua Zhou | Tsinghua University, PRC |
| Zi-Qiang Zhu | University of Sheffield, UK |

## Present and Past of UKACC Control

**Control 2012**

University of Glamorgan, 3-5 September 2012

**Control 2010**

Coventry University, 7-10 September 2010

**Control 2008**

Manchester Conference Centre, 1-4 September 2010

**Control 2006**

Universities of Glasgow and Strathclyde, 30 Aug-1 Sep 2006

**Control 2004**

University of Bath, 6-9 September 2004

**Control 2002**

University of Sheffield, 10-12 September 2002

**Control 2000**

University of Cambridge, 4-7 September 2000

**Control 1998**

Swansea University, September 1998

**Control 1996**

Exeter University, September 1996

# The 2012 UKACC International Conference on Control (UKACC Control 2012)

Cardiff, UK, 3-5 September 2012

## TABLE OF CONTENTS

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Tensegrity-Based Formation Control of Unmanned Vehicles

Sook Yen Lau* and Wasif Naeem†
School of Electronics, Electrical Engineering and Computer Science
Queen's University Belfast, University Road, Belfast BT7 1NN, UK.
Email: slau02@qub.ac.uk*, w.naeem@qub.ac.uk†

***Abstract*—A new formation control methodology modelled by a virtual tendon-driven system using the tensegrity structures is presented. The objective of the work is to regulate the formation of unmanned vehicles within the communications bandwidth and perform point-to-point manoeuvring tasks. The reaction control forces that are experienced by vehicles in the formation are determined by the admissible tendon forces in tensegrity. A control law is designed to stabilize the interspacing between the vehicles in the presence of disturbances by making the combined use of string and spring characteristics. Simulation results demonstrate the effectiveness of the proposed approach in terms of maintaining the formation and avoiding inter-vehicle collisions. Formation shape changing is also performed by varying the relative parameters between the vehicles.**

## I. Introduction

Formation control is a matter of controlling the relative position and orientation of an ensemble of autonomous vehicles while allowing the group to move as a whole in a stable configuration. It is proven that moving a group of vehicles is more beneficial than a single vehicle in the presence of uncertain and adverse environments. This is because vehicles in formation are able to acquire enough and accurate information from the environment, whilst enhancing their power to withstand danger. Hence, the use of groups of multiple autonomous vehicles to perform co-ordinated and co-operative tasks has been attracting a growing amount of attention recently in the area of autonomous robotics and control communities.

It is also well known that formation systems have many advantages such as wide area sensing coverage and system energy conservation due to the reduction of friction in each vehicle. Robustness and efficiency of the system also increases while offering redundancy, reconfiguration ability and structural flexibility for the system. These advantages have led formation control perform a wide variety of functions in land, marine and aerial applications. Specific applications include exploration, search and rescue, microsatellite clusters and transportation of large and heavy objects [1].

Problems to set up in co-operative formation control often involve achieving formation, maintaining formation and dynamic switching between different formation shapes whilst carrying out a task. A number of methodologies have been developed to address the shape dynamics of a group of vehicles in a plane such as implicit polynomial (IP) [2], elliptic Fourier descriptors (EFD) [2], sliding mode controllers [3], extended Kalman filter with an input-state feedback control law [4], Lyapunov function [5], to name a few.

In this paper, formation dynamics of a group of vehicles is synthesised and analysed using a virtual tendon-driven tensegrity system. Tensegrity structures are spatial structural systems with interconnected strings and bars/struts [6], where strings are the tendon members and bars/struts are the compressive members. In literature, tensegrity structures have been mainly used to design mechanical structures and little attention has been paid to their usability to develop formation control algorithms for dynamic systems. In [7], for instance, the authors have demonstrated formation for a group of vehicles using an energy-momentum method in tensegrity models. However, they assumed zero external (disturbance) forces which simplified the problem.

The new control law here is formulated to drive a dynamic group of vehicles into a specified formation with control forces that are represented by admissible tendon forces in tensegrity structures. This control law is designed to prevent strings slacking and yielding and make the structure responsive to the environment disturbance forces. In control terms, this force precludes any two vehicles coming too close to each other (string slacking) in order to avoid collision. The controller also prevents the vehicles moving too far apart from each other (string yielding) in order to keep the vehicles within the communications range. The proposed overall formation system is characterised by three main considerations: vehicles formation geometry which is modelled using the concept of tensegrity, communications topology which is represented by strings and bars of the tensegrity structure, and the interaction control algorithm.

In the remainder of this paper, Section II outlines the benefits of tensegrity properties. Formation shape and its communication topology will be described in Section III whilst tendon controller design and formation system regulation will be explained in Section IV and V respectively. Simulation results are shown in Section VI to demonstrate formation achieving, formation maintaining, formation changing and manoeuvring tasks. Concluding remarks are made in Section VII.

## II. Tensegrity Structures

The artist Kenneth Snelson built the first tensegrity structure [8] and the term tensegrity was coined by Fuller as a

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

contraction of tensional integrity [9]. In nature, bones and tendons of animals and humans are connected in a way that allows for smooth control movements, where the bones provide compressive load-carrying capacity while tendons provide tensions stabilizing the configuration [6]. In structural engineering, controlled tensegrity make buildings responsive to natural environmental disturbances such as earthquakes and winds [6].

In the tensegrity framework, the stability and rigidity of tensegrity structures have been proven using the model of energy function [10]. This has motivated the development of mathematical machinery in the design and analysis of static and dynamic tensegrity systems to achieve shape formation control and other engineering functions. It is possible to form a tensegrity structure by using models of springs, strings and bars which can be combined to provide greater flexibility and control of the whole unit. The extreme shape-changing ability of tensegrity structures has been proven in the approach of a robotic tensegrity worm crawls (which requires stiffness control) while squeezing through crevices requiring large shape changes [11].

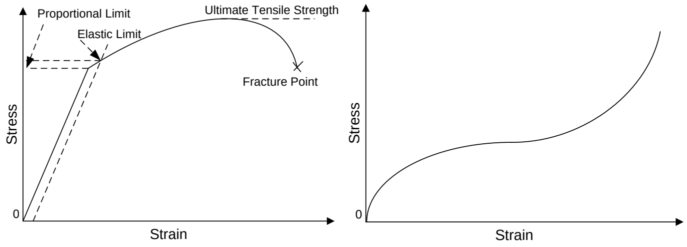
In a dynamic tensegrity-based formation control, one way of controlling the position of each vehicle in the group is by varying the length of the bars. However, for simplicity, the formation controller is designed here based on the relative distance and orientation between the leader and follower vehicles. For example, if the requirement is to expand the formation, the bar lengths could be increased. The same result can also be achieved by varying the interspacing between a designated pair of vehicles by controlling the admissible tendon forces in a tensegrity structure. This tendon controller is designed by assuming a spring with elastic characteristic (non-Hookean) which experiences the properties of both spring and string.

### A. Elastic Spring as Tendon

A linear spring can store energy either by applied compressive force or tensile force on it, however, a string (elastic band) can only store energy in the presence of a tensile force. In a tensegrity system, the tendon/string is always kept in tension, therefore it behaves in a similar manner to the spring. Hooke's Law states that the force that is exerted on a spring is proportional to the extension of the spring. Note that the stress and strain are (uniquely) related to force and extension of the spring, hence, the gradient of a stress - strain graph as shown in Figure 1a is identical to the force versus extension graph. However, a string with elastic characteristics, has a non-linear gradient graph as shown in Figure 1b. This property of the elastic spring is exploited in this paper and is further elaborated in Section IV on formation controller modelling.

## III. Problem Formulation

An example of a 4-vehicle formation connected using the concept of tensegrity is shown in Figure 2. Here, nodes are defined as vehicles and edges are represented by strings/elastic springs(s) and bars(b). The edges correspond to communication topology and control force directions between the vehicles. A chain communication topology is assumed where each vehicle moves according to its neighbour/leader in the formation which may not be its nearest. As shown in Figure 2, $UV_1$ is the leader of $UV_2$ while $UV_2$ act as the leader for $UV_3$ and so on. $\mathbf{f}_{LF}$ represents the force that is exerted on follower ($F$) according to its leader ($L$). The magnitude of this control force is dependant on stress, $\omega$, which is a tensegrity parameter. The edge is a string if $\omega > 0$, and is a bar if $\omega < 0$ [10]. In addition, a centralised control architecture is considered where a virtual leader ($UV_1$) is assigned for leading the whole formation and combined with the chain topology. Hence, the force, $\mathbf{f}_{41}$ and its corresponding communications link in Figure 2 can be ignored.



(a) Spring

(b) Elastomer

Figure 1: The gradient of stress strain curves

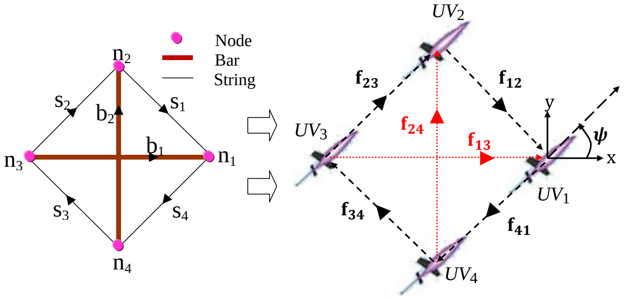


Figure 2: Vehicles communication topology referred to a C2T4 tensegrity system

It has been proven that shape changing task will be simpler to achieve by relaxing the equilibrium of the structure rather than pressing against a fixed equilibrium [6]. The tensegrity control parameter, $\omega$ will affect the edge's tension and cause another new equilibrium interspacing between the nodes. In formation control, the equilibrium interspacing between the vehicles is dependant on $\omega$ and can be represented by a virtual tendon-driven system (spring-mass system) as shown in Figure 3.

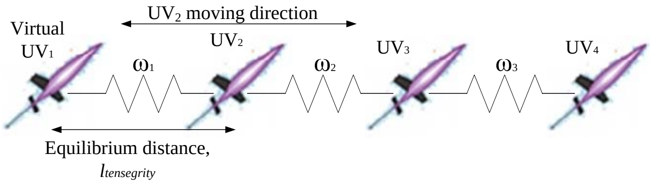


Figure 3: Formation in a virtual tendon system

The force and position of each vehicle in the tendon-driven system can be mathematically defined as

$$m\ddot{\mathbf{q}}+b\dot{\mathbf{q}}+\omega\mathbf{q}+\mathbf{f}(t)=0 \tag{1}$$

where $m$ is the mass of the vehicle, $b$ is the friction coefficient and $\mathbf{q}$ is the position vector of the vehicle defining its local coordinates. $\omega$ is the sum of the two stresses $(\omega_1+\omega_2)$ for vehicle $UV_2$; $(\omega_2+\omega_3)$ for vehicle $UV_3$ and $(\omega_3)$ for vehicle $UV_4$. $\mathbf{f}(t)$ represents the applied force to the tendon system to maintain the interspacing between the vehicles. In order to compensate for uncertain environments, the plant is rewritten as

$$m\ddot{\mathbf{q}}+b\dot{\mathbf{q}}+\omega\mathbf{q}+\mathbf{f}(t)+\mathbf{d}_f=0 \tag{2}$$

where $\mathbf{d}_f$ is an external disturbance force (e.g. wind for aircraft or current for ship). Note that the benefit of this parallel connected formation is that it is easy to keep track of every vehicle in the group. However, a major disadvantage is the error propagation from one pair of vehicles to the other.

## IV. Tendon Controller Modelling

In the formation control framework, there will be $r=n-1$ communication links in a formation containing $n$ vehicles. A separate controller is designed to control each pair of vehicles that are connected by a direct communication link. The signal measured by the controller is the relative distance between leader and follower vehicles and is denoted as $l_{LF}$. The controller output, $\mathbf{f}_{LF}$ is the force function which is modelled by an elastic spring. This tendon force, $\mathbf{f}_{LF}$ that is exerted on the follower vehicle with respect to its leader was designed to have a much larger elastic limit compared to its proportionality limit and is defined mathematically in Equation 3. This equation is essentially a combination of stress-strain relationship of spring and string as illustrated in Figure 1.

$$\mathbf{f}_{LF}=\begin{cases} K\ln\frac{l_{LF}}{l_{tensegrity}}(\mathbf{q}_F-\mathbf{q}_L) & \text{if } 0<l_{LF}\le l_{ultimate} \\ K\exp(-\frac{l_{LF}-l_{tensegrity}}{l_{break}})(\mathbf{q}_F-\mathbf{q}_L) & \text{if } l_{ultimate}<l_{LF}\le l_{break} \\ 0 & \text{if } l_{LF}>l_{break} \end{cases} \tag{3}$$

$$K=K_1\alpha_{LF}\omega_{LF} \tag{4}$$

Where $K_1$ is a gain parameter that is proportional to the disturbance force, $\mathbf{d}_f$ and helps to adapt the controller to external disturbances.

$$K_1\propto\mathbf{d}_f \tag{5}$$

$\omega_{LF}$ represents the stress and $\alpha_{LF}$ is a signed scalar parameter that determines the attracting ($\alpha_{LF}>0$) or repelling ($\alpha_{LF}<0$) force that is exerted on the follower vehicle. This is further elaborated in Section IV A.

The parameter, $l_{tensegrity}$ is the desired distance between a given pair of vehicles. $l_{ultimate}$ is the maximum distance between the controlled pair of vehicles, where ultimate tensile strength (attracting or positive force) that is applied to the follower vehicle increases. After this point, this attracting force starts to reduce. This is done in order to reduce the rebound force that will occur on the follower if the disturbance is suddenly removed.

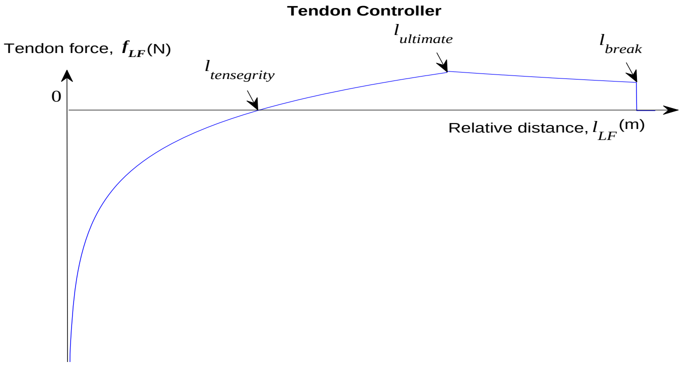


Figure 4: Design of control force with response to displacement

$l_{break}$ is the maximum length of the string; the string is fractured at this point if the disturbance force continues to be added to the string. In formation control terms, $l_{break}$ is the maximum communication length between the vehicles. The control force, $\mathbf{f}_{LF}$, is equal to zero at this point to give up a straying vehicle rather than trying to apply more force on it to pull it back to the formation. This vehicle might collide with the other vehicles in the formation when the disturbance force is suddenly removed due to the large restoring force.

Figure 5 is drawn to show the block diagram representation of the overall tensegrity-based nonlinear tendon-driven control for a pair of vehicles. To ensure zero steady-state tracking error, an integral term is added to the closed loop system.

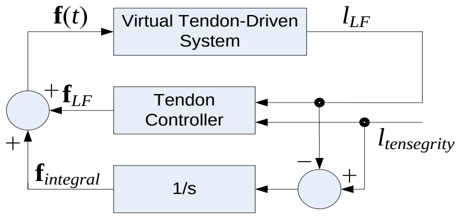


Figure 5: Block diagram of tendon-driven system with integral action

### A. Analysis of Tendon Force

Tension along the spacing edge between any two vehicles is changed according to the edge's extension and can be divided into two categories: attracting force (positive force) and repelling force (negative force). By considering the two vehicles within the $l_{ultimate}$ communication range, the tendon force can be mathematically defined as in Equation 6 with the parameter $K_1$ set to unity for simplicity.

$$\mathbf{f}_{LF}=\alpha_{LF}\omega_{LF}[\ln l_{LF}-\ln l_{tensegrity}](\mathbf{q}_F-\mathbf{q}_L) \tag{6}$$

The equilibrium of the tendon-driven system can be found by equating the sum of all the forces that are acting on vehicles in the system to zero i.e.

$$\sum_{r=1}^{n-1}\alpha_{L_rF_r}\omega_{L_rF_r}[\ln l_{L_rF_r}-\ln l_{tensegrity_r}](\mathbf{q}_{F_r}-\mathbf{q}_{L_r})=0 \tag{7}$$

The subscript $r$ refers to a particular communication link that connected a pair of vehicles and $n$ is the total number of vehicles in the formation. Let the position vector $\mathbf{q}^e = (x^e, y^e)$ represents the equilibrium tensegrity structure. Hence, the equilibrium stress constant of a pair of vehicles can be defined as

$$\widetilde{\omega}_{LF}(x^e, y^e) = \alpha_{LF}\,\omega_{LF}[\ln l_{LF} - \ln l_{tensegrity}] \tag{8}$$

The edge here is modelled as a string, hence, $\omega_{LF} > 0$. Also, by assuming the tendon-driven system in equilibrium, $\widetilde{\omega}_{LF}(x^e, y^e) = \omega_{LF}$, therefore equation 8 can be simplified as

$$1 = \alpha_{LF}[\ln l_{LF} - \ln l_{tensegrity}] \tag{9}$$

Equation 9 proves that the tendon-driven system experiences an attracting force ($\alpha_{LF} > 0$) when the relative distance ($l_{LF}$) between the vehicles is larger than its equilibrium distance ($l_{tensegrity}$); and repelling force ($\alpha_{LF} < 0$) when the relative distance ($l_{LF}$) between the vehicles is smaller than its equilibrium distance ($l_{tensegrity}$).

## V. Regulation of Formation Subsystem

The objective in this section is to regulate the interspacing distance between a pair of leader and follower vehicles. The equilibrium distance, $l_{tensegrity}$ and the maximum communication range, $l_{break}$ are the key control parameters which are assumed to be 16m and 32m, respectively for simulations purposes. The two vehicles formation was simulated in a tendon-driven system using the proposed tendon controller and compared with a linear PI controller. All the parameters in tendon-driven system shown in Equation 2 are considered to be unity. And the parameters of PI controller were optimized using the Ziegler-Nichols (ZN) tuning method in order to give a fast and smooth response. The values estimated for $K_P$ and $K_I$ were 0.4091 and 0.1002, respectively.

Figure 6 shows that the tendon control system with a desired step response of 16m was achieved in a settling time of 50.04s with an overshoot of 1.35%, while the linear controller has a slower settling time of 60.77s and experienced a higher overshoot of 3.56%. The dip at the beginning of step response is due to the higher repelling force that was produced by the linear controller as shown in Figure 7. This repelling force was encountered because the initial spacing between the vehicles was kept very small. However, the tendon controller was able to eliminate this effect and provided a smooth and quick response to the system with a lower repelling force.

Next external disturbance forces, $\mathbf{d}_f$ of magnitude 30N and 110N were added to the subsystem to demonstrate controller performance. These impulse disturbances were triggered at 150s for a period of 150s to force change $l_{LF}$. Both positive and negative disturbances are considered demonstrating the dilation and squeezing of inter-vehicle spacing respectively.

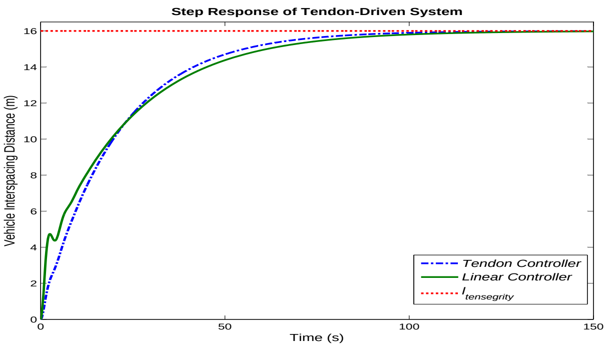


Figure 6: Closed-loop unit step response for formation subsystem

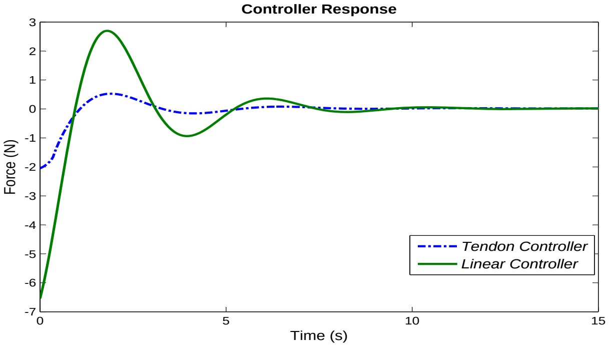


Figure 7: Subsystem controller response

Simulation results in Figure 8 depicts that an increasing positive disturbance force caused the follower and leader vehicles to further repel from each other. It can be observed that as long as the follower vehicle moves within the prescribed communications range, the formation can be maintained (Figure 8(a)). The follower vehicle will lose communication with its leader when the separation between the vehicles is more than $l_{break}$, 32m. This will cause the removal of control force in order to give up the straying follower (Figure 8(b)). However, the PI controller produced a large attracting control force to the follower vehicle when it moved out of the communication to pull it back to its equilibrium position. This had caused a collision between the two vehicles when the disturbance force is suddenly removed. This collision happened due to the rebound force being larger than the vehicle's repelling force.

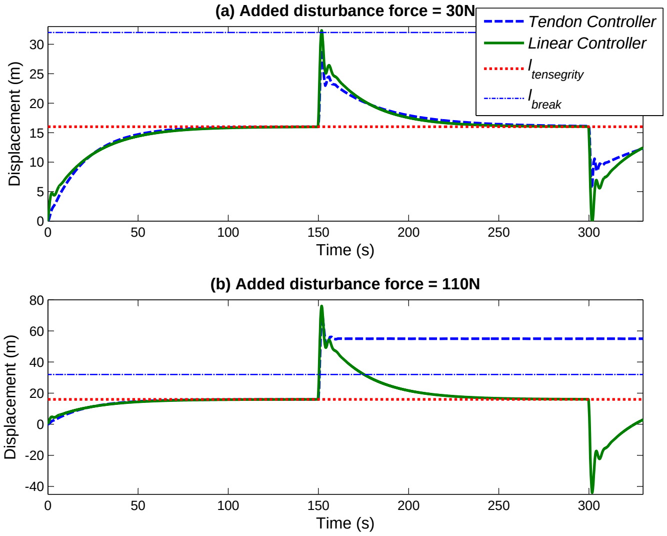


Figure 8: Step responses under positive disturbance force

Table I provides a quantitative comparison between step responses under the positive and negative disturbance forces. In contrast to the positive disturbance force, an increasing negative disturbance force caused the distance between the two vehicles become smaller. Note that the relative distance in tendon control system is never equal to or less than zero due to the force that is exerted on the follower vehicle tends to be infinitely large (see Equation 3). This helps to prevent the collision between the vehicles. However, the linear controller was unable to prevent the collision, shown by negative bold displacement in Table I.

TABLE I: Tendon and linear controller comparison

| Disturbance force $\mathbf{d}_f$, (N) | maximum and minimum vehicles interspacing (m) | | | |
|---|---|---|---|---|
| | $\mathbf{d}_f$ added at 150s | | $\mathbf{d}_f$ removed at 300s | |
| | Tendon controller | Linear controller | Tendon controller | Linear controller |
| 30 | 29.06 | 32.36 | 5.57 | 0.07 |
| 100 | **55.04** | 76.05 | **55.03** | **-43.91** |
| -30 | 5.37 | 0.08 | 28.95 | 32.36 |
| -100 | 1.04 | **-44.07** | **32.08** | 75.97 |

∗ The vehicles undesired spacings are highlighted in bold. Negative values show the vehicles's collision and the communication lost is represented by bold positive value.

## VI. Formation Maintenance and Manoeuvring

In this section, the overall formation system is described as a combination of tendon-driven systems outlined previously. The formation manoeuvring will be described and demonstrated through simulations.

### A. Formation control and dynamic shape changes

In order to adapt to the changes in the environment such as obstacles, the formation control method should be flexible so that shape changes can be conveniently carried out. Here, formation control is achieved by maintaining the distances between nominated pairs of vehicles using the concept of tendon-driven system. Note that each vehicle in the formation is assumed to be autonomous and that it makes decisions based on the relative distance and bearing among individuals. For example, $UV_2$ makes its decision to move according to the distance ($d_1$) and relative orientation ($\theta_{ro1}$) with respect to its leader, $UV_1$ as shown in Figure 9. The topology of closed-loop formation system composed of $r$ communication links is depicted in Figure 10.

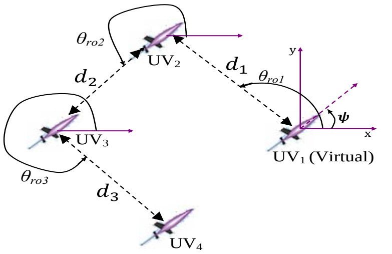


Figure 9: Relative parameters in formation

The synchronisation of two vehicles requires the consideration of the following equations:

$$\begin{aligned} x_{n+1} &= d_n \cos(\theta_{ron} + \psi) + x_n \\ y_{n+1} &= d_n \sin(\theta_{ron} + \psi) + y_n \end{aligned} \quad (10)$$

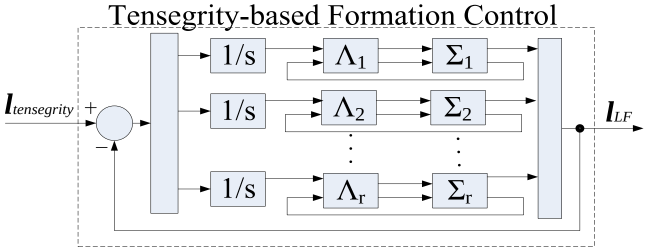


Figure 10: Block diagram of tensegrity-based formation closed-loop system. Tendon-driven systems and tendon controllers are denoted by Σ and Λ, respectively.

where ($x_{n+1}$, $y_{n+1}$) is the updated relative position of $UV_{n+1}$ and ($x_n$, $y_n$) is the current position of its leader, $UV_n$ in the global reference frame of coordinates. For formation shape change, vehicle's relative orientation ($\theta_{ron}$) and the equilibrium spacing ($d_n$=$l_{tensegrity}$) along the edges are set as reference parameters. By controlling these variables, the shape changing task can be achieved.

### B. Autopilot design

In order to design the controller, it is assumed that all the vehicles in the formation have the same dynamics which can be represented by a linear first order Nomoto model given by Equation 11.

$$T\ddot{\psi} + \dot{\psi} = K_n \delta \quad (11)$$

whose transfer function is:

$$\frac{\psi}{\delta}(s) = \frac{K_n}{s(1+Ts)} \quad (12)$$

Where $K_n$ is the gain and $T$ is the system time constant that can be uniquely determined from the input rudder angle ($\delta$) and the output heading angle ($\psi$). From [12], the values of $K_n$ and $T$ are chosen to be 0.049 and 17.78 for simulation purposes.

An autopilot must have the function of maintaining the trajectory of the vehicle by following the desired heading ($\psi$(t)). It must also have the function of performing the change of heading without excessive oscillations and in the minimum possible time. In order to accomplish this, a suitable PD heading controller was developed which was tuned heuristically using trial and error.

The completed formation control setup is shown in Figure 11, where **q** is the position vector of the unmanned vehicle. The task was to move the leader with heading angle, $\psi$ and velocity, $v$ from one way-point to the next.

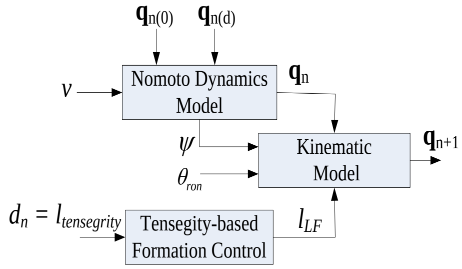


Figure 11: Block diagram of the formation control model

### C. Simulation results

The tensegrity-based formation control design method was simulated using a group of four UVs in a diamond shape formation, as in Figure 9, where $UV_1$ is the virtual vehicle. The objectives were to regulate the inter-UV distances within the prescribed communications range and to perform the formation changing and manoeuvring tasks. Commanded and actual trajectories of the formation along the $x-$ and $y-$axis are shown in Figure 12. The formation was required to move from initial way-point (WP0) to the final waypoint (WP4), through WP1, WP2 and WP3 by following the line of sight (LOS) between successive waypoints. Note that there is no crossover or collisions between any of the vehicle's paths. However, a longer route was taken during turning manoeuvres due to constant velocity assumption in order to maintain the shape of the formation.

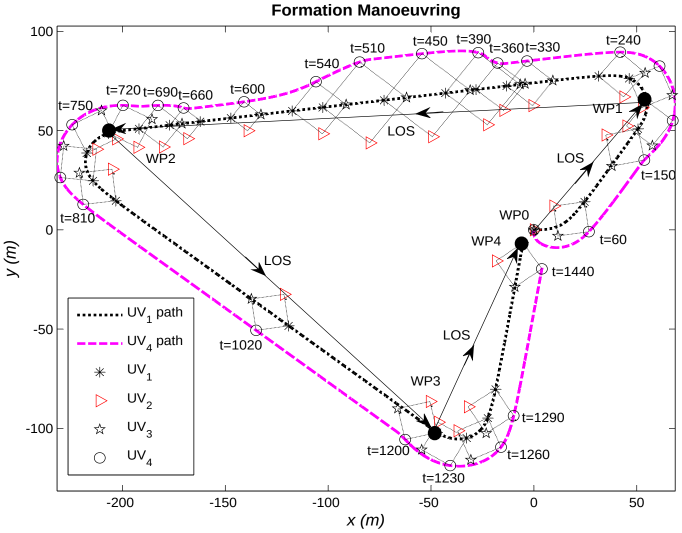


Figure 12: Simulations of formation changing and manoeuvring

The formation's size was changed twice at 350s and 500s for a period of 150s. In order to avoid any sudden changes to the formation and have to avoid inter-vehicle collisions, the reference input ($l_{tensegrity}$) in Figure 10 is taken to be a ramp signal with 0.5 slope. From the controller responses shown in Figure 13, it can be seen that the vehicles experience negative/repelling force when the shape dilation (parameter of $d_n$ or $l_{tensegrity}$ was changed from 16m to 30m) is performed. The subsequent positive force is expected as in Figure 7. Note that a time frame of 30s is needed for this shape dilation task. A positive/attracting force was applied at time 500s to contract the quadrilaterals shape from edge length of 30m to 10m. A longer time (40s) is required for this task due to the nonlinear controller characteristics. The formation was returned to its original shape ($d_r = 16$m) at 650s with a negative force.

## VII. Conclusions

A nonlinear control law that is modelled by the tendon-driven system using the concept of tensegrity structures is presented. It has been demonstrated that this control method can regulate the shape of formation in the presence of disturbances. The proposed approach is also scalable to any number of autonomous vehicles in the formation and is only limited by the communication bandwidth. It is also shown that the shape of the formation is conserved during manoeuvring. The shape changing algorithm can also be modified to include obstacle avoidance. Further research is required to develop more advanced mathematical machinery that can be used to analyse the stability of shape formation based on the tensegrity structures. The management of vehicle failures will also be considered in the future research.

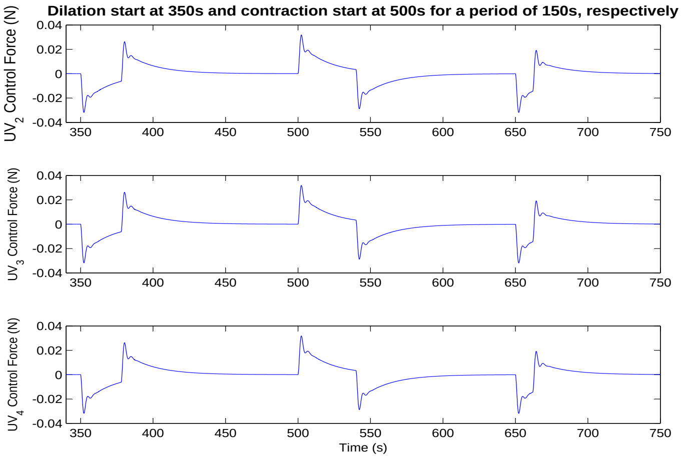


Figure 13: Simulations of input forces

## References

[1] Y. Chen and Z. Wang, "Formation control: a review and a new consideration," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, Edmonton, Canada, 2-6 August 2005, pp. 3181 – 3186.

[2] Y. Esin and M. Ünel, "Formation control of nonholonomic mobile robots using implicit polynomials and elliptic fourier descriptors," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 18, no. 5, pp. 765–780, 2010.

[3] M. Defoot, T. Floquet, A. Kökösy, and W. Perruquetti, "Sliding-mode formation control for cooperative autonomous mobile robots," *IEEE Transactions on Industrial Electronics*, vol. 55, no. 11, pp. 3944–3953, 2008.

[4] G. Mariottini, F. Morbidi, D. Prattichizzo, N. Valk, N. Michael, G. Pappas, and K. Daniilidis, "Vision-based localization for leader-follower formation control," *IEEE Transcation on Robotics*, vol. 25, no. 6, pp. 1431–1438, 2009.

[5] X. Li and J. Xiao, "Formation control in leader-follower motion using direct Lyapunov method," *International Journal of Intelligent Control and Systems*, vol. 10, no. 3, pp. 244–250, 2005.

[6] R. Skelton and M. Oliveira, *Tensegrity Systems*. New York: Springer, 2009, vol. XIV.

[7] B. Nabet and N. Leonard. (2009) Tensegrity models and shape control of vehicle formations. [Online]. Available: http://arxiv.org/pdf/0902.3710v1.pdf [Date accessed: 21 March 2012].

[8] K. Snelson, "Continuous tension, discontinuous compression structures," *U.S. Patent 3, 169, 611*, 1965.

[9] R. Fuller, "Tensile-integrity structures," *U.S. Patent 3, 063, 521*, 1962.

[10] R. Connelly, "Rigidity and energy," *Inventiones Mathematicae*, vol. 66, pp. 11–33, 1982.

[11] J. Aldrich, "Control synthesis for a class of light and agile robotic, tensegrity structures," Ph.D. dissertation, Department of Mechanical and Aerospace Engineering, University of California, San Diego, 2004.

[12] C. Tzeng and J. Chen, "Fundamental properties of linear ship steering dynamic models," *Journal of Marine Science and Technology*, vol. 7, no. 2, pp. 78–88, 1999.

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# Polymer Extrusion Process Monitoring Using Nonlinear Dynamic Model-based PCA

Xueqin Liu, Kang Li, Marion McAfee, Jing Deng

*Abstract*— **Polymer extrusion is one of the final forming stages in the production of many polymeric products in a variety of applications. It is also an intermediate processing step in injection moulded, blown film, thermo-formed, and blow moulded products. However, polymer extrusion is a complex process which is difficult to set up, monitor, and control. As a consequence, high levels of off- specification products and long down-times are the problems facing the plastics industry. This paper proposes a new method for fault detection of the polymer extrusion processes, where the nonlinear finite impulse response (NFIR) model and principal component analysis (PCA) are integrated to form a nonlinear dynamic model-based PCA monitoring scheme. Here the NFIR model is used to capture the nonlinearity and dynamics of the extrusion process. The residuals resulting from the difference between the model predicted outputs and process outputs are then analyzed by PCA to detect process faults. The experimental results confirm the efficacy of the proposed model-based PCA approach for fault detection of polymer extrusion processes.**

**Keywords**
Principal component analysis; nonlinear dynamic model; polymer extrusion process.

## I. Introduction

Extrusion has been used widely as a major method of processing polymer materials for a few decades. However, polymer extrusion is a complex process which is difficult to set up, monitor, and control. As a consequence, high levels of off- specification products and long down-times are the problems facing the plastics industry. Thus the close monitoring of the system performance to provide early detection of significant process changes or disturbances, is recognized by the industry to be of increasing strategic importance. Various fault detection methods have been developed based on the first principles, the identified causal models [1] or the multivariate statistical process control including principal component analysis (PCA) [2] and independent component analysis [3].

It was reported that PCA had been successfully applied to a continuous polymer film production line [4], [5]. However, the conventional PCA is a linear method, it may not be able to describe the nonlinear and dynamic characteristics of the extrusion processes properly as they are complex in nature and nonlinear relationship exists between the process variables. To cope with this problem, extended versions of PCA for describing system nonlinear or dynamic behavior have been developed. For nonlinear process monitoring, nonlinear extensions of PCA have been investigated which include principal curves [6], [7], multi-layer auto-associative neural networks (ANNs) [8], [9], and the kernel function approach [10], [11]. For dynamic process monitoring, linear dynamic PCA (DPCA) was first proposed by augmenting matrix with time-lagged variables [12]. More recently, subspace identification for a state space model was proposed [13] which mainly deals with the linear dynamic system. Again, an attempt has been made to apply the time-lagged data extension with Kernel PCA for handling nonlinearity and dynamics. However, this combination could be computationally expensive due to the augmented data matrix [14]. Recently, Rotem *et. al* [15] presented a model-based PCA approach, where the system nonlinear and dynamic behavior are described by first-principle models. Unfortunately, it is often difficult to obtain such models to describe the complex thermodynamic behavior in the polymer extrusion process. To overcome this problem, this paper proposes a new method for monitoring of the nonlinear dynamic polymer extrusion processes, where the nonlinear finite impulse response (NFIR) model based on the Fast Recursive Algorithm (FRA) and the PCA are integrated to form a nonlinear dynamic model-based PCA monitoring scheme. Different from the DPCA, which makes use of all the potential time-lagged variables, the FRA determines and selects the most important and relevant nonlinear dynamic terms from the potential ones to construct the NFIR model. Here the NFIR model, which is a nonlinear extension of FIR [16], is employed to capture the nonlinear and dynamic behavior of the extrusion process. The residuals resulting from the difference between the model predicted outputs and process outputs are analyzed by PCA to detect process faults. The effectiveness of the proposed model-based PCA approach was demonstrated by the monitoring results for data recorded from a polymer extrusion process.

The paper is organized as follows. The PCA and the dynamic PCA methods are briefly described in Section II. This is followed by the proposed nonlinear dynamic model-based PCA for process monitoring in Section III. The experimental polymer extrusion process and data generation are described in Section IV. Section V then presents the monitoring results of an application study to a polymer extrusion process. A concluding summary is given in Section VI.

This work was supported by the Engineering and Physical Science Research Council (EPSRC), grant number EP/F021070/1 and EP/G059489/1, the UK-China Science Bridge project on sustainable energy and built environment, grant number EP/G042594/1.

X. Liu, K. Li and J. Deng is with the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, BT9 5AH, U.K. k.li@qub.ac.uk

M. McAfee is with the Department of Mechanical and Electronic Engineering, Institute of Technology Sligo, Sligo, Ireland. McAfee.Marion@itsligo.ie

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

## II. PCA And Dynamic PCA

This section provides the description of the PCA and the dynamic PCA for process monitoring.

Suppose a data matrix $\mathbf{X} \in \mathbb{R}^{N\times m}$ consists of $N$ samples and $m$ variables, the PCA decomposition allows the construction of two statistics for a sample vector $\mathbf{x} \in \mathbb{R}^m$, a Hotelling's $\mathrm{T}^2$ statistic and a Q statistic:

$$T^2 = \mathbf{x^T P}\Lambda^{-1}\mathbf{P}^T\mathbf{x} = \mathbf{t^T}\Lambda^{-1}\mathbf{t}, Q = \mathbf{x}^T\left[\mathbf{I}-\mathbf{PP}^T\right]\mathbf{x} \quad (1)$$

for which confidence limits can be calculated based on [2], $\Lambda$ is a diagonal matrix consisting of $r$ eigenvalue of covariance matrix of scaled $\mathbf{X}$. $\mathbf{t} = \mathbf{x^T P}$ is a score vector, and $\mathbf{P} \in \mathbb{R}^{m\times r}$ ($r \leq m$ is the number of retained PCs) is a loading matrix.

Dynamic PCA proposed by [12] arranges the process variables to form an autogressive (AR) structure:

$$\underline{\mathbf{X}} = [\mathbf{X}_0, \mathbf{X}_{-1}, \cdots \mathbf{X}_{-d}] \in \mathbb{R}^{(N-d)\times(d+1)m} \quad (2)$$

where $\underline{\mathbf{X}}$ is an augmented set of variables, representing an AR model structure of order $d$ and the subscript $0, 1, d$ refer to the backshifts. The PCA is then applied to $\underline{\mathbf{X}}$ and the corresponding $T^2$ and $Q$ statistics can be constructed. The above method could be computationally expensive due to the augmented variables. Moreover, the dynamic PCA model is obtained under the assumption that the recorded variables are linear, therefore, it may not be suitable to model the nonlinear dynamics efficiently. To tackle this problem, a new model-based PCA method is proposed to account for both the nonlinear and dynamic behaviors of the extrusion process. If the model is accurate, the residual between the measured values and the model predicted values will be relatively insensitive to the variations caused by the nonlinearity or dynamics of the normal operating conditions [15], [16]. Consequently, PCA is more sensitive to process variation caused by the process faults, such as the disturbance of the material variations. In the following section, a new model-based PCA monitoring scheme will be introduced in more detail.

## III. Nonlinear Dynamic Model-based PCA for Process Monitoring

The proposed model-based PCA involves a nonlinear dynamic modelling approach to identify a non-linear finite impulse response model to capture the underlying relationship between the process input and output variables. The residuals resulting from the difference between the model predicted outputs and process outputs are analyzed by PCA to detect process faults.

### *A. Nonlinear Dynamic Model Based On Fast Recursive Algorithm*

Assuming a general nonlinear dynamic MISO system can be formulated as

$$y(t) = f(u_1(t-1),\cdots,u_1(t-d_{u_1}),\cdots,u_p(t-1),\cdots,u_p(t-d_{u_p}) \quad (3)$$

where $y$ and $u_i$ ($i = 1,\cdots,p,\ p \leq m$) are the system output and input variables respectively. $p$ is the number of input variables. $d_{u_i}$ is the time delay for the process inputs $u_i$. By using a polynomial function, Eq. (3) can be approximated by a linear-in-the-parameter model:

$$\mathbf{y} = \sum_{i=1}^{M} \theta_i \phi_i(\mathbf{u}) + \mathbf{e} \quad (4)$$

where $\phi_i(.)$, $(i = 1,\cdots,M)$ are all candidate model terms, $\mathbf{u} = [\mathbf{u}_1,\cdots,\mathbf{u}_p]^T$, $u_i^T = [u_i(t-1),\cdots,u_i(t-d_{u_i})]$ is the model input vector, and $\mathbf{e}$ is the model residual.

If $N$ data samples $\{y,\mathbf{u}\}^N$ are used for model training, then Eq. (4) can be written in the form

$$\mathbf{y} = \Phi\Theta + \mathbf{e} \quad (5)$$

where $\Phi = [\phi_1,\ldots,\phi_M] \in \mathbb{R}^{(N-d)\times M}$ is the regression matrix, $\phi_i = [\phi_i(u(d-d_{u_i}+1)),\cdots,\phi_i(u(N-d_{u_i}))]^T$; $\mathbf{y} = [y(d+1),\ldots,y(N)]^T \in \mathbb{R}^{N-d}$, $d$ is the maximum delay among $d_{u_i}$. $\Theta = [\theta_1,\cdots,\theta_M] \in \mathbb{R}^M$, and $\mathbf{e} = [e(d+1),\cdots,e(N)]^T \in \mathbb{R}^{N-d}$.

In Eq.(5), $\Theta$ can be estimated using least-squares by minimizing the loss function

$$J(\Theta) = \mathbf{e}^T\mathbf{e} \quad (6)$$

The corresponding solution is given by $\widehat{\Theta} = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{y}$.

Due to the noise and correlation between regressors, the matrix $(\Phi^T\Phi)^{-1}\Phi^T$ is always ill-conditioned, which may lead to inaccurate calculation of the model coefficients. Therefore, a subset selection technique such as the Fast Recursive Algrithm [17] is applied in order to determine the most important and relevant terms of all potential ones with respect to the available data [18].

**Fast Recursive Algorithm**

The FRA employs a residual matrix $\mathbf{R}_k$ defined as:

$$\mathbf{R}_k \triangleq I - \Phi_k(\Phi_k^T\Phi_k)^{-1}\Phi_k^T \qquad \mathbf{R}_0 \triangleq I \quad (7)$$

where $\Phi_k = [\phi_1,\cdots,\phi_k]$, and $k = 1,\cdots,M$. According to [17] and [19], $\mathbf{R}_k$ can be updated recursively:

$$\mathbf{R}_{k+1} = \mathbf{R}_k - \frac{\mathbf{R}_k\phi_{k+1}\phi_{k+1}^T\mathbf{R}_k^T}{\phi_{k+1}^T\mathbf{R}_k\phi_{k+1}},\ k = 0,1,\cdots,M-1 \quad (8)$$

Now, the cost function in (6) can be rewritten as: $J(\Theta,\Phi_k) = tr(\mathbf{y}^T\mathbf{R}_k\mathbf{y})$.

In a forward stepwise stage, the nonlinear model terms are selected one at a time. Thus, suppose at the $k$th step, one more term $\phi_j$, $k+1 \leqslant j \leqslant M$ from the candidate term pool is to be selected. The net contribution of $\phi_j$ to the cost function can be calculated as

$$\Delta J_{k+1}(\Theta,\Phi_k,\phi_j) = \frac{\parallel (\mathbf{y}^{(k)})^T\phi_j^{(k)} \parallel^2}{\parallel \phi_j^{(k)} \parallel^2} \quad (9)$$

where $\phi_j^{(k)} \triangleq \mathbf{R}_k\phi_j$, $\mathbf{y}^{(k)} \triangleq \mathbf{R}_k\mathbf{y}$.

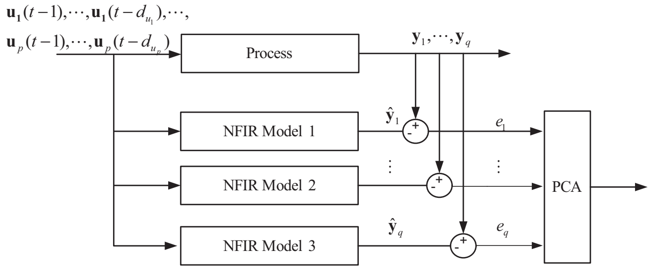

$\mathbf{u}_1(t-1),\cdots,\mathbf{u}_1(t-d_{u_1}),\cdots,$
$\mathbf{u}_p(t-1),\cdots,\mathbf{u}_p(t-d_{u_p})$
Process
$\mathbf{y}_1,\cdots,\mathbf{y}_q$
NFIR Model 1
NFIR Model 2
NFIR Model 3
$\hat{\mathbf{y}}_1$
$\hat{\mathbf{y}}_q$
$e_1$
$e_q$
+
-
PCA

Fig. 1: Nonlinear dynamic model-based PCA

By defining an auxiliary matrix $\mathbf{A} \in \mathbb{R}^{k \times M}$ and $\mathbf{a}_u \in \mathbb{R}^{M \times q}$ with elements

$$a_{i,j} \triangleq \begin{cases} 0, & 1 \leq j < i \\ (\phi_i^{(i-1)})^T \phi_j^{(i-1)}, & i \leq j \leq M \end{cases} \tag{10}$$

$$\mathbf{a}_{i,u} \triangleq \begin{cases} (\phi_i^{(i-1)})^T \mathbf{y}^{(k)}, & 1 \leq i \leq k \\ (\phi_i^{(k)})^T \mathbf{y}^{(k)}, & k < i \leq M \end{cases} \tag{11}$$

the cost function can be updated recursively:

$$\Delta J_{k+1}(\Theta, \Phi_k, \phi_j) = \frac{\| \mathbf{a}_{j,u} \|^2}{a_{j,j}} \tag{12}$$

The model term that provides the largest contribution to (12) is then selected, and this procedure continues until some criterion (e.g., Akaike's information criterion [20]) is met or until a pre-set maximum number of model terms are selected. After a satisfactory model with $M$ terms has been constructed, the model coefficients are then computed recursively from

$$\hat{\theta}_j = \left( \mathbf{a}_{j,u}^T - \sum_{i=j+1}^{M} \hat{\theta}_i a_{j,i} \right) / a_{j,j}, \quad j = k, k-1, \cdots, 1. \tag{13}$$

### B. Nonlinear Dynamic Model-Based PCA Monitoring Scheme

Having discussed the nonlinear dynamic modelling approach based on the FRA, the model-based PCA monitoring scheme, as shown in Fig. 1, will be introduced in detail, where the nonlinear models are employed to predict process outputs. The residuals between the system outputs and predicted outputs are used to construct the PCA model. The following summarizes the construction procedure of the proposed model-based PCA and its application to the process monitoring:

1) Process dynamic data including the input and output variables are recorded for nonlinear dynamic modelling;
2) The nonlinear modelling approach based on the FRA is applied to obtain the NFIR models to describe the nonlinear dynamic relationship between the system input and output variables;
3) The residual resulting from the difference between the model outputs and process outputs is calculated as $\mathbf{e}_i = \mathbf{y}_i - \hat{\mathbf{y}}_i$ ($i = 1, \cdots, q$, $q$ is the number of output variables).
4) The residual matrix $\mathbf{E} = [\mathbf{e}_1, \cdots, \mathbf{e}_q]$ is constructed and scaled;
5) Compute the PCA model for the scaled residual matrix $\mathbf{E}$;
6) Construct the $T^2$ and the $Q$ statistics, and calculate the confidence limit as discussed in [21].
7) For a centered new data sample $\mathbf{z}$ during the on-line monitoring procedure, the scaled residual $\mathbf{e}_z$ between the actual process output and the model predicted ouput is calculated;
8) Calculate the constructed $T^2$ and $Q$ statistics by:

$$T^2 = \mathbf{e}_z^T \mathbf{P} \Lambda \mathbf{P}^T \mathbf{e}_z \qquad Q = \mathbf{e}_z^T \left[ \mathbf{I} - \mathbf{P}\mathbf{P}^T \right] \mathbf{e}_z, \tag{14}$$

9) Check whether $T^2$ or $Q$ exceeds the corresponding control limit; if so, the hypothesis that the soft sensor is violated is accepted, otherwise the soft sensor is reliable.

It is clear that the effectiveness of the proposed model-based PCA method relies on the sensitivity of the system model to the different types of process faults. For instance, if process fault only affects the process output, then it is important that the model-based PCA should be insensitive to the process outputs. Therefore, if the NARX (nonlinear autoregressive with exogenous input) structure is employed, which relies on the past values of the process outputs, then the generated residuals couldn't reflect the influence of the process fault on the the process output. Consequently its monitoring ability can be compromised. However, with the NFIR structure, which only makes use of the past values of the process inputs, the predicted outputs are not affected by the process faults. As a result, the model residuals are capable of reflecting the deviation of process outputs from those obtained without process faults. Based on the above discussion, the NFIR model is preferred over the NARX model to be incorporated by PCA.

## IV. Experimental Polymer Extrusion Process

There are two types of extrusion processes in the polymer industry: continuous and discontinuous. According to the number of screws used in the extruder, there are single-screw, twin-screw and multi-screw extruders. Of these single screw continuous extruders are the most commonly used in polymer industry. The conventional single screw extrusion process has a standard setup including a barrel which is heated by a number of electrical heaters, a rotating screw, and an extrusion die for the final product [22]. Typical temperature and pressure sensors are installed along the barrel and the die to provide continuous data for the process. For process operation, the polymer materials are fed into the barrel through the hopper by gravity, and then they are conveyed and melted along the flights of the screw and finally pushed out through the die to achieve a desired form.

Different types of faults may occur in a polymer extrusion process, such as (i) incoming material variations in terms of size, physical properties and compositions; (ii) process upsets; (iii) equipment faults; (iv) the operator errors.

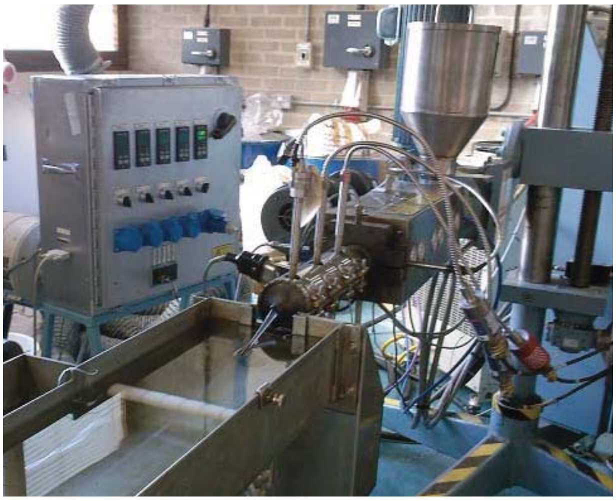


Fig. 2: The laboratory extruder with a in-line-rheometer die

TABLE I: Description Of The Recorded Process Variables

| Number | Description | Variable | Unit | Note |
|---|---|---|---|---|
| 1 | Screw speed | $N$ | rpm | Input |
| 2 | Barrel zone 1 temperature | $T_1$ | $^o$C | |
| 3 | Barrel zone 2 temperature | $T_2$ | $^o$C | |
| 4 | Barrel zone 3 temperature | $T_3$ | $^o$C | |
| 5 | Barrel pressure | $P_b$ | MPa | Output |
| 6 | Die pressure 2 | $P_2$ | MPa | |
| 7 | Viscosity | $\eta$ | Pas | |

Figure 2 depicts the experimental single screw extruder (Killion KTS-100) used in this paper. An in-line-rheometer die is especially instrumented in the experiment for the calculation of the melt viscosity. The analyzed variables are given in Table I. For obtaining some information-rich data sets of process inputs, the screw speed, $N$, and the temperature settings at the three heating zones, $T_1$, $T_2$, $T_3$, were excited using a Pseudo-Random Signal applied in a 'random walk' algorithm respectively. That is the signal excited by a Gaussian sequence and the period of input change was also defined by a Gaussian sequence where the mean and standard deviation were defined on the basis of measured pressure and viscosity response time to step changes in the inputs. Thus a wide operating range was covered in the sequences while consecutive input changes were within practical operating limits [23].

Two low-density polyethylene (LDPE) materials were used in this work, one is LD159AC (LDPE(A)) and the other is DOW352E (LDPE(B)). All signals were acquired at 1Hz using a 16-bit DAQ card through a SC-2345 connector box. At first, 6000 process data samples of LDPE(A) were collected under the above conditions and used as fault free reference samples. These samples were divided into two data sets where the first 2000 samples was used for model identification using the FRA, and the remained 4000 samples for test purpose. A third data set of 4000 samples was collected under the same conditions as above with a different material LDPE(B), which is used as a faulty data set. The generated screw speed, barrel pressure and the viscosity signals under normal conditions are illustrated in Fig. 3.

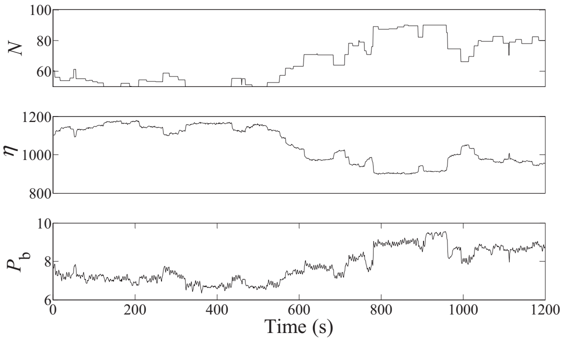


Fig. 3: Generated signals under normal conditions

## V. Application of Model-Based PCA to Polymer Extrusion Process

The description of the polymer extrusion line used in this application study was explained in Section IV. This section discusses the proposed monitoring approach involved in the NFIR model identification for process outputs using the FRA, followed by the PCA applied to the generated residuals.

To build the system model, the screw speed $N$, and the barrel set temperatures $T_1, T_2, T_3$, are employed as the process inputs to predict the three outputs, including the barrel pressure $P_b$, the die pressure $P_2$, and the viscosity $\eta$. As mentioned above, the different model structure employed by model-based PCA has distinct sensitivity to the process faults. To illustrate this point, both NARX and NFIR models are used to predict the process outputs. The three process outputs are generated using the FRA based on the LDPE(A) fault free data with a time series of system variables as below. The NFIR model is given as:

$$\begin{aligned} \mathbf{P}_b(t) &= -2.7 + 10^{-4}\mathbf{N}(t-1)\eta(t-1) \qquad (15)\\ \mathbf{P}_2(t) &= 2.2 + 10^{-4}\mathbf{N}(t-1)\eta(t-1)\\ \eta(t) &= 202.5 - 0.06\mathbf{T_3}(t-1) - 0.05\mathbf{N}(t-1)\mathbf{T}_3(t-5)\\ &\quad +0.03\mathbf{N}(t-1)\mathbf{T}_3(t-8) \end{aligned}$$

The NARX model has the form of

$$\begin{aligned} \mathbf{P}_b(t) &= -0.2 + 0.6\mathbf{P}_b(t-1) + 3\times10^{-5}\mathbf{N}(t-1)\eta(t-1) \quad (16)\\ \mathbf{P}_2(t) &= 0.3 + 0.4\mathbf{P}_b(t-1) + 3\times10^{-5}\mathbf{N}(t-1)\eta(t-1)\\ \eta(t) &= 364.8 + 10^{-4}\eta(t-1) - 0.01\eta(t-1)^2\\ &\quad +0.04\mathbf{N}(t-1)\mathbf{T}_3(t-24) \end{aligned}$$

The error residuals of the above NFIR model has zero mean and variance of 0.04, which can be approximated by a normal distribution. Its performance on the unseen validation data of LDPE(A) including the measured barrel pressure, the melt pressure in the die, and the viscosity is shown in Fig. 4. It shows that the predicted values of the unseen data based on the NFIR model match the measured values very well. For comparison, the NARX model performance on the same data is also illustrated in Fig 4.

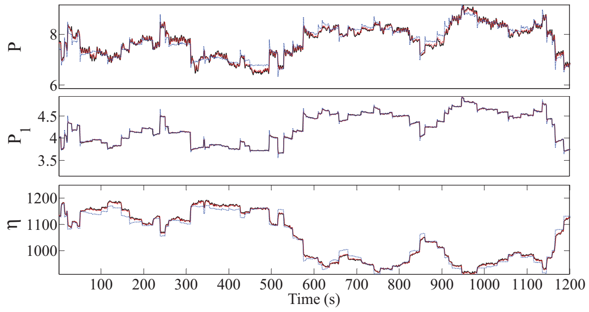


Fig. 4: Modeling result of the NFIR and the NARX on unseen LDPE(A) data under normal conditions. Solid black line: actual output; red dashed line: NARX model output; blue dashed line: NFIR model output.

## A. Process Monitoring

An attempt to monitor an extrusion process using PCA directly will lead to unavoidable difficulties because of the nonlinear and dynamic nature of the process [7], [24], [25]. In this subsection, the monitoring results of applying the proposed model-based PCA on the residual matrix is investigated. A PCA model was produced using the residual matrix **E**, which is composed of the errors between the model predicted outputs and the measured outputs. The proposed NFIR-PCA approach for process monitoring is further compared to both conventional linear DPCA and the NARX-PCA in this section.

The number of principal components, $r$, was determined according to its variance contribution. Thus, two principal components which could capture 79.1% of the variance of the of the total variance of the 3 error variables were chosen for the NFIR-PCA model. Compared to NARX-PCA, two linear principal components captured 86.2% of the total variance. By contrast, one principal component was required for the linear DPCA model and 96.4% of the total variance was explained. After establishing the PCA model, on-line monitoring of the extrusion process requires the $T^2$ statistic to monitor system variations in the PCA model space, and a second $Q$ statistic to monitor system variations in the PCA residual space. The 99% confidence limits for the $T^2$ and $Q$ statistics are determined respectively as discussed in [21].

For the recorded 2000 reference samples, Table II summarizes the Type I error, or false alarm rate, for the $T^2$ and the $Q$ statistics of the linear DPCA, the NARX-PCA and the NFIR-PCA methods for a confidence of 99%. These results imply that the Type I errors for the $T^2$ statistic of linear DPCA is far lower than expected, which is later shown to result in its insensitivity of detecting process fault. This can be attributed to that the principal component in linear DPCA is unable to describe the nonlinear behavior in the extrusion process and that the assumption of the monitored variables follow a Gaussian distribution no longer holds.

TABLE II: Number of Type I errors For Reference Data From Linear DPCA, NFIR-PCA And NARX-PCA With 99% Control Limits

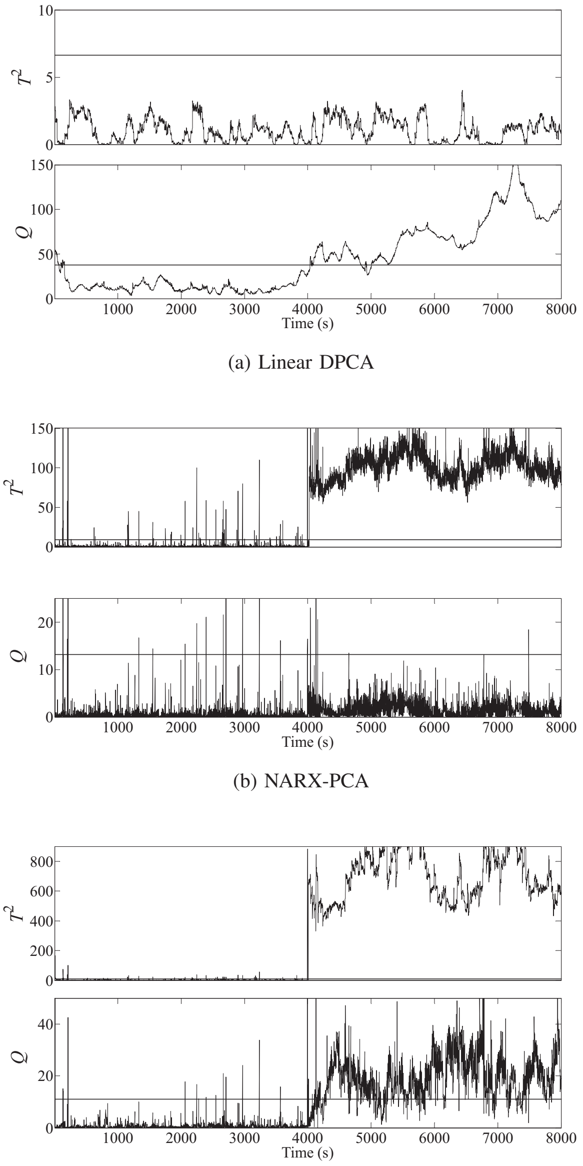
| Method | #PC(s) | Variance Contribution | $T^2$ | $Q$ |
|---|---|---|---|---|
| Linear DPCA | 1 | 96.4% | 0 % | 2.2% |
| NARX-PCA | 2 | 86.2% | 1.9% | 0.9% |
| NFIR-PCA | 2 | 79.1% | 2.0% | 0.7% |

Monitoring results of the proposed method for the testing data set 1) the unseen LDPE(A) fault free data; 2) the unseen LDPE(B) faulty data are shown in Fig. 5. The straight lines represent the 99% control limits. Fig. 5c shows all the unseen LDPE(A) data (the first 4000 samples) is under the control limit, which implies the NFIR model is capable to capture the nonlinear and dynamic behavior. Moreover, for the second fault data set, both $T^2$ and $Q$ statistics of NFIR-PCA in Fig. 5c show the data samples are beyond the control limit (50 samples after the new material introduced) implying the NFIR model is no longer validated for different material. In contrast, the $T^2$ of linear DPCA and the $Q$ of the NARX-PCA are not sensitive to this disturbance. The application study therefore indicated that using linear principal components and incorrect distribution function to describe nonlinear dynamic behavior may render the monitoring statistics insensitive or increase the false alarms. It also demonstrated that the monitoring ability of the NARX-model structure, which relies on the past values of the process outputs, is compromised when the process outputs are affected by process faults.

# VI. Conclusions

This paper has studied the incorporation of NFIR model into the multivariate statistical process control framework, motivated by the fact that monitoring processes with linear dynamic model may lead to insensitive statistics or false alarms. The application of the NFIR model to remove the nonlinear and dynamic information from the monitored variables, together with the use of PCA to monitor the resulting residuals can thus help to circumvent the above problems. The benefit of applying the NFIR model instead of the NARX is that the NFIR model requires no feedback of the process output, and hence its model residual is able to reflect the process faults more closely.

A further contribution of this paper has been to apply the Fast Recursive Algorithm for model identification. In comparison with the traditional subset selection method, such as the orthogonal least square algorithm (OLS), the FRA is able to select the most important and relevant model terms more efficiently without compromising model accuracy [17], [19], [26]. Unlike DPCA, which makes use of all the potential time-lagged data variables, the FRA selects only the most important ones for the NFIR model. The effectiveness of the proposed model-based PCA approach



(a) Linear DPCA

(b) NARX-PCA

(c) NFIR-PCA

Fig. 5: Process monitoring results on the unseen normal data (the first 4000 samples) and the faulty data (after 4000 samples) using (a) DPCA; (b) NARX-PCA; (c) NFIR-PCA

has been demonstrated by the monitoring results for data recorded from a polymer extrusion process.

## References

[1] B. Lennox, G. A. Montague, A. M. Frith, C. Gent, and V. Bevan, "Industrial application of neural networks - an investigation," *Journal of Process Control*, vol. 11, no. 5, pp. 497 – 507, 2001.

[2] J. E. Jackson, *A Users Guide to Principal Components*, ser. Wiley Series in Probability and Mathematical Statistics. New York: John Wiley, 1991.

[3] X. Liu, L. Xie, U. Kruger, T. Littler, and S. Wang, "Statistical-based monitoring of multivariate non-Gaussian systems," *AIChE Journal*, vol. 54, no. 9, pp. 2379–2391, 2008.

[4] S. Valle, S. Qin, and M. Piovoso, "Extracting fault subspaces for fault identification of a polyester film process," in *American Control Conference, 2001. Proceedings of the 2001*, 2001.

[5] Q. P. He, S. J. Qin, and J. Wang, "A new fault diagnosis method using fault directions in fisher discriminant analysis," *AIChE Journal*, vol. 51, no. 2, pp. 555–571, 2005.

[6] D. Dong and T. J. McAvoy, "Nonlinear principal component analysis-based on principal curves and neural networks," *Computers and Chemical Engineering*, vol. 20, no. 1, pp. 65–78, 1996.

[7] X. Liu, K. Li, M. McAfee, and J. Deng, "Application of nonlinear PCA for fault detection in polymer extrusion processes," *Neural computing and applications*, 2011, 10.1007/s00521-011-0581-y.

[8] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE Journal*, vol. 37, no. 3, pp. 233–243, 1991.

[9] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[10] B. Scholkopf, A. J. Smola, and K. R. Muller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, pp. 1299–1399, 1998.

[11] Z. Ge, C. Yang, and Z. Song, "Improved kernel pca-based monitoring approach for nonlinear processes," *Chemical Engineering Science*, vol. 64, no. 9, pp. 2245 – 2255, 2009.

[12] W. Ku, R. H. Storer, and C. Georgakis, "Disturbance rejection and isolation by dynamic principal component analysis," *Chemometrics & Intelligent Laboratory Systems*, vol. 30, pp. 179–196, 1995.

[13] A. Simoglou, E. B. Martin, and A. J. Morris, "Statistical performance monitoring of dynamic multivariate processes using state space modeling," *Computers & Chemical Engineering*, vol. 26, no. 6, pp. 909–920, 2002.

[14] S. W. Choi and I. B. Lee, "Nonlinear dynamic process monitoring based on dynamic KPCA," *Chemical Engineering Science*, vol. 59, pp. 5897–5908, 2004.

[15] Y. Rotem, A. Wachs, and D. R. Lewin, "Ethylene compressor monitoring using model-based PCA," *AIChE Journal*, vol. 46, no. 9, pp. 1825–1836, 2000.

[16] C. Cheng and M.-S. Chiu, "Nonlinear process monitoring using jitl-pca," *Chemometrics and Intelligent Laboratory Systems*, vol. 76, no. 1, pp. 1 – 13, 2005.

[17] K. Li, J. Peng, and G. W. Irwin, "A fast nonlinear model identification method," *IEEE Transactions on Automatic Control*, vol. 50, no. 8, pp. 1211–1216, 2005.

[18] O. Nelles, *Nonlinear System Identification: From Classical Approaches to Neural Networks and Fuzzy Models*. Berlin: Springer-Verlag Berlin Heidelberg, 2001.

[19] K. Li, J. X. Peng, and E. W. Bai, "A two-stage algorithm for identification of nonlinear dynamic systems," *Automatica*, vol. 42, no. 7, pp. 1189–1197, 2006.

[20] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[21] P. Nomikos and J. F. MacGregor, "Multivariate SPC charts for monitoring batch processes," *Technometrics*, vol. 37, no. 1, pp. 41–59, 1995.

[22] C. Rauwendaal, *Polymer Extrusion (4th edition)*. Carl Hanser Verlag, Munich, 2001.

[23] M. McAfee and S. Thompson, "A novel approach to dynamic modelling of polymer extrusion for improved process control," *Proceedings of the Institution of Mechanical Engineers, Part I, Journal of Systems and Control*, vol. 221, no. 4, pp. 617–628, 2007.

[24] X. Liu, K. Li, M. McAfee, and J. Deng, "'soft-sensor' for real-time monitoring of melt viscosity in polymer extrusion process," in *Decision and Control (CDC), 2010 49th IEEE Conference on*, 2010, pp. 3469 –3474.

[25] X. Liu, K. Li, M. McAfee, B. K. Nguyen, and G. M. McNally, "Dynamic gray-box modeling for on-line monitoring of polymer extrusion viscosity," *Polymer Engineering & Science*, vol. 52, no. 6, pp. 1332–1341, 2012.

[26] X. Liu, K. Li, M. McAfee, and G. W. Irwin, "Improved nonlinear PCA for process monitoring using support vector data description," *Journal of Process Control*, vol. 21, no. 9, pp. 1306 – 1317, 2011.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# A New Bandwidth Scheduling Method for Networked Learning Control

Lijun Xu*†, Kang Li†, Minrui Fei*, Dajun Du*
*Shanghai Key Laboratory of Power Station Automation Technology,
School of Mechatronical Engineering and Automation, Shanghai University, Shanghai, China
Email: ruby_mickey@sina.com mrfei@staff.shu.edu.cn
†School of Electronics, Electrical Engineering and Computer Science,
Queen's University Belfast, Belfast, BT9 5AH, UK
Email: k.li@qub.ac.uk

***Abstract*—In this paper, the optimal bandwidth allocation scheduling problem for two-layer networked learning control systems (NLCSs) is studied. In NLCS, multiple networked feedback control loops share a common communication channel and they compete to bid for available bandwidth. A non-cooperative game fairness model is first formulated, which takes into consideration of a number of factors, such as transmission data rate, control sampling strategy and scheduling pattern. Then, a novel two-layer hierarchical market competition algorithm (THMCA) is proposed. Two hierarchical population individuals are defined in the algorithm, namely the holding companies and the subsidiary companies which altogether form conglomerates. Market competitions among these conglomerates lead to the convergence to a monopoly at the end, resulting in an optimal solution of the above problem. The algorithm is shown to have a high convergence rate and the comparison simulation results on a NLCS with up to 100 subsystems have demonstrated the effectiveness of the proposed method.**

## I. Introduction

A networked control system (NCS) [1] is defined as a feedback control system where the control loops are closed via a communication network. Due to its various advantages such as low cost installation, ease of maintenance and great flexibility, NCSs have been widely applied in manufacturing, aircraft and power systems etc. Most researches have focused on the traditional single-layer NCSs in the last decade. However, there exist many complex plants that are composed of a great number of subsystems and the traditional single-layer NCS architecture may not be applicable. Recently, a two-layer Networked Learning Control System (NLCS) architecture has been proposed [2]. In this architecture, the bottom layer is for real-time control, and local controllers communicate with the sensors and actuators attached to the plant through shared networks. The upper layer is used for complex learning and scheduling tasks.

In a real-time NLCS, limited network bandwidth along end-to-end paths [3] inevitably cause network-induced delays [4] or packet dropout which further deteriorates the system performance even causes instability. Therefore, an optimized scheduling for bandwidth allocation among all network links plays a key role in improving the quality of network service (QoS) and the performance of control system (PoC). This can be achieved generally through the following two steps. Firstly, bandwidth scheduling is designed to allocate the available bandwidth to control units to meet communication demand. A non-cooperative game (NG) [5] theory based rational competition mechanism can be employed, which emphasizes individual rational defined by Nash equilibrium (NE) and can simultaneously satisfy individual and collective requirements. Secondly, the exact solution to the bandwidth scheduling problem can be obtained by a complete enumeration.

This is however prohibitive due to its excessive computational time for real-time applications. To tackle this problem, several intelligent optimization methods can be adopted, for example genetic algorithm (GA) [6], particle swarm optimization (PSO) [7], clone evolutionary algorithm [8], bacterial foraging (BF) [9] and shuffled frog leaping algorithm (SFLA) [10] etc. Evolutionary algorithms (EAs), such as GA and PSO, are stochastic based search methods. GA is one of the early proposed evolutionary algorithms which has found many successful applications. However, it is computationally expensive, and the convergence cannot be guaranteed. Bacterial foraging (BF) algorithm is a feature selection method based on a heuristic search strategy with fast computing speed. But a large storage capacity is required to complete the computation. Shuffled frog leaping algorithm (SFLA) is a meta-heuristic optimization method which is based on observing, imitating, and modeling of the behavior of a group of frogs searching for the location where maximum amount of food is available. The SFLA combines the benefits of both the genetic-based memetic algorithm and the social behavior-based PSO algorithm. However SFLA suffers from the curse of dimensionality problem.

In this paper, a new integer-coded two-layer hierarchical market competition algorithm (THMCA) is proposed to efficiently solve the bandwidth scheduling problem in the NLCS applications. The rest of the paper is organized as follows. Section 2 introduces NLCS and the non-cooperative game scheduling scheme. A new two-layer hierarchical market competition algorithm (THMCA) is proposed in Section 3. Section 4 presents the comparative simulation results. Finally, a brief conclusion is given in Section 5.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

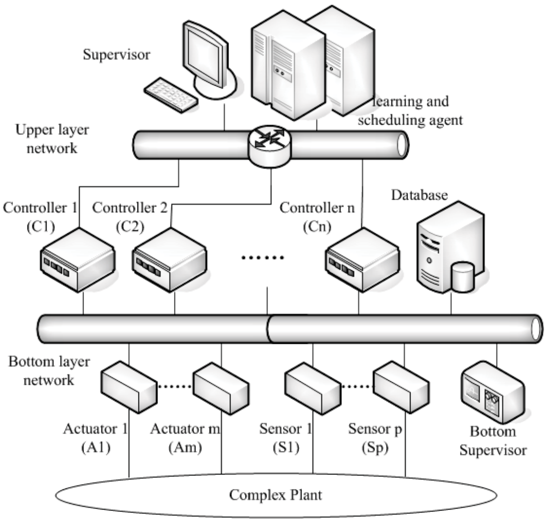


Fig. 1. A two-layer networked learning control system (NLCS)

## II. Bandwidth non-cooperative game model for a networked learning control system

### A. System Architecture

The two-layer networked learning control system architecture [2] is shown in Fig.1, where Ci, Si and Ai represent the i-th controller, sensor and actuator respectively. Local controllers are connected to the sensors and actuators, attached to the complex plant via the bottom layer communication network, typically some fieldbus dedicated to real-time control. Local controllers also communicate with the computer system, which functions as a learning and scheduling agent through the upper layer communication network. This network can be local area network (LAN), wide area network (WAN), or possibly the internet. Control and learning signals at two levels share the available network bandwidth with other consumers. This general architecture can be adopted in many industrial applications with distributed plants and units, such as the power systems.

### B. Bandwidth non-cooperative game model

Suppose the set of subsystems is denoted as $L = \{\ell_i | 1 \leq i \leq n\}$, where $n$ is the total number of subsystems in a NLCS. The network can only provide limited available bandwidth. Each subsystem aims to improve its own bandwidth utilization instead of the overall network performance. Therefore, all the subsystems in a NLCS form a non-cooperative game (NG). The NG model is a triple, namely $G(L, S, u_i)$, where $S$ is the set of strategies including transmission data rate, allotted bandwidth and sampling period for each subsystem, that is, $S = \{s_i | s_i = (b_i, \delta_i, t_i), i \in L\}$, $b_i \in [\delta_i^{min}, U_d]$ is the pre-allocated data rate to the subsystem $l_i$, and $t_i$ is the sampling period. Data rate set available for all subsystems in NLCS is given as $D = \{\delta_i | 1 \leq i \leq n\}$, $\delta_i \in [\delta_i^{min}, \delta_i^{max}]$. The minimum data rate $\delta_i^{min}$ is to guarantee the most basic work. Maximum bandwidth of the network is $U_d$ and $\delta_i^{max} \leq U_d$. $u_i$ is the utility function of subsystem $l_i$ which maps $S$ to real numbers $u : S \to R$. If and only if $u_i(s_i^*) > u_i(s_i)$, the quality of $s_i^*$ is better than $s_i$ for subsystem $l_i$.

The available data rate $\delta_i$ satisfies the probability distribution $d_i(\delta_i)$ defined as follows [11].

$$d_i(\delta_i) = \begin{cases} 0 & \delta_i < \delta_i^{min} \\ (\Delta/\Delta_i)^c & \delta_i^{min} \leq \delta_i \leq \delta_i^{max} \\ 1 & \delta_i > \delta_i^{max} \end{cases} \tag{1}$$

where $\Delta = \delta_i - \delta_i^{min}$, $\Delta_i = \delta_i^{max} - \delta_i^{min}$ and $c \geq 1$ works as an empirical constant which can be adjusted for a specific subsystem.

Given the above definition, the mean value of occupied bandwidth for a subsystem is thus given as

$$Ed_i(\delta_i) = (c\delta_i^{max} + \delta_i^{min})/(c+1) \tag{2}$$

The utility function is defined as follows:

$$u_i(s_i) = \begin{cases} \frac{[\|e_i\|^2 - (\delta_i - Ed_i)^2 - \gamma_i t_i](\frac{b_i}{Ed_i} e)^\mu}{e^{\mu \frac{b_i}{Ed_i}} \|e_i\|^3} \\ if\ 0 \leq b_i < Ed_i \\ [1 - (\frac{b_i - Ed_i}{Ed_i - \delta_i^{max}})^\nu] \frac{\|e_i\|^2 - (\delta_i - Ed_i)^2 - \gamma_i t_i}{\|e_i\|^3} \\ if\ Ed_i \leq b_i \leq \delta_i^{max} \end{cases} \tag{3}$$

Here, the rights law is used, including compounding data rate $\delta_i$ and sample period $t_i$, and $\gamma_i$ is the weight coefficient for trade-off. $\|e_i\|$ is the maximum expectation deviation between optional data rate and the distribution of each subsystem, that is $\|e_i\| = \max(|\delta_i^{min} - Ed_i|, |\delta_i^{max} - Ed_i|)$. $\mu, \nu$ are the empirical constants for specific subsystems to adjust the change rate of utility functions.

## III. Two-layer hierarchical market competition algorithm (THMCA)

The two-layer hierarchical market competition algorithm (THMCA) is inspired by competitions among enterprises in economic activities. The THMCA begins with an initial population called perfect competition companies. Some of the best companies that have the best objective fitness function values are selected to be the holding companies. The rest become the subsidiary companies which are then divided among the holding companies based on their power. The power of a holding company is positively proportional to its fitness value. The holding companies and their subsidiary companies form different conglomerates. Then subsidiary companies move toward their relevant holding companies and the position of the holding companies will be updated if necessary. In the next step, the market competition among the conglomerates begins, and the weak conglomerates are eliminated. This is called monopolistic competition procedure. Then the oligopoly procedure begins. The market competition will gradually lead to an increase in the power of strong conglomerates (oligopolies) and a decrease in the power of weaker ones. Finally the weak conglomerates which could not to improve their performance will collapse. These competitions among the conglomerates will cause all the companies to converge to a state called monopoly where only one conglomerate exits in the market and all the other companies become subsidiary companies of this holding company.

To begin with, an initial population called perfect competition companies is created. In a $D$-dimensional problem, the position of the $i$-th company is defined as $Company_i = [x_{i,1}, x_{i,2}, \cdots, x_{i,D}]$, $i = 1, 2, \cdots, N_c$, where $N_c$ is total number of competition companies.

The fitness function of the $i$-th company is defined as:

$$fit_i(Company_i) = fit_i(x_{i,1}, x_{i,2}, \cdots, x_{i,D}) \tag{4}$$

Then, the cost function of the $i$-th company can be defined as:

$$f_i = \frac{1}{fit_i(x_{i,1}, x_{i,2}, \cdots, x_{i,D})} \tag{5}$$

$N^{holding}$ of the most powerful competition companies are selected as the holding companies, which form different conglomerates. The remaining $N^{sub}$ are the subsidiary companies of these holding companies. At the next step, the subsidiary companies must be divided among the holding companies based on their power. The initial number of subsidiary companies of a conglomerate is directly proportional to its power. Theoretically, the normalized cost of the $n$-th holding company can be defined as:

$$F_n^{holding} = \max_i\{f_i^{holding}\} - f_n^{holding} \tag{6}$$

where $f_n^{holding}$ is the cost of the $n$-th holding company.

The normalized power of each holding company is defined as:

$$P_n^{holding} = \|F_n^{holding} / \sum_{n=1}^{N^{holding}} F_n^{holding}\| \tag{7}$$

Then the initial number of subsidiary companies of a conglomerate becomes

$$N_n^{sub} = round\{P_n^{holding} \times N^{sub}, 0\} \tag{8}$$

where $N_n^{sub}$ is the initial number of subsidiary companies of the $n$-th conglomerate. For each holding company, $N_n^{sub}$ of the subsidiary companies are randomly selected and allocated. These subsidiary companies along with their holding company form the conglomerate.

Then the subsidiary companies start to move toward their relevant holding companies. The positions of the subsidiary companies of the $n$-th conglomerate are updated as follows:

$$SUB_{n,i} = sub_{n,i} + \frac{rand() \times \omega(holding_n - sub_{n,i})}{\cos\theta} \tag{9}$$

where $sub_{n,i}$ is the position of the $i$-th subsidiary company of the $n$-th holding company, $rand()$ is a random number between 0 and 1, $\omega$ is a weight factor, and $holding_n$ is the position of the $n$-th holding company. To search different points around the holding company, a random amount of deviation is added to the direction of movement. The movement of a subsidiary company toward its relevant holding company at its new direction $\theta$ is a random angle between $-\varepsilon$ and $\varepsilon$, where $\varepsilon > 0$ is the parameter that adjusts the deviation from the original direction.

However, a subsidiary company in a conglomerate may reach a position with higher fitness (or lower cost) than its holding company. In such case, the positions of the holding company will be replaced by the higher one. The rest will move toward the new position of the holding company.

The total power of a conglomerate depends on both the power of the holding company and the power of its subsidiary companies. But the holding company has larger weights. This total power is defined by the weighted cost of two hierarchical companies:

$$power_n^{cong} = \frac{1}{p_n^{cong}} = \frac{1}{f_n^{holding} + \tau\overline{f_{n,i}^{sub}}} \tag{10}$$

where $p_n^{cong}$ is the total cost of the $n$-th conglomerate and $0 < \tau < 1$. Cost $\overline{f_{n,i}^{sub}}$ is the geometry mean of $i$-th subsidiary company in $n$-th conglomerate. In fact, $\tau$ represents the role of the subsidiary companies in determining the total power of a conglomerate.

The market competition among conglomerates begins and all the conglomerates try to take possession of the subsidiary companies of other conglomerates. This competition is modeled by picking some of the weakest subsidiary companies of the weakest conglomerates and making a competition among all conglomerates to possess these subsidiary companies. Each of the conglomerates will have a likelihood of taking possession of these subsidiary companies based on its total power; therefore, powerful conglomerates have greater chance to possess subsidiary companies. The possession probability of each conglomerate must be found. The normalized total cost of each conglomerate is calculated as:

$$P_n^{cong} = \max_i\{p_i^{cong}\} - p_n^{cong} \tag{11}$$

where $P_n^{cong}$ is the normalized total cost of the $n$-th conglomerate.

The possession probability of each conglomerate is given by

$$pp_n^{cong} = \|P_n^{cong} / \sum_{n=1}^{N^{holding}} P_n^{cong}\| \tag{12}$$

where $pp_n^{cong}$ is the possession probability of the $n$-th conglomerate. A vector is formed to divide the relevant subsidiary companies among the conglomerates:

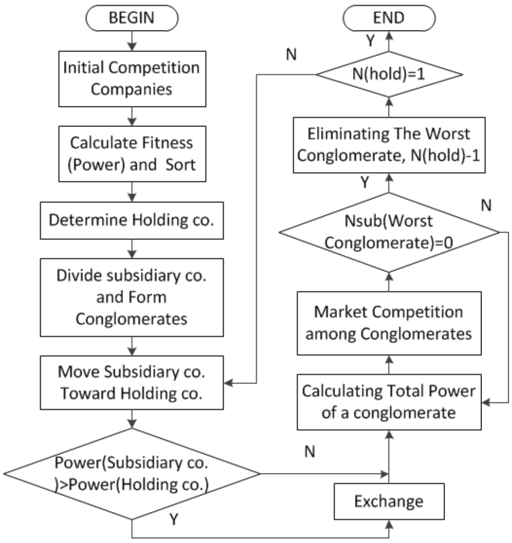

BEGIN
Initial Competition Companies
Calculate Fitness (Power) and Sort
Determine Holding co.
Divide subsidiary co. and Form Conglomerates
Move Subsidiary co. Toward Holding co.
Power(Subsidiary co. )>Power(Holding co.)
Y
N
Exchange
Calculating Total Power of a conglomerate
Market Competition among Conglomerates
Nsub(Worst Conglomerate)=0
N
Y
Eliminating The Worst Conglomerate, N(hold)-1
N(hold)=1
N
Y
END

Fig. 2. Flowchart of two-layer hierarchical market competition algorithm

$$\mathbf{PP}^{cong} = [pp_1^{cong}, pp_2^{cong}, \cdots, pp_{N^{holding}}^{cong}] \tag{13}$$

Then a random vector with the same size is generated:

$$\mathbf{RAND} = [rand_1, rand_2, \cdots, rand_{N^{holding}}] \tag{14}$$

where $rand_1, rand_2, \cdots, rand_{N^{holding}}$ are randomly generated numbers between 0 and 1.

Finally, vector $\mathbf{Dif}$ is formed by subtracting $\mathbf{RAND}$ from $\mathbf{PP}^{cong}$:

$$\begin{aligned} \mathbf{Dif} &= \mathbf{PP}^{cong} - \mathbf{RAND} \\ &= \begin{pmatrix} PP_1^{cong} - rand_1, \\ PP_2^{cong} - rand_2, \\ \cdots \\ PP_{N^{holding}}^{cong} - rand_{N^{holding}} \end{pmatrix}^{\mathrm{T}} \end{aligned} \tag{15}$$

The mentioned subsidiary companies will be given to an conglomerate which has the maximum relevant index in $\mathbf{Dif}$ vector.

The powerless conglomerates will collapse in the market competition. Different criteria can be defined for collapse mechanism. In this paper, a conglomerate is assumed to be collapsed when it loses all of its subsidiary companies.

After the market competitions, all the conglomerates will collapse except the most powerful one and all companies under their possession become subsidiary companies of this conglomerate. All the subsidiary companies also have the same positions and the same fitness. In such case, the algorithm stops. The flowchart of the procedure is illustrated in Fig.2.

In the integer coded THMCA, the company position consists of a sequence of integer numbers, representing the sequence of the bandwidth allocation cycle duration of each subsystem during the scheduling horizon. A positive integer represents an increased bandwidth allocation, while a negative integer represents the bandwidth decrease of the subsystem.

TABLE I
RANGE OF THE OUTPUT DATA RATES OF SUBSYSTEMS FOR TWO DISTRIBUTIONS

| Data Rate | DND | | SD | |
|---|---|---|---|---|
| Subsystems | $\delta_i^{min}$ | $\delta_i^{max}$ | $\delta_i^{min}$ | $\delta_i^{max}$ |
| $l_1$ | 20.58 | 75.51 | 32 | 51 |
| $l_2$ | 41.63 | 88.68 | 56 | 85 |
| $l_3$ | 53.45 | 97.20 | 19 | 31 |
| $l_4$ | 64.02 | 106.71 | 98 | 120 |
| $l_5$ | 76.84 | 111.69 | 28 | 42 |
| $l_6$ | 77.68 | 115.60 | 102 | 135 |
| $l_7$ | 65.97 | 108.10 | 49 | 76 |
| $l_8$ | 56.32 | 99.59 | 84 | 116 |
| $l_9$ | 46.24 | 90.73 | 156 | 196 |
| $l_{10}$ | 32.30 | 80.18 | 246 | 380 |

## IV. SIMULATION STUDIES

To verify the effectiveness of the proposed bandwidth scheduling method, numerical simulations were carried out using MATLAB on an Intel(R) Core(TM)-i5-3.10-GHZ desktop personal computer (PC) with Windows 7 Enterprise.

The proposed THMCA was tested on NLCSs with $Num$ = 10, 20, 50 and 100 subsystems. The required data for ten subsystems is given in table I, which shows the distributions of sending data rates for the first ten subsystems. Two different distributions, namely the Discrete Normal Distribution (DND) and Stochastic Distribution (SD) are generated to cover general cases. The simulation test system first has 10 subsystems. Initially, the network bandwidth is $U_d = 1000Kbps$. For NLCS with 20 subsystems, the data of the ten subsystems was duplicated and the total bandwidth available was multiplied by two. For the problem with more subsystems, the data was scaled appropriately.

The optimal THMCA parameters for ten subsystems which were chosen after several runs are given as $N^{holding} = 10$, $N^{sub} = 200$, $\tau = 0.2$, $\omega = 2$. For NLCS with more subsystems, the same parameters were utilized, except for $N^{holding}$ and $N^{sub}$ which increase correspondingly. Another parameter to be selected is $\theta \in (45°, 90°)$, which adjusts the deviation from the original direction. These values were found suitable to produce good solutions in terms of the processing time and the quality of the solutions. The fitness function is set to be the utility function of NG, that is, $fit_i = u_i$.

For the comparison purpose among the THMCA and other optimization methods, all the simulation experiments used the same basic parameter settings.

The convergence of the algorithm for ten subsystems in NLCS with 50, 100, 150, 200, 250 and 300 initial competition companies with discrete nomal distribution data rate is shown in Fig.3. It is clear that the initial number of 200 competition companies is notably the best on both convergence (mean solution time) and the bandwidth consumption.

From the results of 10 simulation runs, it is found that the optimal solution can be obtained after 8-th to 10-th market competition interactions. Fig.4 shows the convergence of the

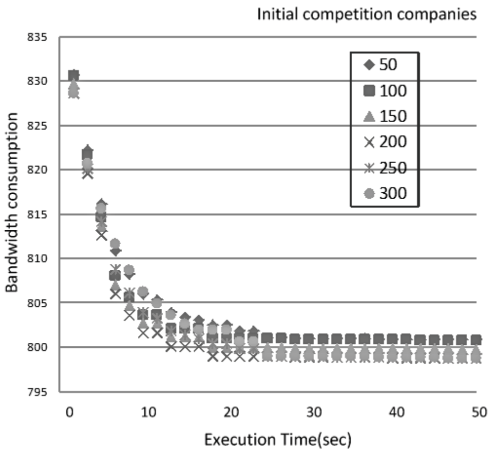


Fig. 3. Convergence characteristics of market competition algorithm with different initial competition companies in 10-subsystem NLCS

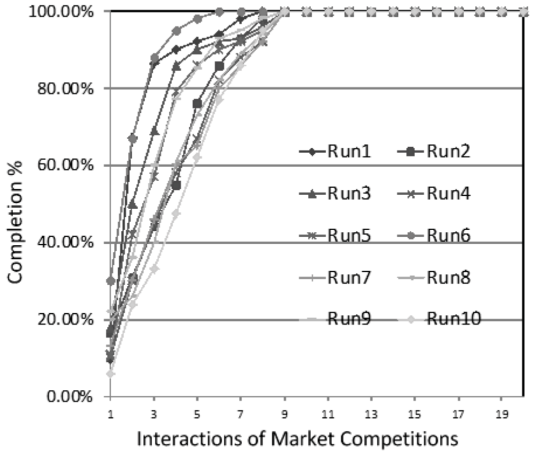


Fig. 4. 10-runs convergence of market competition algorithm in 10-subsystem NLCS

iteration for the 10-subsystem test system with stochastic distribution data rate. This indeed verifies the high convergence rate of the algorithm.

When the overall fitness value is stabilized, the nash equilibrium point is reached. For discrete normal distribution, this is: $\{b_1, b_2, \cdots, b_{10}\} = \{55.90,\ 68.12,\ 77.56,\ 93.37,\ 101.34,\ 105.82,\ 94.70,\ 84.18,\ 60.79,\ 56.84\}$. For stochastic distribution, this is $\{38,\ 78,\ 27{,}110,\ 36,\ 117,\ 62,\ 101,\ 107,\ 318\}$.

Figure 5 and 6 illustrate the proportion of statistical output data rates with lower limit, the reserved data rate and output data limit of the two distributions. It is evident that the game model based method is able to produce a fair network resource allocation in NLCS under different constraints. The overall performance of the system remained stable by effectively restricting large network resource utilization.

Figure 7 shows the ratio between the scheduled sampling period of subsystems and the original sampling period $t_i^*/t_i$, where the lower ratio means a better optimization scheme. The optimization results of the sampling period after the adjustment within the requirements of the two data rates distributions is also shown. Further, when the discrete normal distribution data rates was used, larger probability distribution leads to better optimization results, indicating that this method is adequate for meeting the requirements of most data rates.

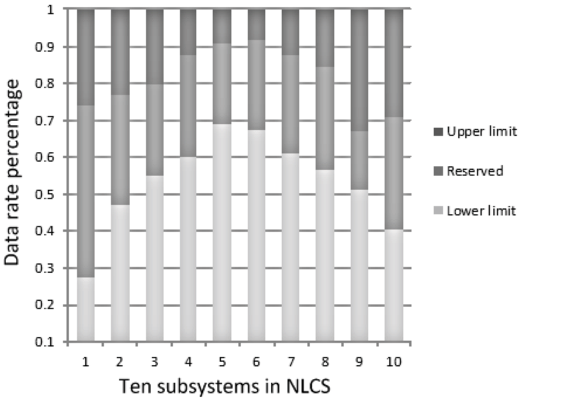


Fig. 5. Data rate percentage plot of DND in 10-subsystem NLCS

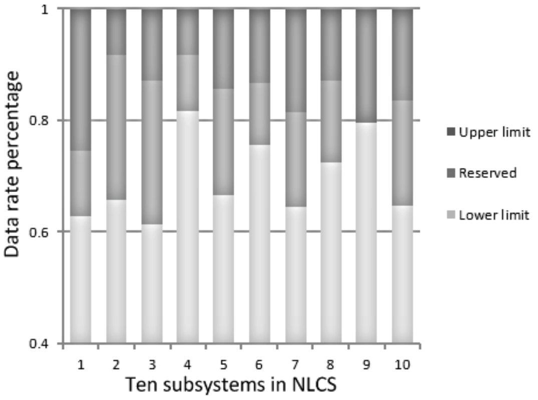


Fig. 6. Data rate percentage plot of SD in 10-subsystem NLCS

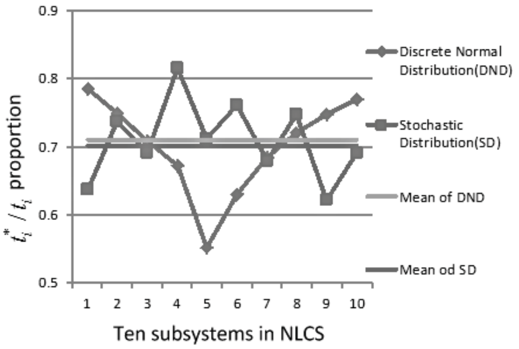


Fig. 7. $t_i^*/t_i$ proportion plot in 10-subsystem NLCS

The best bandwidth results obtained by THMCA are compared with those obtained by the shuffled frog leaping algorithm (SFLA) [11], BF [9], GA [6], hybrid quantum clone evolutionary algorithm (HQCE) [8], and quantum inspired

TABLE II
COMPARISON OF TOTAL BANDWIDTH CONSUMPTION(KBPS) WITH DIFFERENT NUMBER OF SUBSYSTEMS

| $Num$ | GA | HQCE | Q-PSO | BF | SLFA | THMCA |
|---|---|---|---|---|---|---|
| 10 | 825 | 817 | 807 | 823 | 803 | 798 |
| 20 | 1643 | 1630 | 1611 | 1641 | 1601 | 1591 |
| 50 | 4105 | 4074 | 4026 | 4099 | 4002 | 3976 |
| 100 | 8209 | 8149 | 8053 | 8196 | 8005 | 7952 |

TABLE III
COMPARISON OF AVERAGE EXECUTION TIME(SEC) WITH DIFFERENT NUMBER OF SUBSYSTEMS

| $Num$ | GA | HQCE | Q-PSO | BF | SLFA | THMCA |
|---|---|---|---|---|---|---|
| 10 | 30 | 24 | 22 | 28 | 21 | 17 |
| 20 | 82 | 65 | 61 | 75 | 58 | 46 |
| 50 | 214 | 169 | 158 | 196 | 147 | 119 |
| 100 | 1504 | 1189 | 1112 | 1376 | 1076 | 836 |

PSO (Q-PSO) [7] for NLCS with up to 100 subsystems in table II.

The execution time is also an important factor. Table III lists the execution time with different size of subsystem obtained by THMCA, SFLA, BF, GA, HQCE, and Q-PSO. It is obvious that the execution time of THMCA increases linearly with the size of the bandwidth scheduling problem. The overall execution time obtained by THMCA is less than that of other methods.

## V. CONCLUSIONS

Resource allocation in a networked learning control system is a constrained nonlinear optimization problem. A fair non-cooperation game model has been proposed in this paper. Then, the resource allocation problem is transformed into a problem of settling the equilibrium point of the game model. A new optimization algorithm namely two-layer hierarchical market competition algorithm (THMCA) has been proposed to solve the problem effectively. The proposed method has been tested and compared with a few alternatives. Simulation results show that the computational time and bandwidth consumptions of THMCA are less than other algorithms such as SFLA, GA, BF, quantum-inspired PSO, and hybrid quantum clone evolutionary algorithm. However, the performance of the THMCA also depends on certain parameters selected. Future research includes developing more efficient algorithms and addressing uncertainties in the proposed non-cooperation game model.

## ACKNOWLEDGMENT

This work was financially supported by National Natural Science Foundation of China under Grant No.61074032,61104089, Project of Science and Technology Commission of Shanghai Municipality under Grant No.10JC1405000, 11ZR1413100, and UK-China Science Bridge Project under RCUK grant EP/G042594/1.

## REFERENCES

[1] T. Yang, *Networked control system: a brief survey*, IEE Proc.-Control Theory Appl, vol. 153, no. 4, pp. 403412, 2006.

[2] D. Du, M. Fei and K. Li, *A two-layer networked learning control system using actor-critic neural network*,Applied Mathematics and Computation, vol. 205, no. 1,pp.2636,2008.

[3] R. M. Salles and J. A. Barria, *Lexicographic maximin optimisation for fair bandwidth allocation in computer networks*, European Journal of Operational Research., vol. 185, no. 2, pp.778794,2008.

[4] W. Deng, K. Li, G. W. Irwin and M. Fei, *Identification and control of Hammerstein systems via wireless networks*, Int. J. of Systems Science, pp. 113, published online Feb 2012.

[5] A. Ganesh, K. Laevens and R. Steinberg, *Congestion Pricing and Noncooperative Games in Communication Networks*, Operations Research, vol. 55, no. 3, pp. 430438,2007.

[6] Q. Chen, Y. Zhong and X. Zhang, *A pseudo genetic algorithm*, Neural Comput & Applic, vol.19, no. 1, pp.77 83, 2010.

[7] J. Zhang, J. Wang and C. Yue, *Small Population-Based PSO for Short-Term Hydrothermal Scheduling*, Power Systems, IEEE Transactions on, vol. 27, no. 1, pp. 142152,2012.

[8] L. Xu and M. Fei, *A Hybrid Quantum Clone Evolutionary Algorithm-Based Scheduling Optimization in a Networked Learning Control System*, Chinese Control and Decision Conference (CCDC) 2010, pp.36323637,2010.

[9] M. Eslamian, S. H. Hosseinian and B. Vahidi, *Bacterial foragingbased solution to the unit-commitment problem*, IEEE Trans. on Power Syst., vol. 24, no. 3, pp. 14781488, 2009.

[10] J. Ebrahimi, S. H. Hosseinian and G. B. Gharehpetian, *Unit Commitment Problem Solution Using Shuffled Frog Leaping Algorithm*, IEEE Transactions on Power Systems, vol. 26, no. 2, pp.19,2011.

[11] L. Xu, M. Fei, T. Jia and T. Yang, *Bandwidth scheduling and optimization using non-cooperative game model-based shuffled frog leaping algorithm in a networked learning control system*, Neural Computing & Applications, published online September 2011.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# An Improved Conjugate Gradient Algorithm for Radial Basis Function (RBF) Networks Modelling

Long Zhang and Kang Li
School of Electronics, Electrical Engineering
and Computer Science,
Queen's University Belfast, UK
Email: lzhang14@qub.ac.uk,
k.li@qub.ac.uk

Shujuan Wang
School of Electrical Engineering and Automation,
Harbin Institute of Technology, China,
Email: wsj603@hit.edu.cn

***Abstract*—This paper proposes a new nonlinear optimization algorithm for the construction of radial basis function (RBF) networks in modelling nonlinear systems. The main objective is to speed up the learning convergence of the conventional conjugate gradient method. All the hidden layer parameters of RBF networks are simultaneously optimized by the conjugate gradient method while the output weights are adjusted accordingly using the orthogonal least squares (OLS) method. The derivatives used in the conjugate gradient algorithm are efficiently computed using a recursive sum squared error criterion. Numerical examples show that the new method converges faster than the previously proposed continuous forward algorithm (CFA).**

## I. Introduction

The Radial Basis Function (RBF) networks are well known for strict interpolation for approximating scattered data in multi-dimensions [1]. It has been proved that a RBF network can approximate any multi-variate continuous function if a sufficient number of RBF nodes are provided [2]. The RBF networks have been successfully applied in nonlinear system identification [3], signal processing [4] and fault diagnosis [5]. A standard RBF network consists of a nonlinear hidden layer and a linear output layer. The training of such RBF networks involves the optimization of hidden layer parameters (centers and widths) and linear output weights, with the aim of minimizing the cost function like sum squared error (SSE) or Akaike information criterion (AIC) [6].

A variety of methods for training RBF networks have been proposed. Unsupervised clustering methods and the supervised least squares based subset selection methods are widely used. Though these methods, like $k$-means clustering [7], orthogonal least squares (OLS) [8] and the fast recursive algorithm (FRA) [9], are fast and powerful techniques, they may not build a compact model. To address this problem, methods combining the OLS with evolutionary algorithms [10], [11] have been proposed. However, they are often computationally expensive and suffer from slow convergence due to their random search nature [12].

Alternative solutions are gradient based methods, like conjugate gradient and Newton algorithms [13]. However, they treat all the hidden layer parameters and output weights separately without the consideration of the correlation between these two sets of parameters, therefore, they may converge slowly or not at all if the initial guess is far from the minimum. To speed up the convergence and simplify the computational complexity, the continuous forward algorithm (CFA) [14] has been proposed recently. This method employs the conjugate gradient algorithm to optimize the hidden layer parameters while the optimal output weights are transformed into a set of dependant parameters based on the least squares method, thus without explicit optimizing the output weights. The CFA can be considered as a stepwise algorithm as it optimizes one hidden node at a time. Its advantage is that the model parameters and model size can be determined simultaneously and the computational complexity is not high for each iteration of optimization. The disadvantage is that the learning convergency can be slow and the parameters are not necessarily optimal as the previously obtained parameters are fixed and become the constraints when optimizing the new hidden layer nodes. In other words, it does not optimize all the hidden layer parameters simultaneously.

In this paper, a conjugate gradient method combined with OLS approach is introduced to construct RBF networks, in which all the hidden layer parameters are optimized simultaneously by the conjugate gradient method while the output weights are adjusted accordingly using the OLS method in the continuous space, leading to an improved model performance with fast learning convergence. This is achieved effectively using a recursive sum squared error (SSE) and all the derivatives are updated recursively at each iteration. A numerical example confirms that the new method could converge faster than the CFA.

## II. Problem formulation and preliminaries

A RBF network for modelling a nonlinear system can be formulated as [14]

$$y(t) = \sum_{i=1}^{m} \mathbf{p}_i(\mathbf{x}(t), d_i, \mathbf{s}_i)\theta_i + \xi(t) \tag{1}$$

where $\{\mathbf{x}(t), y(t)\}$ are the system input and output variables at time instant $t$. $\mathbf{x}(t)$ is of assumed known dimension of $q$, $t = 1, 2, \ldots, N$, $N$ being the size of the training data set. $\mathbf{p}_i(\mathbf{x}(t), d_i, \mathbf{s}_i)$ denotes the RBF network function of the $i$th hidden node with the width $d_i \in \Re^1$ and centers $\mathbf{s}_i \in \Re^q$, here $i = 1, 2, \ldots, m$, $m$ being the number of RBF nodes. $\theta_i$ is the weight parameter. $\xi(t)$ is a model residual sequence. For a RBF network, all adjustable parameters are the hidden layer parameters and output weights. More specifically, all the hidden layer parameters can be expressed as

$$\mathbf{V} = \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_m \end{bmatrix} = \begin{bmatrix} d_1 & s_{11} & s_{12} & \ldots & s_{1q} \\ d_2 & s_{21} & s_{22} & \ldots & s_{2q} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ d_m & s_{m1} & s_{m2} & \ldots & s_{mq} \end{bmatrix} \tag{2}$$

where the vector $\mathbf{v}_i$ includes the $i_{th}$ hidden layer node parameters. $d_i$ and the set $\{s_{i1}, ..s_{iq}\}$ are the width and centers of the $i_{th}$ hidden layer node, respectively. For simplicity, $\mathbf{v}_i = [d_1, s_{i1}, \ldots, v_{iq}]$ can be denoted as $\mathbf{v}_i = [v_{i1}, \ldots, v_{ir}]$ where $r = q + 1$.

Equation (1) can be expressed in the matrix form as

$$\mathbf{y} = \mathbf{P\Theta} + \Xi \tag{3}$$

where $\mathbf{y} = [y(1), \ldots, y(N)]^T$ is the output vector, $\mathbf{\Theta} = [\theta_1, \ldots, \theta_m]^T$ is the unknown parameter vector, $\Xi = [\xi(1), \ldots, \xi(N)]^T$ is the residual vector, and $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_m]$ is a $N$-by-$m$ matrix with $\mathbf{p}_j = [p_j(\mathbf{x}(1), \mathbf{v}_j), \ldots, p_j(\mathbf{x}(N), \mathbf{v}_j)]^T$. The RBF network modeling aims to minimize the sum squared error (SSE)

$$\Xi^T\Xi = (\mathbf{y} - \mathbf{P\Theta})^T(\mathbf{y} - \mathbf{P\Theta}) \tag{4}$$

by optimizing $d_i$'s, $\mathbf{s}_i$'s and $\theta_i$'s.

If all the hidden layer parameters are determined, the output weights can then be obtained by least squares methods. The orthogonal least squares (OLS) is the most popular algorithm for determining both the linear parameters and model structure. It computes the output weights using the orthogonal decomposition given by [15]

$$\mathbf{P} = \mathbf{WA} \tag{5}$$

where $\mathbf{A}$ is a $m$-by-$m$ triangular matrix,

$$\mathbf{A} = \begin{bmatrix} 1 & \alpha_{12} & \ldots & \alpha_{1m} \\ 0 & 1 & \ldots & \alpha_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{6}$$

and $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_m]$ is a $N$-by-$m$ matrix with orthogonal columns $\mathbf{w}_i$,

$$\mathbf{W} = \begin{bmatrix} w_{11} & w_{12} & \ldots & w_{1m} \\ w_{21} & w_{22} & \ldots & w_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ w_{N1} & w_{N2} & \ldots & w_{Nm} \end{bmatrix} \tag{7}$$

which satisfies

$$\mathbf{W}^T\mathbf{W} = diag\{\mathbf{w}_1^T\mathbf{w}_1, \ldots, \mathbf{w}_m^T\mathbf{w}_m\}. \tag{8}$$

The orthogonal decomposition involved in OLS can be carried out by the classic Gram-Schimdt (CGS) method. This computes one column of $\mathbf{A}$ at a time and factorizes $\mathbf{P}$ as follows:

$$\left.\begin{array}{l} \mathbf{w}_1 = \mathbf{p}_1 \\ \left.\begin{array}{l} \alpha_{ik} = \dfrac{< \mathbf{w}_i, \mathbf{p}_k >}{< \mathbf{w}_i, \mathbf{w}_i >},\ 1 \leq i < k \\ \mathbf{w}_k = \mathbf{p}_k - \displaystyle\sum_{i=1}^{k-1} \alpha_{ik}\mathbf{w}_i \end{array}\right\} k = 2, \ldots, m \end{array}\right\} \tag{9}$$

The model (3) can thus be expressed as

$$\mathbf{y} = (\mathbf{PA}^{-1})(\mathbf{A\Theta}) + \Xi = \mathbf{Wg} + \Xi \tag{10}$$

where $\mathbf{g} = [g_1, g_2, \ldots, g_m]^T = \mathbf{A\Theta}$ is the orthogonal weight vector. The original model weight vector $\mathbf{g}$ can then be calculated by [8]

$$\mathbf{g} = (\mathbf{W^TW})^{-1}\mathbf{W}^T\mathbf{y} - (\mathbf{W^TW})^{-1}\mathbf{W}^T\Xi. \tag{11}$$

and the sum squares of $\mathbf{y}$ is given by

$$\mathbf{y}^T\mathbf{y} = \mathbf{g}^T\mathbf{W}^T\mathbf{Wg} + \Xi^T\Xi + \mathbf{g}^T\mathbf{W}^T\Xi + \Xi^T\mathbf{Wg} \tag{12}$$

If $\Xi$ is a zero mean white sequence and is uncorrelated with $\mathbf{P}$ and all stochastic processes of interest are ergodic, then $\mathbf{W}^T\Xi = 0$ [8]. Hence (11) and (12) can be further simplified to

$$\mathbf{g} = (\mathbf{W^TW})^{-1}\mathbf{W}^T\mathbf{y}, \tag{13}$$

and

$$\mathbf{y}^T\mathbf{y} = \mathbf{g}^T\mathbf{W}^T\mathbf{Wg} + \Xi^T\Xi, \tag{14}$$

respectively. Finally, the weight $\theta_i$ can be computed using backward elimination

$$\left.\begin{array}{l} \theta_m = g_m \\ \theta_i = g_i - \displaystyle\sum_{k=i+1}^{m} \alpha_{ik}\theta_k, i = m-1, \ldots, 1 \end{array}\right\} \tag{15}$$

Substituting $\mathbf{g}$ in (10) and (14) using (13), we get

$$\Xi = \mathbf{y} - \mathbf{W}(\mathbf{W^TW})^{-1}\mathbf{W}^T\mathbf{y}. \tag{16}$$

and

$$\Xi^T\Xi = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{W}(\mathbf{W^TW})^{-1}\mathbf{W}^T\mathbf{y}. \tag{17}$$

Using the property in (8), the (17) is converted into a recursive form given by

$$\begin{aligned} \Xi^T\Xi &= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\{\sum_{j=1}^{m} \frac{\mathbf{w}_j\mathbf{w}_j^T}{\mathbf{w}_j^T\mathbf{w}_j}\}\mathbf{y} \\ &= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{Ry} \end{aligned} \tag{18}$$

where $\mathbf{R} = \sum_{j=1}^{m} \frac{\mathbf{w}_j\mathbf{w}_j^T}{\mathbf{w}_j^T\mathbf{w}_j}$ is a recursive matrix.

## III. THE PROPOSED METHOD

To address the problem of slow convergence in the conventional conjugate gradient method, the CFA [14] considers the coupling relationship between $\mathbf{V}$ and $\mathbf{\Theta}$. It converts $\mathbf{\Theta}$ and the estimation error sequence $\mathbf{\Xi}$ into $(\mathbf{P}^T\mathbf{P})^{-1}\mathbf{P}^T\mathbf{y}$ and $\mathbf{y}-\mathbf{P}(\mathbf{P}^T\mathbf{P})^{-1}\mathbf{P}^T\mathbf{y}$ using least-squares method respectively, and then computes the derivatives of the estimated error sequence with respect to $\mathbf{V}$. After the updated $\mathbf{V}$ are determined using derivative information, the output weights $\mathbf{\Theta}$ are recalculated by the least squares methods. However, It optimizes each vector parameter $\mathbf{v}_i, i = 1, ..., M$ at a time while other parameters $\mathbf{v}_j, j = 1, ..., M, j \neq i$ being fixed and become the constraints, hence the resultant model may be not optimal, though the computational complexity is significantly reduced in each iteration..

The proposed algorithm not only considers the coupling relationship between $\mathbf{V}$ and $\mathbf{\Theta}$ and but also computes all the hidden layer parameters $\mathbf{V}$ simultaneously. The proposed algorithm first optimizes all the hidden layer parameters in (2) using conjugate gradient algorithm and then adjusts the output weights $\mathbf{\Theta}$ by the OLS method.

The conjugate gradient involves computing the first derivative of the SSE with respective to the hidden layer parameters $\mathbf{V}$, which is given by [16]

$$\nabla(\mathbf{\Xi}^T\mathbf{\Xi}) = 2\left(\frac{\partial\mathbf{\Xi}}{\partial\mathbf{V}}\right)^T\mathbf{\Xi} \tag{19}$$

where the first order derivative of the error sequence $\frac{\partial\mathbf{\Xi}}{\partial\mathbf{V}}$ is also called the Jacobian matrix, given by

$$\left.\begin{aligned} \mathbf{J} &= \frac{\partial\mathbf{\Xi}}{\partial\mathbf{V}} = -\left[\frac{\partial\mathbf{R}}{\partial\mathbf{v}_1}\mathbf{y}, ..., \frac{\partial\mathbf{R}}{\partial\mathbf{v}_m}\mathbf{y}\right] \\ \frac{\partial\mathbf{R}}{\partial\mathbf{v}_k} &= \left[\frac{\partial\mathbf{R}}{\partial v_{k1}}, ..., \frac{\partial\mathbf{R}}{\partial v_{kr}}\right], k = 1, ..., m \\ \frac{\partial\mathbf{R}}{\partial v_{ki}} &= \sum_{j=1}^{m}\frac{\partial\left\{\frac{\mathbf{w}_j\mathbf{w}_j^T}{\mathbf{w}_j^T\mathbf{w}_j}\right\}}{\partial v_{ki}}, i = 1, ..., r \end{aligned}\right\} \tag{20}$$

where $\mathbf{v}_k = \{\mathbf{v}_{k1}, ..., \mathbf{v}_{kr}\}$.

The key step for the conjugate gradient algorithm is the computation of the Jacobian matrix. For each element in the Jacobian matrix shown in (20), the derivative of $\mathbf{w}_k$ with respect to $v_{ki}, k = 1, ..., m, i = 1, ..., r$ needs to be computed. The mapping between $\mathbf{w}_k$ and $v_{ki}$ can be obtained from the relations among $v_{ki}$, $\mathbf{p}_k$ and $\mathbf{w}_k$. Each RBF term $\mathbf{p}_k$ is determined by a vector $\mathbf{v}_k = [v_{k1}, ..., v_{kr}]$. Further, (9) indicates $\mathbf{w}_k$ is dependant on $\mathbf{p}_1, ..., \mathbf{p}_k$ and independent of $\mathbf{p}_{k+1}, ..., \mathbf{p}_m$, hence,

$$\frac{\partial\mathbf{w}_k}{\partial v_{ij}} = \begin{cases} 0, & k < i, j = 1, ..., r \\ \dfrac{\partial\mathbf{w}_k}{\partial v_{ij}}, & k \geq i, j = 1, ..., r \end{cases} \tag{21}$$

and thus, the element in (20) is simplified as

$$\frac{\partial\mathbf{R}}{\partial v_{ki}} = \sum_{j=1}^{m}\frac{\partial\left\{\frac{\mathbf{w}_j\mathbf{w}_j^T}{\mathbf{w}_j^T\mathbf{w}_j}\right\}}{\partial v_{ki}} = \sum_{j=k}^{m}\frac{\partial\left\{\frac{\mathbf{w}_j\mathbf{w}_j^T}{\mathbf{w}_j^T\mathbf{w}_j}\right\}}{\partial v_{ki}}, i = 1, .., r \tag{22}$$

while $k = m$, (22) becomes

$$\frac{\partial\mathbf{R}}{\partial v_{mi}} = \frac{\partial\left\{\frac{\mathbf{w}_m\mathbf{w}_m^T}{\mathbf{w}_m^T\mathbf{w}_m}\right\}}{\partial v_{mi}}, i = 1, .., r \tag{23}$$

Compare (22) and (23), and it is clear that to compute the derivative with respect to the last term parameters $\mathbf{v}_m = [v_{m1}, ..., v_{mi}]$ only $\frac{\partial\mathbf{w}_m}{\partial\mathbf{w}_{mi}}, i = 1, ..., r$, is needed while for other derivatives, say $\mathbf{v}_k = [v_{k1}, ..., v_{ki}]$, only $\frac{\partial\mathbf{w}_k}{\partial\mathbf{v}_{ki}}, ..., \frac{\partial\mathbf{w}_m}{\partial\mathbf{w}_{ki}}, i = 1, ..., r$, need to be computed. To simplify the computational complexity for derivatives, each term except the last one is moved to the last position and its derivatives is then computed as proposed in [12]. If any term position is changed, the orthogonal decomposition procedure needs to be repeated. Suppose $\mathbf{p}_k, k = 1, ..., m-1$, is to be moved to the last position. This leads to shifting the columns $\mathbf{p}_{k+1}, ..., \mathbf{p}_m$, left by one place. This is given by

$$\{\mathbf{p}_1, ..., \mathbf{p}_k, ..., \mathbf{p}_m\} \to \{\mathbf{p}_1, ..., \mathbf{p}_{k-1}, \mathbf{p}_{k+1}, ...\mathbf{p}_m, \mathbf{p}_k\} \tag{24}$$

and consequently their corresponding orthogonal basis is changed. If $k = 1$, then all the orthogonal terms $[\mathbf{w}_1^{'}, ..., \mathbf{w}_m^{'}]$ need to be computed using (9). If $k > 1$, the terms $[\mathbf{w}_1, ..., \mathbf{w}_{k-1}]$ are left unchanged and only $[\mathbf{w}_k^{'}, ..., \mathbf{w}_m^{'}]$ need to alter. In details, shift the position by

$$\left.\begin{aligned} \mathbf{p}_j^{'} &= \mathbf{p}_{j+1}, \ k \leq j < m \\ \mathbf{p}_m^{'} &= \mathbf{p}_k \end{aligned}\right\} \tag{25}$$

and then re-decompose $\mathbf{p}_k^{'}, ..., \mathbf{p}_m^{'}$ by

$$\left.\begin{aligned} \alpha_{nj}^{'} &= \frac{<\mathbf{w}_n, \mathbf{p}_k^{'}>}{<\mathbf{w}_n, \mathbf{w}_n>}, \ 1 \leq n < k \\ \alpha_{nj}^{'} &= \frac{<\mathbf{w}_n^{'}, \mathbf{p}_k^{'}>}{<\mathbf{w}_n^{'}, \mathbf{w}_n^{'}>}, \ k \leq n < j \\ \mathbf{w}_j^{'} &= \mathbf{p}_j^{'} - \sum_{n=1}^{k-1}\alpha_{nj}^{'}\mathbf{w}_n - \sum_{n=k}^{j-1}\alpha_{nj}^{'}\mathbf{w}_n^{'} \end{aligned}\right\} j = k, ..., m. \tag{26}$$

The derivative with respect to the last term $\mathbf{p}_m$ in Jacobian matrix is calculated by

$$\left.\begin{aligned} \frac{\partial\mathbf{p}_m}{\partial v_{mi}} &= \left[\frac{\partial p_m(x(1))}{\partial v_{mi}}, ..., \frac{\partial p_m(x(N))}{\partial v_{mi}}\right] \\ \frac{\partial\mathbf{w}_m}{\partial v_{mi}} &= \frac{\partial\mathbf{p}_m}{\partial v_{mi}} - \sum_{j=1}^{j=m-1}\frac{\mathbf{w}_j^T\frac{\partial\mathbf{p}_m}{\partial v_{mi}}}{\mathbf{w}_j^T\mathbf{w}_j}\mathbf{w}_j \\ J_{mi} &= -\frac{\partial\mathbf{R}}{\partial v_{mi}}\mathbf{y} \\ &= -\frac{\frac{\partial\mathbf{w}_m}{\partial v_{mi}}(\mathbf{w}_m^T\mathbf{y}) + \mathbf{w}_m(\frac{\partial\mathbf{w}_m^T}{\partial v_{mi}}\mathbf{y})}{\mathbf{w}_m^T\mathbf{w}_m} \\ &\quad + 2\frac{\mathbf{w}_m(\mathbf{w}_m^T\mathbf{y})(\frac{\partial\mathbf{w}_m^T}{\partial v_{mi}}\mathbf{w}_m)}{(\mathbf{w}_m^T\mathbf{w}_m)^2} \end{aligned}\right\} i = 1, ..., r \tag{27}$$

Other terms $\mathbf{p}_k, k = 1, ..., m-1$, moved to the last position also use the formula (27) where $\frac{\partial \mathbf{p}_m^{'}}{\partial \mathbf{v}_{mi}^{'}}$ equals $\frac{\partial \mathbf{p}_k}{\partial \mathbf{v}_{ki}}$. To compute $\frac{\partial \mathbf{p}_k}{\partial \mathbf{v}_{ki}}$, the basis function $\mathbf{p}_k = [p_k(\mathbf{x}(1), \mathbf{v}_k), \dots, p_k(\mathbf{x}(N), \mathbf{v}_k)]$ should be given. In this paper, the Gaussian function is considered

$$p_k(\mathbf{x}(t), \mathbf{v}_k) = p_k(\mathbf{x}(t), d_k, \mathbf{s}_k) = \text{exp}(-\eta) \qquad (28)$$

where $\eta = \sum_{i=1}^{i=q} (\frac{x_i - s_{ki}}{d_k})^2$, $t = 1, \dots, N$ and $k = 1, \dots, m$. The first-order partial derivatives with respect to widths and centers are

$$\frac{\partial p_k}{\partial d_k} = \frac{2}{d_k} \eta \text{exp}(-\eta) \qquad (29)$$

and

$$\frac{\partial p_k}{\partial s_{ki}} = \frac{2}{d_k^2} (x_i - s_{ki}) \text{exp}(-\eta) \qquad (30)$$

respectively.

The procedure of the new method is summarized as follows:

*Step 1*: Initialize all the hidden layer parameters $\mathbf{V}$ randomly or using OLS based subset selection method [15], [17]. It should be pointed out that to choose appropriate values could speed up the model convergence.

*Step 2*: Compute Jacobian and Hessian matrices using (25)-(27).

*Step 3*: Update the $\mathbf{V} = \mathbf{V} + \Delta \mathbf{V}$ and construct the basis functions terms $\mathbf{P}$ in the form of (28) using the new parameters $\mathbf{V}$, and then compute the SSE using (9) and (18).

*Step 4*: Repeat step 3 until the SSE is reduced to some error target or the iteration number reaches a given number. After the hidden layer parameters $\mathbf{V}$ are determined, the output weights are calculated by back elimination (15).

## IV. A NUMERICAL EXAMPLE

To verify the efficacy of the new method, a numerical examples is used to test its convergence rate. The comparison with CFA will be also discussed. All the tests were carried out using MATLAB R2010a on a desktop Intel E8400 PC with Windows XP system.

The following nonlinear system model [3]

$$\begin{aligned} y(t) = & -0.6377y(t-1) + 0.07298y(t-2) \\ & + 0.03597u(t-1) + 0.06622u(t-2) \\ & + 0.06568u(t-1)y(t-1) + 0.02357u^2(t-1) \\ & + 0.05939 \end{aligned} \qquad (31)$$

Here $t$ denotes the time series and $u$ and $y$ represent the system input and output, respectively. A data sequence of length 500 was generated for training, with input $u$ being uniformly distributed within $[-1, 1]$.

For comparison, the CFA was also used to model the nonlinear system. For a fair comparison, these two methods used the same initialization procedure. The iteration stops when the SSE reduction rate is less than 0.01. Fig. 1, Fig. 2 and Fig. 3 illustrate the training curves for the CFA and new methods with model size of 5, 6 and 7 separately. These results are averages over 100 different trials. It is shown that the new method converges faster than the CFA.

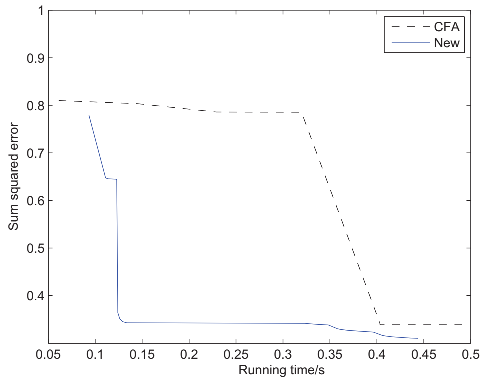


Fig. 1. Variation in training SSE for CFA and the new method with size 5

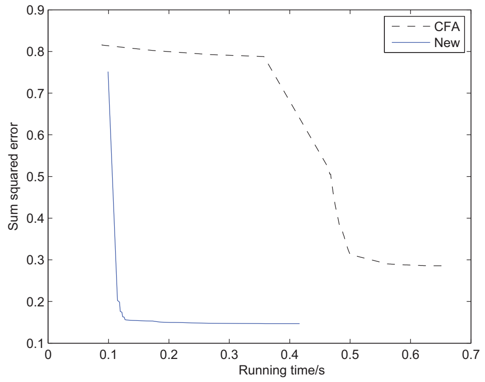


Fig. 2. Variation in training SSE for CFA and the new method with size 6

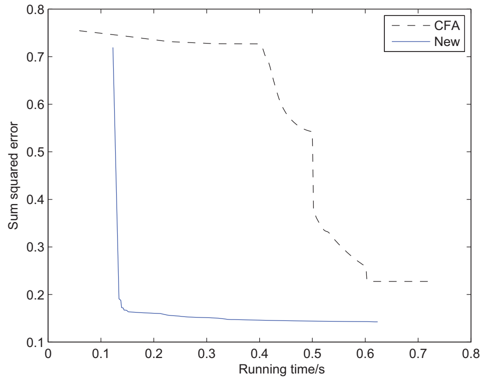


Fig. 3. Variation in training SSE for CFA and the new method with size 7

## V. CONCLUSION

An improved conjugate gradient algorithm for constructing RBF models has been proposed. Unlike previously proposed CFA method which optimizes one hidden node parameters at a time, the new method optimizes all the RBF parameters simultaneously, leading to a faster convergence rate. A numerical example has confirmed the effectiveness of the proposed method.

## Acknowledgment

This work was supported by the RCUK under grant EP/G042594/1 and EP/G059489/1, and partially supported by the Chinese Scholarship Council, National Natural Science Foundation of China under grant 51077022, and Science and Technology Commission of Shanghai Municipality under Grant 11ZR1413100.

## References

[1] M. Buhmann, *Radial basis functions: theory and implementations*. Cambridge University Press, 2003.

[2] W. A. Light, “Some aspects of radial basis function approximation,” *Approximation Theory, Spline Functions and Applications*, vol. 356, pp. 163–190, 1992.

[3] G. Zheng and S. Billings, “Radial basis function network configuration using mutual information and the orthogonal least squares algorithm,” *Neural Networks*, vol. 9, no. 9, pp. 1619–1637, 1996.

[4] S. Chen, C. Cowan, and P. Grant, “Orthogonal least squares algorithm for radial basis funtion networks,” *International Journal of Control*, vol. 2, no. 2, pp. 302–309, 1991.

[5] K. Narendra, V. Sood, K. Khorasani, and R. Patel, “Application of a radial basis function (rbf) neural network for fault diagnosis in a hvdc system,” *Power Systems, IEEE Transactions on*, vol. 13, no. 1, pp. 177–183, 1998.

[6] H. Akaike, “A new look at the statistical model identification,” *IEEE Transaction on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[7] S. Billings, H. Wei, and M. Balikhin, “Generalized multiscale radial basis function networks,” *Neural Networks*, vol. 20, no. 10, pp. 1081–1094, 2007.

[8] S. Billings, S. Chen, and M. Korenberg, “Identification of mimo non-linear systems using a forward-regression orthogonal estimatorn,” *International Journal of Control*, vol. 49, no. 6, pp. 2157–2189, 1989.

[9] K. Li, J. Peng, and G. W. Irwin, “A fast nonlinear model identification method,” *IEEE Transcations on Automatic Control*, vol. 8, no. 50, pp. 1211–1216, 2005.

[10] S. Billings and Y. Yang, “Identification of the neighborhood and ca rules from spatio-temporal ca patterns,” *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, vol. 33, no. 2, pp. 332–339, 2003.

[11] K. Mao and S. Billings, “Algorithms for minimal model structure detection in nonlinear dynamic system identification,” *IEEE Transcations on Neural Networks*, vol. 68, no. 2, pp. 311–330, 1997.

[12] K. Li, J. Peng, and E. Bai, “A two-stage algorithm for identification of non-linear dynamic systems,” *Automatica*, vol. 42, no. 7, pp. 1189–1197, 2006.

[13] D. Marquardt, “An algorithm for least-squares estimation of nonlinear parameters,” *Journal of the society for Industrial and Applied Mathematics*, vol. 11, no. 2, pp. 431–441, 1963.

[14] J. Peng, K. Li, and G. W. Irwin, “A novel continous forward algorithm for rbf neural network,” *IEEE Transactions on Automatic control*, vol. 52, no. 1, pp. 117–122, 2007.

[15] S. Chen, S. Billing, and W. Luo, “Orthogonal least squares methods and their application to non-linear system identification,” *International Journal of Control*, vol. 50, no. 5, pp. 1873–1896, 1989.

[16] M. Hagan and M. Menhaj, “Training feedforward networks with the marquardt algorithm,” *Neural Networks, IEEE Transactions on*, vol. 5, no. 6, pp. 989–993, 1994.

[17] J. Peng, K. Li, and D. Huang, “A hybrid forward algorithm for rbf neural network construction,” *IEEE Transactions on Neural Networks*, vol. 17, no. 6, pp. 1439–1451, 2006.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Heuristically optimized RBF neural model for the control of section weights in stretch blow moulding

Jing Deng*, Ziqi Yang†, Kang Li†, Gary Menary*, Eileen Harkin-Jones *
*School of Mechanical & Aerospace Engineering
Queen's University Belfast, Belfast, BT9 5AH, UK
Email: {j.deng, G.Menary, e.harkinjones}@qub.ac.uk
†School of Electronics, Electrical Engineering and Computer Science,
Queen's University Belfast, Belfast, BT9 5AH, UK
Email: {zyang06, k.li}@qub.ac.uk

***Abstract*—The injection stretch-blow Moulding (ISBM) process is typically used to manufacture PET containers for the beverage and consumer goods industry. The process is somehow complex and users often have to heavily rely on trial and error methods to setup and control it. In this paper, a novel identification method based on a radial basis function (RBF) network model and heuristic optimization methods, such as particle swarm optimization (PSO), deferential evolution (DE), and extreme learning machine (ELM) is proposed for the modelling and control of bottle section weights. The main advantage of the proposed method is that the non-linear parameters are optimized in a continuous space while the hidden nodes are selected one by one in a discrete space using a two-stage selection algorithm. The computational complexity is significantly reduced due to a recursive updating mechanism. Experimental results on simulation data from ABAQUS are presented to confirm the superiority of the proposed method.**

## I. Introduction

Over last few decades, the use of plastics has experienced healthy growth due to its many high performances including lightweight with high tensile/impact/tear strengths, high temperature/chemical resistance, high clarity/modulus/plasticity and low cost [1]. The injection stretch-blow moulding (ISBM) process is a kind of blow-moulding process in making thin-walled polyethylene terephthalate (PET) bottles for the carbonated soft drink and mineral water industries. In such a process, polymer granules is first extruded and injected into a hollow tube to produce structurally amorphous preforms. The resultant preforms are then loaded and conveyed in an infrared oven for reheat. Finally, the heated preforms are simultaneously stretched by a rod and blown with high pressure air to produce to the finished article shape. This stretch blowing and subsequent cooling takes around one second. Due to the fast production rate, the ability to mould complex part and some other attractive features, ISBM process has become one of the most popular methods adopted in the polymer industry.

The quality of made bottles is usually indicated by the bottle wall thickness distribution, top load, burst, and the section weights. All these variables are correlated to the process parameter settings, including preform temperature distribution, blowing air pressure and the delay between rode stretching and air blowing [2]. Unfortunately, the current ISBM process is still an open-loop system where optimal process settings are found by trial and error [3]. This is not only time consuming but also leads to wasting of materials, and the process is sensitive to both internal and external interruptions. Therefore, the design of feedback control for ISBM process is extremely urgent. The bottle wall thickness and its section weights are clearly the best options for feedback signals, but they are either difficult to measure in real-time or the cost is too high. Alternatively, the 'soft-sensor' method to infer these parameters based on the mathematical process model becomes an affordable and cost-effective approach.

In non-linear system identification, the radial basis function networks are recognized as an universal approximation model that has been widely applied in data mining, pattern recognition, signal processing, and system modelling and control [4]. The RBF network has a simple topological structure, and it is easy to be trained compared to the multilayer perceptron (MLP) neural network alternative. The construction of RBF network model mainly involves two steps, the optimization of basis function parameters and the estimation of output layer weights. The latter can be easily achieved by least-square estimation while the former is difficult to implement as it involves non-linear optimization. The conventional way to handle those non-linear parameter are either by exhaustive search or gradient-based methods, which can be computationally expensive, and the global best solution cannot be guaranteed [5].

Heuristic approaches, such as simulated annealing (SA) [6], evolutionary algorithm (EA) [7], Tabu search (TS) [8], particle swarm optimization (PSO) [9], differential evolution (DE) [10], ant colony optimization (ACO) [11] and harmony search (HS) [12] have offered the alternatives for such an optimization problem. Unlike conventional calculus-based methods, heuristic approaches randomly generate some new solutions from which the best one is selected. The employment of heuristic methods in the RBF network construction can be implemented at either the global or the local level. In the former case all the non-linear and linear parameters are optimized simultaneously while in the latter case only the non-linear parameters of a single hidden node are regarded as one solution, and the RBF model is built step by step through subset selection algorithm [13]. Simultaneous optimization

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

of the whole RBF network can archive the global optimal solution. However, it can be computationally expensive, and a large number of iterations are required in the optimization process. By contrast, subset selection involves fewer parameters to be optimized at each step, thus it is more efficient. The recently proposed two-stage subset selection algorithm has also shown its ability in reaching the near global optimal solution while retaining the computational efficiency of the forward alternative [14, 15].

In this paper, heuristic optimization methods, such as the Particle Swarm Optimization, Differential Evolution and the recently developed Extreme Learning Machine [16, 17] are effectively integrated with our two-stage selection (TSS) algorithm, leading to a flexible and efficient construction scheme for RBF neural modelling. Non-linear parameters in the hidden layer will be continuously optimized at each step while the output layer weights are estimated by the least-squares. The main advantages over the conventional subset selection methods are that both the centre vector and width vector in a RBF function are optimized separately for each hidden node. Thus the model size can be significantly reduced, leading to improved generalisation performance. The experimental results on the simulation data illustrate the compactness and performances of the obtained model. It also shows that the PSO and DE have similar capabilities in non-linear parameter optimization while the extreme learning machine provides an efficient alternative with significantly reduced computational effort.

The rest of this paper is organized as follows. Section II and III give an overview of the ISBM process and the heuristic methods employed. The new two-stage selection algorithm is then introduced in Section IV with the experimental results given in V. Finally, VI concludes the paper and provides with suggested future work.

## II. The ISBM process

Practically, it is difficult to generate enough data for system identification. Thus a commercial finite element (FE) package known as ABAQUS is used to simulate the ISBM process. Several conditions need to be pre-set before the simulation process. These include the geometry of the preform, stretch rod and mould, material model, pre-blow process conditions (mass flow rate, pre-blow air pressure, preform temperature) and final-blow process conditions (final-blow air pressure and cooling time).

The pre-blow process conditions have a major influence on the final thickness and mechanical properties distributions in the ISBM process. At this stage, the simulations are only focused on modelling wall thickness distribution at different pre-blow process conditions. The preform reheating stage and cooling stage are beyond the scope of this study. Four parameters of the process were identified to be important, namely mass flow rate, pressure, temperature, and timing. The mass flow rate indicates the velocity of air blowing to the pre-form while the pressure affects the total amount of air consumed. The pre-form temperature was assumed uniform and equivalent to the setting value in the simulation, and the timing reflects the delay between stretching and blowing. Each of the four parameters were given different settings as shown in table I, and totally 144 possible combinations were obtained. Meanwhile the bottle wall volume is divided into three parts: shoulder, side-wall and base. These volume distribution indicators can generally represent the weight distribution. The frictional force between mould and bottle was infinite during the simulation.

TABLE I
DOE settings of each parameter in ISBM simulation

| Mass flow rate(g/s) | Pressure(MPa) | Temperature(°C) | Timing(ms) |
|---|---|---|---|
| 5 | 0.6 | 95 | 0 |
| 17 | 0.8 | 100 | 50 |
| 29 | 1 | 105 | 100 |
| 40 |  | 110 |  |

## III. Review of RBF network model and heuristic optimization methods

### A. Radial Basis function network model

A general RBF neural model can be expressed as

$$y(t) = \sum_{k=1}^{n} \theta_k \varphi_k(\mathbf{x}(t); \mathbf{c}_k; \mathbf{\Sigma}_k) + \varepsilon(t) \tag{1}$$

where $y(t)$ is the actual output at sample time $t$, $\mathbf{x}(t) \in \Re^p$ is the input vector, $\varphi_k(\mathbf{x}(t); \mathbf{c}_k; \mathbf{\Sigma}_k)$ denotes the nonlinear activation function, $\mathbf{c}_i = [c_{i1}, c_{i2}, \cdots, c_{ip}]^T$ is the centre vector, and $\mathbf{\Sigma}_i$ is the associated norm matrix. Finally, $\theta_k$ represents the output layer weight for each RBF node, and $\varepsilon(t)$ is the network error at sample time $t$. By using a set of $N$ data samples $\{\mathbf{x}(t), y(t)\}_{t=1}^{N}$ for model training, (1) can then be re-written in a matrix form as

$$\mathbf{y} = \mathbf{\Phi}\boldsymbol{\theta} + \mathbf{e} \tag{2}$$

If the regression matrix $\mathbf{\Phi}$ is of full column rank, the Least-Squares estimate of the regression coefficients in (2) is given by

$$\hat{\boldsymbol{\theta}} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{y} \tag{3}$$

where $\mathbf{\Phi}^T\mathbf{\Phi}$ is sometimes called the information matrix. The associated minimal cost function is

$$J_n(\hat{\boldsymbol{\theta}}_n) = \mathbf{y}^T(\mathbf{I} - \mathbf{\Phi}_n)(\mathbf{\Phi}_n^T\mathbf{\Phi}_n)^{-1}\mathbf{\Phi}_n^T\mathbf{y} \tag{4}$$

### B. Particle Swarm Optimization

PSO involves a number of particles which move through the problem search space seeking an optimal or satisfactory solution. The position of each particle is adjusted according to its velocity and the difference between its current position, the best one it has found so far, and the best position to date found by its neighbours.

Suppose $\mathbf{x}_i$ denotes the $i^{th}$ particle in the swarm, $\mathbf{v}_i$ represents its velocity, $\mathbf{p}_i$ is its best position to date, while

$\mathbf{p}_g$ denotes the best position from the entire swarm. In inertia-weighted PSO, $\mathbf{v}_{(i+1)}$ and $\mathbf{x}_{(i+1)}$ are updated as:

$$\mathbf{v}_{(i+1)} \leftarrow w_0\mathbf{v}_i + c_1r_1(\mathbf{p}_i - \mathbf{x}_i) + c_2r_2(\mathbf{p}_g - \mathbf{x}_i) \quad (5)$$
$$\mathbf{x}_{(i+1)} \leftarrow \mathbf{x}_i + \mathbf{v}_i \quad (6)$$

where $w_0$ is the inertia weight used to scale the previous velocity term, $c_1$ and $c_2$ are acceleration coefficients, and $r_1$ and $r_2$ are two uniform random number generated between 0 and 1. The acceleration coefficients $c_1$ and $c_2$ can be fixed or varied during the iterative procedure. In order to ensure that each updated particle is still inside the search space, it is also necessary to define a value range, and check both the position and velocity for each particle at the end of an iteration.

### C. Deferential Evolution

As a population based optimization technique, DE [10] also starts with some initial points which are randomly generated in the search space, and then pushes the populations toward the global optimum point through repeated operations of mutation, crossover and selection. New populations are obtained by adding the weighted difference of two vectors to a third one, where the vectors are mutually different random points from the last generation.

Suppose $\mathbf{x}_i^{(l)}$ $(i = 1, 2, \cdots, p)$ is the solution vector in generation $l$, the operations in the classic DE method can be summarised as follows:

- **Mutation**: A mutant vector is generated by:
$$\mathbf{v}_i^{(l+1)} = \mathbf{x}_{r1}^{(l)} + F(\mathbf{x}_{r2}^{(l)} - \mathbf{x}_{r3}^{(l)}) \quad (7)$$
where $r1$, $r2$, $r3$ are random indices from $[1, 2, \cdots, p]$ and $F \in [0, 2]$ is a real constant which controls the amplification of the added differential variation. Larger values for $F$ lead to higher diversity in new populations, while lower values cause faster convergence.
- **Crossover**: This operation is implemented to increase the diversity of the population. A trial vector is defined as
$$\mathbf{u}_i^{(l+1)} = [u_{i1}^{(l+1)}, u_{i2}^{(l+1)}, \cdots, u_{ip}^{(l+1)}]^T \quad (8)$$
with elements given by
$$u_{ij}^{(l+1)} = \begin{cases} v_{ij}^{(l+1)} & rand_i(0,1) \leq C_r \ or \ i = b_r \\ x_{ij}^{(l)} & otherwise \end{cases} \quad (9)$$
where $p$ is the vector dimension, $C_r \in [0, 1]$ is the predefined crossover constant, $rand_i(0, 1)$ uniformly generates a scaler from $[0, 1]$ at the $i^{th}$ evaluation, and $b_r$ is a random index chosen from $[1, 2, \cdots, p]$ so that $\mathbf{u}_i^{(l+1)}$ contains at least one parameter from $\mathbf{v}_i^{(l+1)}$.
- **Selection**: The last step is to compare all the trial vectors $\mathbf{u}_i^{(l+1)}$ with the target ones $\mathbf{x}_i^{(l)}$ using a criterion, such as their contribution to a loss function, and then decide which one becomes a member of the next generation.

The above procedure continues until a pre-set number of iterations is reached or the desired accuracy is obtained.

### D. Extreme Learning Machine

The Extreme Learning Machine (ELM) was first introduced for the training of Single-hidden Layer Feedforward neural Network (SLFN). It builds the SLFN model by randomly assigning non-linear parameters for each hidden node instead of iterative training. The target is then simply a linear combination of the hidden nodes, and the output layer weights can be easily estimated by Least-Squares. As a result, the learning speed in ELM can be several orders of magnitude faster than traditional learning. Using incremental methods, it has been proven that the ELM can be regarded as a universal approximator [18].

The SLFN has a similar structure to a RBF network. For a multi-input, multi-output (MIMO) system, it can be expressed as

$$\mathbf{y}(t) = \sum_{k=1}^{n} \boldsymbol{\theta}_k \varphi_k(\mathbf{w}_k \cdot \mathbf{x}(t) + b_k) \quad (10)$$

where $\mathbf{x}(t) = [x_1(t), x_2(t), \cdots, x_p(t)]$ and $\mathbf{y}(t) = [y_1(t), y_2(t), \cdots, y_m(t)]$ is the system input and output vector; $\mathbf{w}_k = [w_{k1}, w_{k2}, \cdots, w_{kp}]$ is the weight vector between the $p$ inputs and the $k^{th}$ hidden node; $b_k$ is the threshold of the $k^{th}$ hidden node; $(\cdot)$ denotes the inner product, and $\varphi$ is the activation function. Finally $\boldsymbol{\theta}_k = [\theta_{k1}, \theta_{k2}, \cdots, \theta_{km}]$ is the output layer weight vector between the $k^{th}$ hidden node and $m$ outputs.

In ELM, the non-linear parameters $\mathbf{w}_k$ and $b_k$ in (10) are assigned randomly, and it has been proven that the the required number of hidden nodes $n \leq N$ if the activation function $\varphi$ is infinitely differentiable [16].

According to the above theorem, the ELM is also valid for the RBF network in (1). The construction process can be summarised in two steps:

1) Randomly assign the hidden nodes parameters, including the number of hidden nodes $n$, and non-linear parameters $\mathbf{c}_i$ and $\boldsymbol{\sigma}_i$ for $i = 1, 2, \cdots, n$;
2) Form the regression matrix $\boldsymbol{\Phi}$, and estimate the output layer weights using (3);

The main issue in using ELM is that the sparsity of the constructed model cannot be guaranteed due to its stochastic characteristics. Fortunately, subset selection methods can be easily integrated to solve this problem, and better generalization performance can be achieved.

## IV. Two-stage selection based on Heuristic optimization

The two-stage selection algorithm includes a forward model construction stage and a backward model refinement stage. At its first stage, one RBF centre is selected and added to the model at each step. The significance of each centre is measured by its contribution to the cost function. This process continues until some pre-defined modelling criteria are met (such as Akaike's information criterion or a pre-defined number of hidden nodes is reached), the algorithm then moves to the second stage where the importance of previously selected centres are reviewed, and any insignificant ones are replaced. The computational efficiency involved in such process is achieved by defining a residual matrix $\mathbf{R}_k$ which can be updated recursively. Moreover, PSO and DE are used to find the best centre at each step while ELM is used to form the centre pool before the selection process.

### A. First stage - forward selection

Refer to (4), a recursive matrix $\mathbf{M}_k$ and a residual matrix $\mathbf{R}_k$ are defined to simplify the computation.

$$\mathbf{M}_k \triangleq \mathbf{P}_k^T \mathbf{P}_k \qquad k = 1, \cdots, n \tag{11}$$

$$\mathbf{R}_k \triangleq I - \mathbf{P}_k \mathbf{M}_k^{-1} \mathbf{P}_k^T \qquad \mathbf{R}_0 \triangleq I \tag{12}$$

where $\mathbf{P}_k^T \in \Re^{N \times k}$ contains the first $k$ columns of the regression matrix $\mathbf{\Phi}$ in (2).

By substituting (12) into (4), the cost function becomes

$$J(\mathbf{P}_k) = \mathbf{y}^T \mathbf{R}_k \mathbf{y} \tag{13}$$

At this forward stage, the RBF centres are optimized one at a time, and given by the best solution from the entire population of the heuristic methods after several iterations. Suppose at the $k^{th}$ step, one more centre $\mathbf{p}_{k+1}$ is to be added. The net contribution of $\mathbf{p}_{k+1}$ to the cost function can then be calculated as [14]:

$$\Delta J_{k+1}(\mathbf{P}_k, \mathbf{p}_{k+1}) = \frac{(\mathbf{y}^T \mathbf{p}_{k+1}^{(k)})^2}{\mathbf{p}_{k+1}^T \mathbf{p}_{k+1}^{(k)}} \tag{14}$$

where $\mathbf{p}_{k+1}^{(k)} \triangleq \mathbf{R}_k \mathbf{p}_{k+1}$. According to [14], an auxiliary matrix $\mathbf{A} \in \Re^{n \times n}$ and a vector $\mathbf{b} \in \Re^{n \times 1}$ need to be defined to reduce the computational complexity. Their elements are given by:

$$a_{i,j} \triangleq (\mathbf{p}_i^{(i-1)})^T \mathbf{p}_j, \quad 1 \le i \le j \tag{15}$$

$$b_j \triangleq (\mathbf{p}_j^{(j-1)})^T \mathbf{y}, \quad 1 \le j \le n \tag{16}$$

where $(\mathbf{p}_j^{(0)} = \mathbf{p}_j)$. The efficiency of the this forward stage then follows from updating these terms recursively as:

$$a_{i,j} = \mathbf{p}_i^T \mathbf{p}_j - \sum_{l=1}^{i-1} a_{l,i} a_{l,j} / a_{l,l} \tag{17}$$

$$b_j = \mathbf{p}_j^T \mathbf{y} - \sum_{l=1}^{j-1} (a_{l,j} b_l) / a_{l,l} \tag{18}$$

By substituting (15) and (16) into (14), the net contribution of a new RBF centre $\mathbf{p}_{k+1}$ to the cost function can then be expressed as:

$$\Delta J_{k+1}(\mathbf{p}_{k+1}) = \frac{b_{k+1}^2}{a_{k+1,k+1}} \tag{19}$$

In heuristic methods, (19) provides the formula to evaluate each solution in the population. For instance, the selection of local best and global best particles in PSO or the selection between trial vector and target vector in DE.

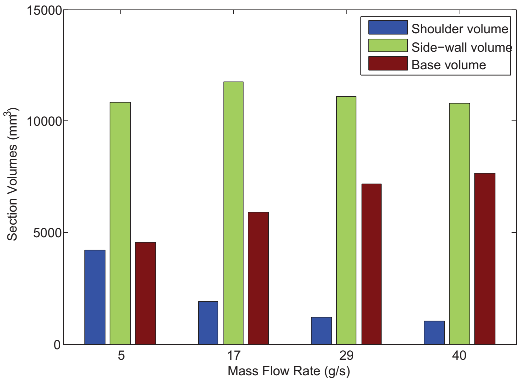


Fig. 1. Effect of mass flow rate on the section volumes (The other three parameters were kept constant, and set as Pressure = 10bar, Temperature = 100°C, Timing = 50ms

### B. Second stage - backward network refinement

The model from the first stage is not optimal due to the correlations between selected terms, that is the prior selected centres introduced a constraint on the latter selections. This second stage is therefore adopted to eliminate such constraint and replace any insignificant centres by new one generated from the population. Clearly, the last selected centre in the forward construction has always been maximally optimized for the whole network, the backward refinement can be divided into two main steps: Firstly, a selected centre $\mathbf{p}_k$, $k = 1, \cdots, n-1$ is shifted to the $n^{th}$ position as if it were the last selected one; then the optimization is implemented to find an alternative centre at the $n^{th}$ position based on the re-ordered $n-1$ centres. If the shifted one is less significant than the new centre from the population it will be replaced, leading to a reduced training error and potential improvement in the generalisation capability. This review is repeated until a pre-defined number of check loops is reached. The detailed algorithm and its computational analysis can be found in [13].

## V. Experimental results

According to table I, a total number of 144 experiments need to be carried out to generate the required data. However, running these experiments and measuring the section weights are practically difficult. Therefore, a manufacturing simulation software, known as ABAQUS/standard version 6.10 [2], was adopted as a substitute of the real process. Each simulation lasted around 20 minutes, and when all finished, four experiments were regarded as failure with their associated results being removed from the final data set. For the remaining 140 data points, 100 were used for RBF model training, and another 40 were reserved for model validation. The effect of mass flow rate is illustrated in Fig. 1 while other three parameters were kept constant.

The two-stage selection algorithm integrated with PSO, DE, and ELM was then applied on the training data. For heuristic optimization methods, their algorithm parameter settings are shown in table II. From the implementation, it is found that

TABLE II
PARAMETER SETTINGS FOR THE HEURISTIC APPROACHES (THE INITIAL CENTRE VECTORS WERE RANDOMLY SELECTED FROM THE TRAINING DATA)

| Method | Parameter | Value | Description |
|---|---|---|---|
| PSO | $\mathbf{x}_i$ | $[c_i; \sigma_i]$ | $i^{th}$ particle in the swarm |
| | $\sigma_i$ | $\in [0.1, 4]$ | Range of the width of $i^{th}$ RBF centre |
| | $S$ | 20 | Number of particles |
| | $G$ | 30 | Number of updating cycles |
| | $w_0$ | 0.8 | Inertia weight in velocity updating |
| DE | $\mathbf{x}_i$ | $[c_i; \sigma_i]$ | $i^{th}$ solution vector in the population |
| | $\sigma_i$ | $\in [0.1, 4]$ | Width of $i^{th}$ RBF centre randomly generated from the specified range; if $\sigma_i < 0$ in the mutation step, $\lvert\sigma_i\rvert$ is used |
| | $S$ | 10 | Population size |
| | $G$ | 20 | Maximum number of generations |
| | $F$ | 0.8 | Weight of vector difference |
| | $C_r$ | 0.6 | Crossover constant |
| ELM | $\sigma_i$ | $\in [0.1, 4]$ | Width of $i^{th}$ RBF centre randomly generated from the specific range; |

TABLE III
COMPARISON OF BOTTLE SHOULDER VOLUME MODELLING PERFORMANCES (RMSE)

| Algorithm | Model size | Training error | Test error |
|---|---|---|---|
| PSO+FRA | 4 | 0.2395 | 0.2770 |
| PSO+TSS | 4 | 0.1687 | 0.1998 |
| DE+TSS | 4 | 0.1780 | **0.1869** |
| ELM+TSS | 4 | 0.2531 | 0.2784 |
| | | | |
| PSO+FRA | 5 | 0.2333 | 0.2319 |
| PSO+TSS | 5 | 0.1441 | 0.1575 |
| DE+TSS | 5 | 0.1448 | **0.1216** |
| ELM+TSS | 5 | 0.2354 | 0.2408 |
| | | | |
| PSO+FRA | 6 | 0.1791 | 0.1972 |
| PSO+TSS | 6 | 0.1297 | **0.1272** |
| DE+TSS | 6 | 0.1089 | 0.1293 |
| ELM+TSS | 6 | 0.1896 | 0.2261 |

the increase of swarm size normally affects the performance more than the increase of updating cycles in PSO, while in differential evolution, these two control parameters have the similar effects.

The resultant RBF models are usually evaluated by the Root Mean Squared Error (RMSE) on validation data, table III - V illustrate such performances under different algorithms. Due to the stochastic characteristics in heuristic optimization, the best results from five runs are chosen for comparison. The fast recursive algorithm (first stage of TSS) usually gives worse results than the TSS alternative, so it was only tested in the modelling of bottle shoulder volume. The results also indicate that DE works better than PSO in the shoulder volume modelling. However, PSO outperforms DE in the other two section volume modelling. Extreme learning machine is definitely the most efficient methods which runs much faster than PSO or DE, however the corresponding model performances are less favourable. The prediction of bottle base volume is also shown in Fig. 2, while the associated RBF model parameters are given in table VI.

TABLE IV
COMPARISON OF BOTTLE SIDE WALL VOLUME MODELLING PERFORMANCES (RMSE)

| Algorithm | Model size | Training error | Test error |
|---|---|---|---|
| PSO+TSS | 4 | 0.4084 | **0.4023** |
| DE+TSS | 4 | 0.4201 | 0.4531 |
| ELM+TSS | 4 | 0.4506 | 0.5505 |
| | | | |
| PSO+TSS | 5 | 0.3761 | **0.3761** |
| DE+TSS | 5 | 0.3481 | 0.3935 |
| ELM+TSS | 5 | 0.4194 | 0.4996 |
| | | | |
| PSO+TSS | 6 | 0.3523 | 0.4926 |
| DE+TSS | 6 | 0.3379 | **0.4049** |
| ELM+TSS | 6 | 0.4156 | 0.4425 |

TABLE V
COMPARISON OF BOTTLE BASE VOLUME MODELLING PERFORMANCES (RMSE)

| Algorithm | Model size | Training error | Test error |
|---|---|---|---|
| PSO+TSS | 4 | 0.1300 | **0.1281** |
| DE+TSS | 4 | 0.1518 | 0.1508 |
| ELM+TSS | 4 | 0.2500 | 0.2959 |
| | | | |
| PSO+TSS | 5 | 0.1148 | **0.1021** |
| DE+TSS | 5 | 0.1363 | 0.1260 |
| ELM+TSS | 5 | 0.1907 | 0.2215 |
| | | | |
| PSO+TSS | 6 | 0.0898 | **0.0938** |
| DE+TSS | 6 | 0.1054 | 0.0990 |
| ELM+TSS | 6 | 0.1834 | 0.2104 |

TABLE VI
OPTIMIZED PARAMETERS OF RBF MODEL FOR THE BOTTLE BASE VOLUME PREDICTION (REFER TO (1))

| Parameters | Optimized values |
|---|---|
| $\theta_1$ | 1.5419 |
| $\theta_2$ | -2.6327 |
| $\theta_3$ | 2.5896 |
| $\theta_4$ | -3.0357 |
| $\theta_1$ | 3.0150 |
| $\boldsymbol{c}_1$ | [-0.051, 0.244, 1.635, 1.499] |
| $\boldsymbol{\sigma}_1$ | [0.100, 4.000, 4.000, 0.100] |
| $\boldsymbol{c}_2$ | [-1.605, 1.475, -0.112, 1.499] |
| $\boldsymbol{\sigma}_1$ | [0.244, 4.000, 4.000, 4.000] |
| $\boldsymbol{c}_3$ | [1.575, 1.475, 1.635, 0.451] |
| $\boldsymbol{\sigma}_3$ | [0.487, 4.000, 2.986, 0.276] |
| $\boldsymbol{c}_4$ | [0.946, -1.475, -1.574, 1.499] |
| $\boldsymbol{\sigma}_4$ | [4.000, 4.000, 2.008, 4.000] |
| $\boldsymbol{c}_5$ | [1.185, 0.632, -0.332, 1.499] |
| $\boldsymbol{\sigma}_5$ | [2.966, 4.000, 2.454, 1.030] |

## VI. CONCLUSION AND FUTURE WORK

Injection stretch-blow moulding is a typical process to produce plastic bottles in the industry. The process is usually controlled by several adjustable parameters, such as the mass flow rate, blow pressure and preform temperature. The quality of resultant bottles can be measured by its top loads, or the section weights distribution. Due to the lack of closed-loop control, large variations can be observed on the quality indicators.

Section weights are found to be the most prospective variable for feed-back control, but cannot be measured directly in a typical process. Thus, a soft-sensor approach based on

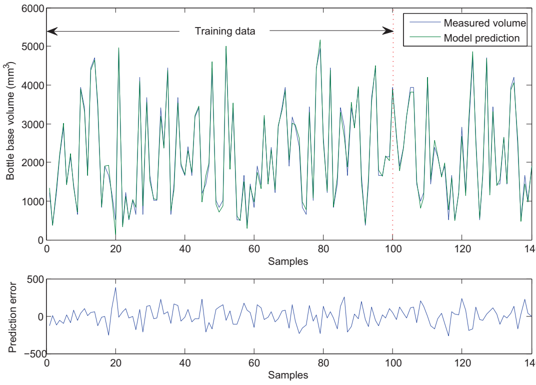


Fig. 2. The bottle base volume model prediction by TSS+PSO (5 hidden nodes were used, the first 100 samples show the training performance while the rest 40 points are model prediction)

an inferential mathematical model becomes an affordable alternative for control implementation. This paper uses a radial basis function network model to predict the bottle section weights. The main issue involved in RBF model construction is the determination of non-liner parameters in the hidden nodes. As the gradient-based approaches require large computational effort, heuristic optimization methods, such as particle swarm optimization, deferential evolution, and extreme learning machine become appropriate alternatives. In this paper, these heuristic optimization methods are effectively integrated with our recently proposed two-stage subset selection algorithm, leading to an efficient RBF model construction algorithm. Experimental results on simulation data has successfully verified the effectiveness of the proposed method. Future work will use practical data to build the models and implement iterative learning control for the ISBM process based on the soft-sensors.

## Acknowledgment

This work was financially supported by Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/G059489/1 and partly supported by EPSRC grant EP/F021070/1, EP/H020756/1, and EP/G042594/1, and Science and Technology Commission of Shanghai Municipality (11ZR1413100).

## References

[1] C. Abeykoon, K. Li, M. McAfee, P. J. Martin, and G. W. Irwin, "Extruder melt temperature control with fuzzy logic," in *Proceedings of the 18th IFAC World Congress*, 2011, pp. 8577–8582.

[2] Z. Yang, E. Harkin-Jones, G. Menary, and C. Armstrong, "A non-isothermal finite element model for injection stretch-blow molding of pet bottles with parametric studies," *Polymer Engineering & Science*, vol. 44, no. 7, pp. 1379–1390, 2004.

[3] G. Menary, C. Tan, C. Armstrong, Y. Salomeia, M. Picard, N. Billon, and E. Harkin-Jones, "Validating injection stretch-blow molding simulation through free blow trials," *Polymer Engineering & Science*, vol. 50, no. 5, pp. 1047–1057, 2010.

[4] S. Chen and S. Billings, "Neural networks for nonlinear dynamic system modelling and identification," *International Journal of Control*, vol. 56, no. 2, pp. 319–346, 1992.

[5] S. McLoone, M. Brown, G. Irwin, and G. Lightbody, "A hybrid linear/nonlinear training algorithm for feedforward neural networks," *IEEE Transactions on Neural Networks*, vol. 9, no. 4, pp. 669–684, 1998.

[6] S. Kirkpatrick, "Optimization by simulated annealing: Quantitative studies," *Journal of Statistical Physics*, vol. 34, no. 5, pp. 975–986, 1984.

[7] Z. Michalewicz, *Genetic algorithms + data structures = evolution programs*. Springer, 1996.

[8] F. Glover and R. Marti, "Tabu search," *Metaheuristic Procedures for Training Neutral Networks*, vol. 36, pp. 53–69, 2006.

[9] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proc. IEEE International Conference on Neural Networks*, vol. 4, Perth, Australia, 1995, pp. 1942–1948.

[10] R. Storn and K. Price, "Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces," *Journal of global optimization*, vol. 11, no. 4, pp. 341–359, 1997.

[11] M. Dorigo and L. Gambardella, "Ant colony system: A cooperative learning approach to the traveling salesman problem," *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 53–66, 2002.

[12] G. Loganathan, "A New Heuristic Optimization Algorithm: Harmony Search," *SIMULATION*, vol. 76, no. 2, pp. 60–68, 2001.

[13] J. Deng, K. Li, G. Irwin, and M. Fei, "Two-stage RBF network construction based on PSO," *Transactions of the Institute of Measurement and Control*, vol. 0, no. 0, pp. 1–9, 2011.

[14] K. Li, J. Peng, and G. Irwin, "A fast nonlinear model identification method," *IEEE Transactions on Automatic Control*, vol. 50, no. 8, pp. 1211–1216, 2005.

[15] J. Deng, K. Li, and G. Irwin, "Locally regularised two-stage learning algorithm for RBF network centre selection," *International Journal of Systems Science*, no. 1, pp. 1–14, 2011.

[16] G. Huang, Q. Zhu, and C. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, pp. 489–501, 2006.

[17] J. Deng, K. Li, and G. Irwin, "Fast automatic two-stage nonlinear model identification based on the extreme learning machine," *Neurocomputing*, vol. 74, no. 16, pp. 2422–2429, 2011.

[18] G. Huang, L. Chen, and C. Siew, "Universal approximation using incremental constructive feedforward networks with random hidden nodes," *IEEE Transactions on Neural Networks*, vol. 17, no. 4, pp. 879–892, 2006.

# How to Reduce Congestion on TCP/AQM Networks with Simple Adaptive PID Controllers

Teresa Alvarez, Anuar Salim
Department of Engineering Science and Automatic Control,
Escuela de Ingenierías Industriales (Sede Doctor Mergelina), University of Valladolid
Universidad de Valladolid, Valladolid, Spain
e-mail: tere@autom.uva.es

*Abstract*—**Congestion is a problem in real networks. Users do not want to lose information and data should be delivered as fast and reliably as possible. This is really difficult to achieve. Moreover, networks work in changing environments: number of users, type of traffic, delays in transmission, etc. So this paper presents how to design PID controllers that take network changes into account. Non-linear simulations using ns-2 will show the goodness of the approach when compared with classical PID, drop tail and RED.**

***Keywords-component; congestion control; PID; adaptive control; robustness; TCP/AQM***

## I. INTRODUCTION

Internet is very important in our society. Information, images, videos and data are transmitted from one place to another in a matter of seconds. Nevertheless, there are problems in transmission that affect the quality of the process. Congestion is one of the most annoying obstacles ([1], [2]). It is important to detect this and to act as fast as possible to solve it. Thus, techniques to reduce congestion are of great interest. There are two basic approaches ([3], [4]): c*ongestion control,* which is used after the network is overloaded, and c*ongestion avoidance,* which takes action before the problem appears.

Feedback control can help to solve congestion control. The end-to-end transmission control protocol (TCP), and the active queue management (AQM) scheme define the two parts implemented at the routers' transport layer where congestion control is carried out. The AQM objectives ([5]) are: to minimize the occurrences of queue overflow and underflow, to minimize the time required for a data packet to be serviced by the routing queue and to maintain closed-loop performance in spite of changing conditions (robustness).

AQM schemes enhance the performance of TCP, although they do not work perfectly in every traffic situation. Numerous algorithms have been proposed (a good survey can be found in [4]). RED [6] is the basic algorithm used for comparison. It can detect and respond to long-term traffic patterns, but it cannot detect congestion caused by short-term traffic load changes. The most widely used AQM mathematical models were published in [4]. Since then, several control approaches have been applied, such as fuzzy, predictive control or robust control. The automatic control techniques that have received most attention are the PI controller, followed by the PID ([5]).

AQM techniques should be robust and give good results when the network is not operating in a nominal situation, when the delays are changing, or the number of users or the intensity of the packets' flow vary. So the AQM congestion control algorithm should be robust, stable and do better in a changing environment. If PID is chosen as the AQM algorithm, there are many useful references in the literature ([7]-[9]). These works do not consider what happens with the tuning when the traffic conditions change. They explicitly consider delays in the design of the PIDs and this is the situation that arises when working with networks: a system with delay.

Motivated by these issues, this paper presents how to derive PIDs, with guaranteed stability and robustness, which perform well in a network with changing traffic conditions. The work presented in this paper applies the generic method by [7] for finding stable PID controllers for time delay systems to the dynamic TCP/AQM router model. A linear gain scheduling is included in the design (reducing the gain variation) to enhance the robustness and ensure an adequate behaviour in extreme working scenarios. A similar approach was described in [10], but further study has shown that it can be simplified, as shown in this paper. The metrics applied to study the controller performance are: the router queue size, the link utilization and the probability of packet losses.

Non-linear simulations with ns-2 will show the performance of the technique by applying it to a problem of two routers connected in a Dumbbell topology, which represents a single bottleneck scenario. A comparison between the new PID, a classic PID and the RED approach is performed.

This paper is organized as follows. Section II briefly describes the TCP NewReno modeling. Section III describes a fluid model for the system. Next, the gain scheduling PID, classic PID and RED are explained and their methodology is described. Simulation results are shown in section V. Finally, some conclusions are presented.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

## II. TCP/IP Network Modelling

### A. TCP/IP NewReno dynamic model

Now, the TCP/IP NewReno network dynamics is presented. The dynamics of an AQM router are complex due to the number of variables that come into play: packet sources, protocols, etc. Nevertheless, it is possible to obtain a nonlinear model that represents the dynamics of the system ([4]), considering that the protocol used is TCP. The model relates to the average value of the network variables and is described by the following coupled, nonlinear differential equations:

$$\dot{W}(t) = \frac{1}{R(t)} - \frac{W(t)}{2}\frac{W(t-R(t-R(t)))}{R(t-R(t))}p(t-R(t))$$

$$\dot{q}(t) = \begin{cases} -C + \dfrac{N_{TCP}(t)}{R(t)}W(t), & q > 0 \\ \max\left\{0,\ \ -C + \dfrac{N_{TCP}(t)}{R(t)}W(t)\right\}, & q = 0 \end{cases},$$

(1)

where

$W$: average TCP window size (packets),

$\dot{q}$ : average queue length (packets),

$R$: round-trip time = $q/C+T_p$ (secs),

$C$: link capacity (packets/sec),

$T_p$: propagation delay (secs),

$N_{TCP}$: load factor (number of TCP sessions) and

$p$: probability of packet mark.

As explained in [4], the first differential equation in (1) describes the TCP window control dynamic, and the second equation models the bottleneck queue length, as an accumulated difference between the packet arrival rate and the link capacity. The queue length and window size are positive, bounded quantities, i.e., $q \in [0,\bar{q}]$ and $W \in [0,\overline{W}]$, where $\bar{q}$ and $\overline{W}$ denote buffer capacity and maximum window size, respectively. In this formulation, the congestion window size $W(t)$ is increased by one every round-trip time if no congestion is detected, and is halved when congestion is detected.

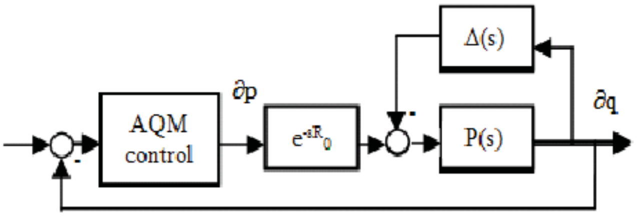


Fig. 1. Block diagram of AQM as a feedback control system

### B. Linearized model

Although an AQM router is a non-linear system, in order to analyze certain types of properties and design controllers, a linearized model is used. To linearize (1), it is assumed that the number of active TCP sessions and the link capacity are constant, i.e., $N_{TCP}(t)= N$ and $C(t)=C$.

The dependence of the time delay argument $t-R$ on queue length $q$ is ignored ([4]) and it is assumed to be fixed at $t-R_0$. This is a big assumption and there will be situations where it may not be acceptable. In [11], the authors deduced that this is a good approximation when the round-trip time is dominated by the propagation delay. This occurs when the capacity $C$ is large. $R_0$ should be chosen such that the above hypothesis can be ensured, so a value that works in the worst case scenario should be advisable. When the model is linearized, the same supposition is made. It cannot be denied that calculations are significantly simplified, but further research in the matter would be advisable. Local linearization of (1) around the operating point results in the following differential equations:

$$\left.\begin{aligned} \partial\dot{W}(t) &= -\frac{N}{R_0^2 C}(\partial W(t) - \partial W(t-R_0)) \\ &\quad - \frac{1}{R_0^2 C}(q(t) - \partial q(t-R_0)) - \frac{R_0^2 C}{2N^2}\partial p(t-R_0) \\ \partial\dot{q}(t) &= \frac{N}{R_0}\partial W(t) - \frac{1}{R_0}\partial q(t) \end{aligned}\right\}, \quad (2)$$

where $\partial\dot{W}(t) = W - W_0$ , $\partial q = q - q_0$ and $\partial p = p - p_0$ , represent the perturbed variables. The operating point for a desired equilibrium queue length $q_0$ is given by:

$$R_0 = \frac{q_0}{C} + T_p\,,\ W_0 = \frac{R_0 C}{N_{TCP}} \text{ and } p_0 = \frac{2}{W_0^2}. \qquad (3)$$

The linearized model (2) can be rewritten by separating the low frequency ('nominal') behaviour $P(s)$ of the window dynamic from the high frequency behaviour Δ(s) which is considered as parasitic.

$$P(s) = \frac{C^2/(2N_{TCP})}{\left(s + (2N_{TCP})/(R_0^2 C)\right)\left(s + 1/R_0\right)},$$

$$\Delta(s) = \frac{2N_{TCP}{}^2}{R_0 C^3}\left(1 - e^{-R_0 s}\right)$$

(4)

Taking (4) as a starting point, Hollot et al. (2002) give a feedback control description of AQM (Fig. 1). The action implemented by an AQM control law is to mark packets with a discard probability $p(t)$, as a function of the measured queue length $q(t)$. The larger the queue, the greater the discard probability becomes.

## III. Design of AQM Controllers

This section introduces the control formulation of an AQM router and how RED, classic PID and the gain scheduling PID can be applied. The latter will be described in depth, as it is the approach that outperforms the other two.

### A. AQM as feedback control

Taking (4) as the starting point, [5] gives a feedback control system of AQM (Figure 1). The action of an AQM control law is to mark packets with probability *p*, as a function of the measured queue length *q*. Following (4), the transfer function Δ(s) denotes the high-frequency window dynamics and *P(s)* (plant dynamics) relates how *p* dynamically affects *q*.

### B. AQM using RED

Random Early Detection (known as RED) was presented in [6]. A RED gateway calculates the average queue size, using a low-pass filter with an exponential weighted moving average. The average queue size is compared to two thresholds (minimum and maximum). When the average queue size is less than the minimum threshold, no packets are marked. When the average queue size is greater than the maximum threshold, every arriving packet is marked. If marked packets are in fact dropped, or if all source nodes are co-operative, this ensures that the average queue size does not significantly exceed the maximum threshold. When the average queue size is between the minimum and the maximum threshold, each arriving packet is marked with probability *p*, where *p* is a function of the measured queue length *q*.

RED does not work with a proper reference, set-point. As explained in [7], RED introduces a range of reference input values, rather than a proper set-point. We will use it as a classic AQM algorithm for the sake of comparison.

### C. AQM using PID control

PID controllers [5] are the most common form of feedback. They can be described by (5). This is the classic formulation.

$$\begin{aligned} u(t) &= K_p\left(e(t)+\frac{1}{T_i}\int_0^t e(\tau)d\tau + T_d\frac{de(t)}{dt}\right) \\ &= K_p e(t) + K_i\int_0^t e(\tau)d\tau + K_d\frac{de(t)}{dt} \end{aligned} \qquad (5)$$

### D. AQM using gain scheduling PID controller

This subsection presents the approach that has been followed to deal with the AQM congestion control problem in a network working with TCP Newreno (it should be noted that the approach can be applied to other TCP variants, [12]) under varying traffic. There are parameters, such as the link capacity *C*, which do not usually change, but the round trip delay, $R_0$, and the number of users, $N_{TCP}$, frequently vary.

The proposed technique is simple, but works well in networks with a wide range of users and varying delays. As network traffic changes constantly, this affects the performance of the system. We tune the PID for a certain set of conditions, but they change. The method proposed here will improve the router's performance no matter how many users or delay there may be in the network.

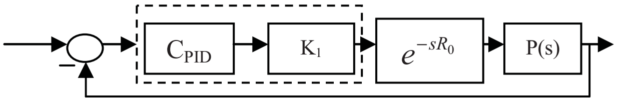


Figure 2: Block diagram of the AQM feedback control system

The PID is tuned following a model-based approach, taking as its starting point the linearized dynamic model of TCP described in the previous section. Figure 2 depicts the block diagram corresponding to the feedback control approach followed in the paper: P(s) is the transfer function obtained in the previous section. The delay $e^{-sR_0}$ is explicitly considered in the design. There are two elements in the controller C(s):

- $C_{PID}(s)$ is a PID controller and
- $K_1$ is a variable gain; a simple gain scheduling that allows the controller to work satisfactorily in a broad range of situations.

The steps in the design are:

- Choose the worst network scenario in terms of $N_{TCP}$ and $R_0$: i.e., the biggest delay and the smallest number of users.
- Choose the best network scenario in terms of $N_{TCP}$ and $R_0$: i.e., the smallest delay and the biggest number of users.
- Design a controller $C_{PID}(s)$ that is stable in these two situations and the scenarios in between.
- Add a simple gain scheduling $K_1$ to improve the system's response, based on the expected variations of the system's gain.

We have chosen the PID controller (6) as the basic AQM congestion control algorithm, as it is the most common form of feedback control technique. The PID should be tuned to be stable in the above mentioned situations. It is also required to have good responses in terms of settling time and overshoot. Taking this into account, the tuning of the controller is done following the method presented in [7]. It is relevant to note that the system can be of any order and the roots can be real or complex. Using the provided Matlab toolbox, a three dimensional region is obtained. If $K_p$, $K_i$ and $K_d$ are chosen inside this region, the closed-loop response is stable.

The simple gain scheduling approach that we propose takes into consideration the big gain variation that the AQM dynamic model has:

$$\frac{-C^2}{2N} \qquad (7)$$

Moreover, if a fair comparison among all the network configurations is to be done, the open-loop gain of the systems should be the same, so the independent term of the denominator of the transfer function should be included:

$$\frac{2N}{C\cdot R_0^3} \qquad (8)$$

This term greatly affects the size of the stable region and the magnitude of the PID's tuning parameters (as will be shown in next section). So we take out this term from P(s) and calculate the stable region for:

$$P_{normalized}(s)=P_n(s)=\frac{1}{s^2+a\cdot s+1},\qquad a=\frac{2N}{R_0^2C}+\frac{1}{R_0} \tag{9}$$

And:

$$\begin{aligned} P(s)&=\frac{-\dfrac{C^2}{2N}}{\dfrac{2N}{C\cdot R_0^3}}P_n(s)\\ &=-\frac{C^3R_0^3}{4N^2}P_n(s) \qquad (10)\\ &=P_{cte}\cdot P_n(s) \end{aligned}$$

So, the gain scheduling term is given by (11).

$$K_1=-\frac{4N^2}{C^3R_0^3} \tag{11}$$

It is important to choose the worst and best scenarios properly. A good knowledge of the network is required. Section IV defines the network parameters. The stable regions are depicted and the PID tuned. Then linear and non-linear simulations will show the goodness of the method.

## IV. Simulations

This Section presents how the new PID is tuned. A comparison between the proposed approach and a classical PID is shown using a linear simulation. Finally, ns-2 is used to test the new controller and to compare it with a classical PID and the AQM/RED congestion control technique.

### A. *Tuning the Controller*

The basic network topology, used as an example to test the controller, is depicted in Figure 3. It is a typical, single bottleneck topology. Although the network capacity C could change, its adjustments are more related with network updates than with everyday traffic conditions. The different scenarios come from changing the number of users (*N=N<sub>TCP</sub>*) and the round trip time (*R<sub>0</sub>*) (see Table I). For each of the values in Table 1, the link capacity takes two values: 1875 and 940 packets/sec.

TABLE I.

|  | N | $R_0$ | Results (Fig. 4) |
|---|---|---|---|
| Case 1 | 100 | 0.5 | Red |
| Case 2 | 150 | 0.4 | Green |
| Case 3 | 40 | 0.9 | Blue |
| Case 4 | 60 | 0.24 | Yellow |

The situations deal with a broad range of conditions: few users and big delay, many users and small delay and in-between situations. In [13] and [14], the effect of N and $R_0$ on the size of the controller's stability region was studied. They concluded (and our experiments also agree) that the smallest stability region is obtained for the smallest number of users (N) and the minimum round trip delay ($R_0$)

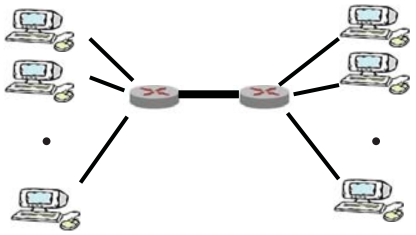


Figure 3: Dumbbell topology

Figure 4 shows the open loop poles of the system, without considering the delay. The lower graph depicts the open loop poles and zeros if a Padé approximation of first order is used for the delay of each of the scenarios.

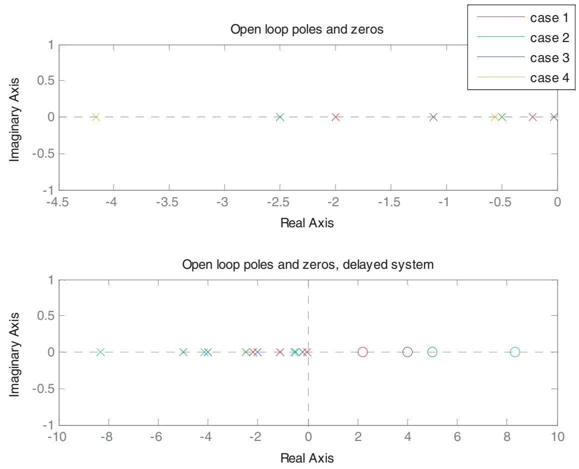


Figure 4: Open loop poles and zeros

The open loop step response for each of the studied scenarios of the original linear system (*P(s)*) and the normalized one (*P<sub>n</sub>(s)*) are shown in Figure 5 (upper and lower, respectively). Both systems have the same settling time, but the behaviour will be very different once the loop is closed. Case 3 gives the slowest open loop response (settling time: 150 sec.) and case 4 gives the fastest (settling time: 7.53 sec.).

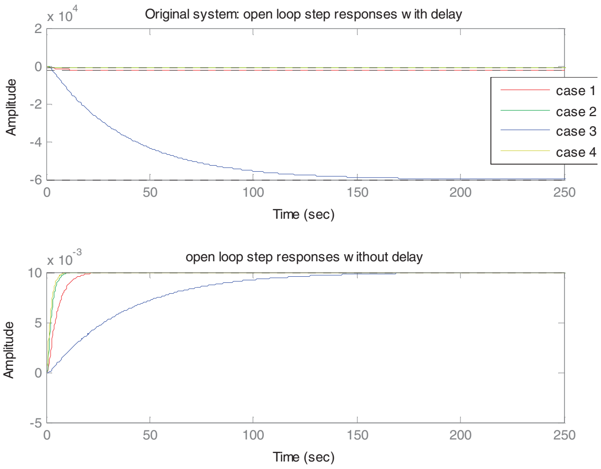


Figure 5: Step responses

TABLE II. PID PARAMETERS

| | Normalized | Standard |
|---|---|---|
| Kp | 1.2 | -0.000006 |
| Ki | 0.2 | -0.0000012 |
| Kd | 0.05 | -0.00000432 |

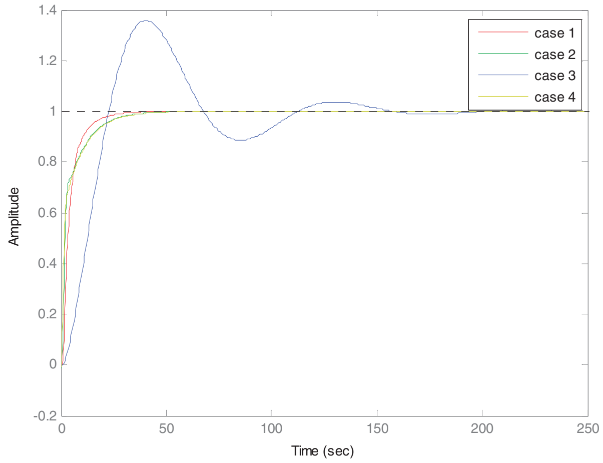


Figure 8: Closed-loop step response with proposed method

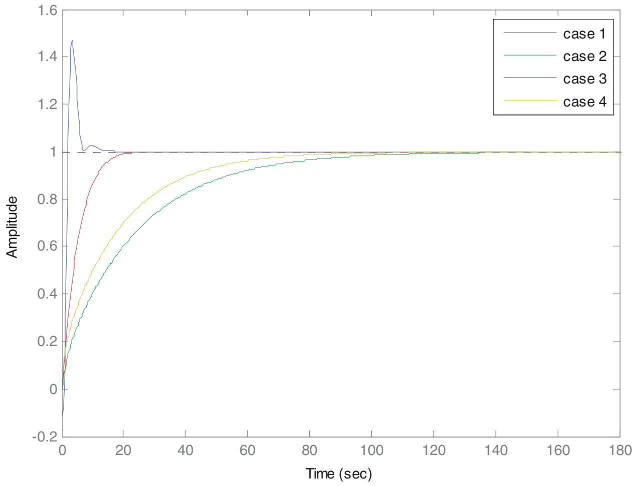


Figure 9: Closed-loop step response with classical approach

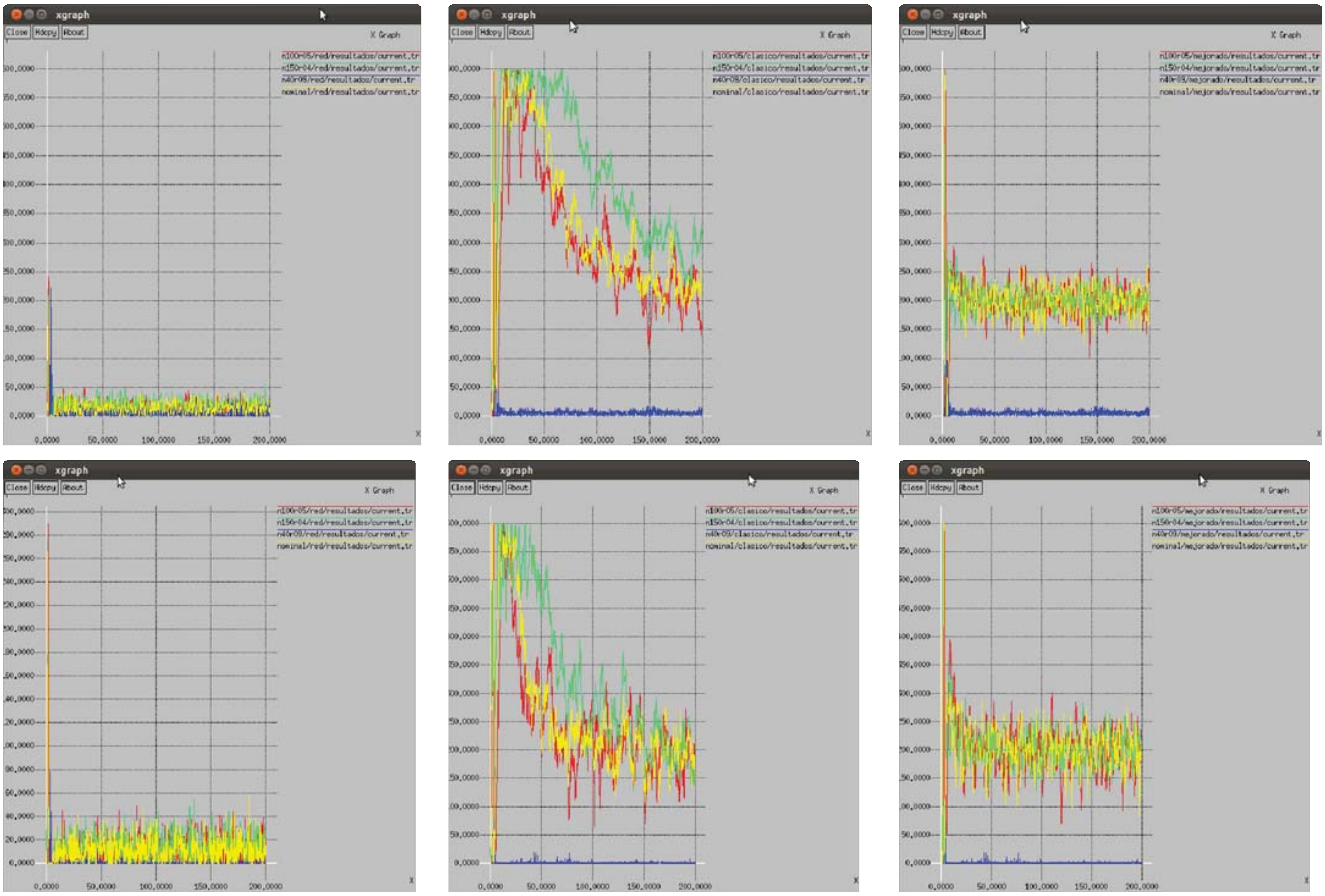


Figure 10: Non-linear simulations in ns-2

Using the approach in [7], a PID is tuned for both linear systems (Table II). The closed loop for these settings is stable for both approaches: the classical and the one presented in the paper. This system has negative gain. With the new PID, the sign is included in the gain scheduling, but the classic PID should be taken into account.

### B. Linear Simulations

Figures 8 and 9 show the closed loop step response of the normalized and classical systems. The proposed PID greatly improves the response when the network deals with cases 1, 2 and 4. Case 3 results are slower than when the classical approach is used.

### C. Non-linear simulations

This sub-section shows a comparison between the PID proposed in the paper, a classical PID and RED. Table I summarizes the different network parameters that have been used. The PIDs have been tuned as in Table II. RED minimum and maximum thresholds are set to 180 and 220, respectively.

Figure 10 depicts the results of the non-linear simulations that have been carried out using the very well known ns-2. The results with the link capacity C set to 1875 packet/sec are shown in the first row of Figure 10. The second row depicts the results when C=940 packets/sec. In both cases the reference is set to 200 packets, i.e., the queue at the router should take this value. The simulation lasts 200 seconds. The first column illustrates the RED algorithm; the classic PID is in the second column and the results of the gain scheduling discussed in the paper correspond to the third column.

The experiment has been carried out for the four scenarios described at the beginning of this section. The RED algorithm cannot deal properly with the requirements. This is logical because this technique is not intended for following a reference. Packets are marked following the steps described in Section III.

The classic PID gives good results, but it is slow. If C=940 packets/sec., the queue does not reach the reference value at any of the scenarios until t=150 sec. This would be almost unacceptable in a real network. When, in case 3 (blue line), the link is not congested, no packets are queued at the router.

Finally, the gain scheduling PID gives the best results. The response is very fast and there is almost no difference (in terms of settling time or final value) for any of the scenarios. As in the classic PID approach, when N=40 and the delay is 0.9 sec., the link is not congested. Results are very promising and further tests would give us a better insight.

## V. Conclusion

This paper has presented a comparison between a PID with linear gain scheduling and normalized values, a classic PID and an AQM/RED algorithm using the non-linear simulation environment ns-2. The controller is tuned for the worst scenario and works properly in a wide range of situations.

The main goal was to show that when traffic network changes and the delay, the number of users or the link's capacity are not the nominal values, the controller gives good results. The proposed PID outperforms the other approaches, as linear and non-linear simulations confirm.

The technique presented in the paper gives more uniform results than the classical approach. It does not matter if the scenario is close to the nominal situation or if it is a worst case setting, the controller reacts adequately and the settling times are all of the same order, whereas the classic approach cannot deal well with extreme situations. The RED algorithm gives the worst results of the three.

Future work will include testing the technique in a network with more congested routers.

## References

[1] Azuma, T., T. Fujita, M. Fujita (2006). Congestion control for TCP/AQM networks using State Predictive Control. Electrical Engineering in Japan, 156, 1491-1496.

[2] Deng, X., S. Yi, G. Kesidis. C.R. Das (2003). A control theoretic approach for designing adaptive AQM schemes. GLOBECOM'03, 5, 2947 – 2951.

[3] Jacobson, V. (1988). Congestion avoidance and control. ACM SIGCOMM'88.

[4] Ryu, S., C. Rump, C. Qiao (2004). Advances in Active Queue Management (AQM) based TCP congestion control. Telecommunication Systems, 25, 317-351.

[5] Hollot, C.V., V. Misra, D. Towsley, W. Gong (2002). Analysis and Design of Controllers for AQM Routers Supporting TCP flows. IEEE Transactions on Automatic Control, 47, 945-959.

[6] Aström K.J. and T. Hägglund (2006), Advanced PID control. ISA. NC.

[7] S. Floyd &V. Jacobson, Random early detection gateways for congestion avoidance. *IEEE/ACM Transactions on Networking*, *1*(4), 1993, 397-413.

[8] Hohenbichler, N. (2009). All stabilizing PID controllers for time delay systems. Automatica, 45, 2678-2684.

[9] Silva, G., A. Datta and S. Bhattacharyya. PID controllers for time-delay systems. Birkhäuser, 2005.

[10] Long, G.E., B. Fang, J.S. Sun and Z.Q. Wang (2010). Novel Graphical Approach to Analyze the Stability of TCP/AQM Networks. Acta Automatica Sinica, 36, 314-321.

[11] Alvarez, T. (2010). "Practical Design of PID Controllers for TCP/AQM". UKACC Control 2010, Coventry, United Kingdom.

[12] Hollot, C.V. and Y. Chait (2001). Nonlinear stability analysis for a class of TCP/AQM networks". In Proceedings of the 40th IEEE Conference on Decision and Control, Orlando, USA.

[13] Alvarez, T. (2012). Design of PID Controllers for TCP/AQM Wireless Networks. WCE 2012, London.

[14] Al-Hammouri, A.T. , V. Liberatore, M.S. Branicky, and S.M. Phillips (2006b). Parameterizing PI congestion controllers. Proc. First International Workshop on Feedback Control Implementation and Design in Computing Systems and Networks, Vancouver, CANADA.

[15] Al-Hammouri, A.T., V. Liberatore, M.S. Branicky, and S.M. Phillips (2006a). Complete stability region characterization for PI-AQM. SIGBED Review, 3(2).

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Robust adaptive actuator failure compensation controller for systems with unknown time-varying state delays

Marzieh Kamali, Javad Askari, Farid Sheikholeslam
Department of Electrical and Computer Engineering
Isfahan University of Technology
Isfahan, Iran
E-mail: m.kamaliandani@ec.iut.ac.ir

Ali Khaki Sedigh
Department of Electrical Engineering
Khajeh Nasir Toosi University of Technology
Tehran, Iran
Email: sedigh@kntu.ac.ir

*Abstract*—**An output feedback model reference adaptive controller is developed for a class of linear systems with multiple unknown time-varying state delays and in the presence of actuator failures. The adaptive controller is designed based on SPR-Lyapunov approach and is robust with respect to multiple unknown time-varying plant delays and to an external disturbance with unknown bounds. Closed-loop system stability and asymptotic output tracking are proved using suitable Lyapunov-Krasovskii functional and Simulation results are provided to demonstrate the effectiveness of the proposed controller.**

***Keywords- Robust adaptive control, Output feedback, Actuator failure, State delay systems.***

## I. INTRODUCTION

Component failures occur in many practical systems and may cause performance deterioration and catastrophic accidents. There have been many studies in the literature on control of systems with component failures [1]-[11]. In these papers, different design methods including multiple model, switching and tuning designs, fault detection and diagnosis designs, robust control designs and adaptive designs are used. Compared to other approaches, the direct adaptive control approach has the key advantage that it can provide theoretically provable asymptotic tracking in addition to stability, in the presence of large parameter variation and uncertainties. Important results for direct adaptive control of systems with actuator failures exist in [8]-[11].

Delay phenomena are frequently encountered in mechanics, physics, applied mathematics, biology, economics and engineering systems. In the presence of time delay, the design of fault tolerant controller becomes more complex. Therefore, the problem of fault tolerant adaptive control of delay systems has received little attention. For example, in [12] a fault detection and accommodation method is considered for nonlinear state delay systems, based on an iterative design of an observer which monitors the variations of the system dynamics and the control signal is formed by treating component failures as uncertainties. In [13] and [14], state feedback controllers are developed within the framework of Linear Matrix Inequalities for a class of linear systems with time delay in control inputs and constant actuator failures of stuck-type. A direct state feedback adaptive control scheme is introduced in [15] for linear state delay systems with unknown plant dynamics and unknown constant stuck failures in actuators. The same problem is solved for decentralized systems in [16]. Based on a linear matrix inequality technique and an adaptive method, [17] suggests adaptive reliable controllers against loss of effectiveness unknown actuator failures, but with the assumption that the system parameters are known.

In a recent work [18], an output feedback adaptive controller is designed for state delay systems with unknown parameters and unknown constant failures of stuck-type. To the best knowledge of authors, it is the first output-feedback model reference adaptive controller for compensating actuator failures in time delay systems. In this study, a new controller structure is developed to have robustness with respect to an external bounded disturbance with unknown bounds in addition to multiple unknown time-varying plant delays.

## II. PROBLEM FORMULATION

In this section, the control problem is formulated. Consider a linear state delay plant described by

$$\begin{aligned}&\dot{x}(t)=Ax(t)+Bu(t)+\sum_{l=1}^{M}A_{dl}x(t-d_l(t))+b_f f(t),\\&x(\theta)=x_0,\ \ \theta\in[-d_{\max},0]\\&y(t)=cx(t)\end{aligned}\tag{1}$$

where $x(t)\in\Re^n$ is the state vector, $y(t)\in\Re$ is the plant output and $u(t)=[u_1,...,u_m]^T\in\Re^m$ is the input vector whose elements may fail during system operation. $f(t)\in\Re$ is the external disturbance with $|f(t)|\le f^*$ . The constant matrices $A\in\Re^{n\times n}$ , $A_{dl}\in\Re^{n\times n}$, $l=1,...,M$ and $B=[b_1,...,b_m]\in\Re^{n\times m}$ and the vectors $c\in\Re^{1\times n}$ and $b_f\in\Re^{n\times 1}$ are unknown. The time delays $d_l(t)$ are nonnegative differentiable functions, satisfying

$$0 \le d_l(t) \le d_{\max}, \qquad \dot{d}_l(t) \le \bar{d} \; < 1, \quad l = 1,...,M \tag{2}$$

where $d_{\max}$ and $\bar{d}$ are some unknown positive constants.

In this paper, one important type of actuator failure modeled as

$$u_j(t) = \bar{u}_j, \qquad j \in \{1,2,...,m\} \tag{3}$$

is considered, where the constant values $\bar{u}_j$ and the failure time instants $t_j$ are unknown. With this type of actuator failure, the input $u(t)$ is defined as

$$u(t) = v(t) + \sigma(\bar{u} - v(t)) \tag{4}$$

where $\bar{u} = [\bar{u}_1,...,\bar{u}_m]^T$,

$$\sigma = diag\{\sigma_1, \sigma_2,...,\sigma_m\}, \quad \sigma_i = \begin{cases} 1, & u_i(t) = \bar{u}_i \\ 0, & u_i(t) \ne \bar{u}_i \end{cases}$$

and $v(t)$ is an applied control input to be designed. For this important type of actuator failure it is a basic assumption that [8]:

(A1)- If the system parameters and actuator failures (up to $m-1$ failures) are known, the remaining actuators can still achieve a desired control objective.

The adaptive control objective is to determine an output feedback $v(t)$ for the plant (1) with unknown parameters and unknown actuator failures (3) such that despite the control error $u - v = \sigma(\bar{u} - v)$, all signals of the closed-loop system remain bounded and the plant output $y(t)$ follows the output $y_m(t)$ of a stable reference model with the transfer function

$$y_m(t) = W_m(s)r(t), \qquad W_m(s) = \frac{1}{D_m(s)} \tag{5}$$

asymptotically; i.e., $\lim_{t\to\infty} e(t) = \lim_{t\to\infty}(y(t) - y_m(t)) = 0$. In the above equation, $D_m(s)$ is a monic Hurwitz polynomial and $r(t)$ is the reference input which is assumed to be uniformly bounded and piecewise continuous.

For simplicity the "equal control" design

$$v_1(t) = v_2(t) = ... = v_m(t) \equiv v_o(t) \tag{6}$$

which assumes that the control inputs to all actuators are the same, is used. It is reasonable in many practical applications.

Consider the transfer function of the system without delay as

$$y(t) = \sum_{j=1}^{m} \frac{k_{pj} N_j(s)}{D(s)} u_j(t), \quad c(sI - A^{-1})b_j = k_{pj} \frac{N_j(s)}{D(s)} \tag{7}$$

where $k_{pj}$ is a scalar, $N_j(s)$ is a monic polynomial and $D(s)$ is a monic polynomial of degree $n$.

With the assumption that $p$ (up to $m-1$) actuators fail and from (4) and (6) the closed-loop system (7) can be expressed as

$$y(t) = W_o(s)v_o(t) + \bar{y}(t) \tag{8}$$

where

$$\bar{y}(t) = \sum_{j=j_1,...,j_p} \frac{k_{pj} N_j(s)}{D(s)} \bar{u}_j(t),$$

$$W_o(s) = \frac{\sum_{j \ne j_1,...,j_p} k_{pj} N_j(s)}{D(s)} \equiv \frac{k_p N_{uf}(s)}{D(s)} = c(sI - A)^{-1} b, \tag{9}$$

$$b = \sum_{j \ne j_1,...,j_p} b_j \;,$$

To achieve control objective, it is assumed that $W_o(s)$ satisfies the following assumptions:

(A2)- $N_{uf}(s)$ is stable for each number of failures.

(A3)- The high-frequency gain $k_p$ has the same sign for each number of failures and $sign[K_r^*] = sign[k_p]$.

(A4)- The relative degree of the transfer function $W_o(s)$ is $n^* = 1$ for each number of failures.

## III. Error Equation

In order to design the adaptive controller, a suitable error equation parameterization is obtained in this section. For this purpose, first consider the system without delay (7) and when the plant parameters and actuator failures are known. The controller structure is defined as

$$\begin{aligned} &v_o^*(t) = K_e^* y(t) + K_1^{*T} x_1(t) + K_2^{*T} x_2(t) + K_r^* r(t) + K_3^*, \\ &x_1(t) = H(s)v_o(t), \;\; x_2(t) = H(s)y(t), \\ &H(s) = \frac{\alpha(s)}{\Lambda(s)}, \end{aligned} \tag{10}$$

in which $\Lambda(s)$ is a monic Hurwitz polynomial of degree $n$, $K_1^* \in \Re^{n-1}, K_2^* \in \Re^{n-1}, K_r^* \in \Re$, $K_e^* \in \Re$ are the parameters of the standard controller structure for MRAC of systems without delay and $\alpha(s) = [1, s, \ldots s^{n-2}]^T$.

In order to drive the error equation, $y(t)$ from (8) is substituted in (10) and the closed-loop system is obtained as

$$\begin{aligned} y(t) = W_o(s)(1 - K_1^{*T} H(s) - K_2^{*T} H(s) W_o(s) \\ - K_e^* W_o(s))^{-1} \times [K_2^{*T} H(s)\bar{y} + K_e^* \bar{y} + K_r^* r + K_3^*] + \bar{y}(t). \end{aligned} \qquad (11)$$

Now suppose that there exist constant matrices $a_{dl}^*$ and $F$ of appropriate dimensions such that

$$A_d = b a_{dl}^{*T}, \quad b_f = bF.$$

Using directly the analysis in [18] and [19], the tracking error $e(t) = y(t) - y_m(t)$ can be written as

$$\begin{aligned} e(t) = \frac{k_p}{D_m(s)}[v_0(t) - K_1^{*T} x_1(t) - K_2^{*T} x_2(t) \\ - K_e^* y(t) - K_3^* + (1 - K_1^{*T} H(s)) F f(t) \\ + \sum_{l=1}^{M} a_{dl}^{*T} x(t - d_l) - K_1^{*T} H(s) \sum_{l=1}^{M} a_{dl}^{*T} x(t - d_l)]. \end{aligned} \qquad (12)$$

To find a suitable error equation parameterization the dynamic system

$$z(t) = \sum_{l=1}^{M} K_1^{*T} H(s)[a_{dl}^{*T} x(t - d_l)] = \sum_{l=1}^{M} K_{zl}^{*T} z_{xl}(t) \qquad (13)$$

is defined in which,

$$\begin{aligned} K_{zl}^{*T} &= [K_1^{*1T} a_{dl}^{*T}, K_1^{*2T} a_{dl}^{*T}, \ldots, K_1^{*(n-1)T} a_{dl}^{*T}], \\ z_{xl}(t) &= H_n(s)[x(t - d_l)], \\ H_n(s) &= \frac{[I_{n\times n} s^{n-2}, \ldots, I_{n\times n} s, I_{n\times n}]}{\Lambda(s)} \in \Re^{n(n-1)\times n}. \end{aligned}$$

By decomposing $z_{xl}(t)$ into two components we have

$$\begin{aligned} z_{xl}(t) &= z_{el}(t) + z_{ml}(t), \\ z_{ml}(t) &= H_n(s)[x_{ml}(t - d_l)], \\ z_{el}(t) &= H_n(s)[e_x(t - d_l)], \\ e_x(t - d_l) &= x(t - d_l) - x_m(t - d_l), \end{aligned} \qquad (14)$$

Using (13) and (14), the error equation (12) can be rewritten as follows

$$\begin{aligned} e(t) = \frac{k_p}{D_m(s)}[v_0(t) - K^{*T} w(t)] - \frac{k_p}{D_m(s)}[K_m{}^{*T} w_m(t) \\ + \sum_{l=1}^{M} K_{dl}{}^{*T} e_x(t - d_l(t)) + \sum_{l=1}^{M} K_{zl}{}^{*T} z_{el}(t) \\ + (1 - K_1^{*T} H(s)) F f(t)] \end{aligned} \qquad (15)$$

where $K_{dl}^* = -a_{dl}^*$ and

$$\begin{aligned} K^* &= [K_e^*, K_1{}^{*T}, K_2{}^{*T}, K_r^*, K_3{}^*]^T, \\ w(t) &= [e, x_1{}^T, x_2{}^T, r, 1]^T, \\ K_m^* &= [K_{x_m}^{*T}, K_d{}^{*T}, K_z^{*T}]^T, \;\; K_{x_m}^* = c_m^T K_e^{*T} \\ w_m(t) &= [x_m^T(t), x_m{}^T(t - d(t)), z_m{}^T(t)]^T. \end{aligned}$$

## IV. Adaptive Controller design

For system with state delays and unknown parameters and actuator failures, the adaptive controller is designed in this section. Using the error equation (17) the controller structure

$$v_0(t) = K^T(t) w(t) - K_I \operatorname{sgn}(e(t)) \int_0^t |e(t)| dt, \qquad (16)$$

is suggested, in which $K(t) = [K_e, K_1{}^T, K_2{}^T, K_r, K_3]^T$ is the estimate of the unknown parameter vector $K^*$ and $K_I$ is a positive constant scalar. This control law is composed of two terms. The first component $K^T(t)w(t)$ has the standard structure for output feedback control of systems with actuator failures [9]. The integral term $K_I \operatorname{sgn}(e(t)) \int_0^t |e(t)| dt$ is used to achieve robustness with respect to unknown plant delays and external disturbance [20].

Introducing the parameter error $\tilde{K}(t) = K(t) - K^*$ and using (16), the tracking error (15) can be expressed as

$$\begin{aligned} e(t) = \frac{k_p}{D_m(s)}[\tilde{K}^T w(t) - K_I \operatorname{sgn}(e(t)) \int_0^t |e(t)| dt] \\ - \frac{k_p}{D_m(s)}[K_m{}^{*T} w_m(t) + \sum_{l=1}^{M} K_{dl}{}^{*T} e_x(t - d_l(t)) \\ + \sum_{l=1}^{M} K_{zl}{}^{*T} z_{el}(t) + (1 - K_1^{*T} H(s)) F f(t)] \end{aligned} \qquad (17)$$

As the usual design method of MRAC for systems without delay, the augmented state vector $\hat{x}=[x^T, x_1^T, x_2^T]^T$ is defined. Let $\hat{e}=\hat{x}-\hat{x}_m$ where $\hat{x}_m(t)$ is the state of a nonminimal realization $\hat{c}(sI-\hat{A})^{-1}\hat{b}K_r^*$ of $W_m(s)$. Then the state space representation

$$\begin{aligned}
\dot{\hat{e}}(t) &= \hat{A}\hat{e}(t)+\hat{b}[\tilde{K}^T(t)w(t)-K_I\,\mathrm{sgn}(e(t))\int_0^t |e(t)|\,dt] \\
&\quad -\hat{b}[K_m^{*T}w_m(t)+\sum_{l=1}^{M} K_{dl}^{*T}L^T\hat{e}(t-d_l(t)) \\
&\quad +\sum_{l=1}^{M} K_{zl}^{*T}C_e\hat{z}_{el}(t)+(1-K_1^{*T}H(s))Ff(t)], \qquad (18)\\
\dot{\hat{z}}_{el}(t) &= A_e\hat{z}_{el}(t)+B_eL^T\hat{e}(t-d_l(t)),\\
z_{el}(t) &= C_e\hat{z}_{el}(t),\\
e(t) &= \hat{c}\hat{e}(t),
\end{aligned}$$

is obtained for (17), where $L=[I_{n\times n}, 0_{n\times(n-1)}, 0_{n\times(n-1)}]^T$ and the triple $(A_e, B_e, C_e)$ is a minimal state space realization for the stable transfer matrix $H_n(s)$.

Because the relative degree is assumed to be one, $W_m(s)=\hat{c}(sI-\hat{A})^{-1}\hat{b}K_r^*$ is SPR. Therefore, according to MKY lemma [21], for any symmetric positive definite matrix $L_c = L_c^T > 0$, there exist a symmetric positive definite matrix $P=P^T>0$, a positive scalar $\upsilon>0$ and a vector $q$ such that

$$\begin{aligned}
\hat{A}^TP+P\hat{A} &= -qq^T-\upsilon L_c,\\
P\hat{b}K_r^* &= \hat{c}^T.
\end{aligned} \qquad (19)$$

Since $A_e$ is stable, there exist symmetric positive definite matrices $P_{zl}=P_{zl}^T>0$ and $Q_{zl}=Q_{zl}^T>0$ that satisfy

$$A_e^TP_{zl}+P_{zl}A_e=-Q_{zl}, \quad l=1,\ldots,M \qquad (20)$$

Now we are ready to state the following theorem.

**Theorem 1:** Consider the system (1) with actuator failures (3) and the reference model (5). Suppose that assumptions (A1) to (A4) hold. Then for any positive number $\gamma$ and positive definite matrix $\Gamma=\Gamma^T>0$, the adaptive control (16) with coefficients

$$\begin{aligned}
\dot{K}(t) &= -\mathrm{sgn}(\rho^*)\Gamma e(t)w(t),\\
K_I &= \gamma\,\mathrm{sgn}(\rho^*),
\end{aligned} \qquad (21)$$

assures that all the closed-loop signals are bounded and the tracking error $e(t)$ converges to zero asymptotically.

**Proof** To prove this theorem, choose the Lyapunov-Krasovskii functional

$$\begin{aligned}
V(t) &= \hat{e}^T(t)P\hat{e}(t)+\sum_{l=1}^{M}\hat{z}_{el}^T(t)P_{zl}\hat{z}_{el}(t)\\
&\quad +\sum_{l=1}^{M}\int_{t-d_l(t)}^{t}\hat{e}^T(s)\upsilon L_c\hat{e}(s)\,ds+\frac{|\rho^*|}{\gamma}(-\gamma\int_0^t|e(t)|\,dt+\eta^*)^2 \qquad (22)\\
&\quad +(\tilde{K}(t)-\hat{K}_1)^T\Gamma^{-1}(\tilde{K}(t)-\hat{K}_1)|\rho^*|
\end{aligned}$$

in which $\gamma>0$ is a constant scalar and $\rho^*=K_r^{*-1}$. The vector $\hat{K}_1$ is defined as $\hat{K}_1=-\frac{r}{2}[\rho^*, 0, \ldots, 0]^T$. The parameters $r>0$ and $\eta^*>0$ with arbitrary values will be defined later.

According to the update law (21) and using (19) and (20), the time derivative of $V(t)$ along (18) is

$$\begin{aligned}
\dot{V}(t) &= -\hat{e}^T(t)qq^T\hat{e}(t)-r\hat{e}^TP\hat{b}\hat{b}^TP\hat{e}\\
&\quad -\sum_{l=1}^{M}\nu(1-\dot{d}_l(t))\hat{e}^T(t-d_l(t))L_c\hat{e}(t-d_l(t))\\
&\quad -\sum_{l=1}^{M}\hat{z}_{el}^T(t)Q_{zl}\hat{z}_{el}(t)-2\sum_{l=1}^{M}\hat{e}^TP\hat{b}K_{dl}^{*T}L^T\hat{e}(t-d_l(t))\\
&\quad -2\sum_{l=1}^{M}\hat{e}^TP\hat{b}K_{zl}^{*T}c_e\hat{z}_{el}+2\sum_{l=1}^{M}\hat{z}_{el}^TP_{zl}B_eL^T\hat{e}(t-d_l(t)) \qquad (23)\\
&\quad -2\hat{e}^TP\hat{b}K_m^{*T}w_m-2\hat{e}^TP\hat{b}K_I\,\mathrm{sgn}(e(t))\int_0^t|e(t)|\,dt\\
&\quad -2|\rho^*||e(t)|(-\gamma\int_0^t|e(t)|\,dt+\eta^*)+2\hat{e}^TP\hat{b}(1-K_1^{*T}H(s))Ff(t)
\end{aligned}$$

for $t\in(T_i, T_{i+1})$, $i=0,1,\ldots,m_0$.

Because $D_m(s)$ and $H_n(s)$ are stable and the reference input $r(t)$ is bounded, the reference signals $x_m(t)$, $x_m(t-d)$ and $z_m(t)$ are bounded. Therefore there exists a constant $w_m^*$ such that $\|w_m(t)\|\le w_m^*$ and the following inequality can be written for the eighth term of (23).

$$\begin{aligned}
-2\hat{e}^TP\hat{b}K_m^{*T}w_m &\le 2|\hat{e}^TP\hat{b}K_r^*||\rho^*|\|K_m^{*T}\|\|w_m\|\\
&\le 2|e(t)||\rho^*|\|K_m^{*T}\|w_m^*
\end{aligned} \qquad (24)$$

By choosing $\eta_1^* = \left\|K_m^{*T}\right\| w_m^*$,

$$-2\hat{e}^T p\hat{b}K_m{}^{*T} w_m \le 2\eta_1^* \left|\rho^*\right|\left|e(t)\right|. \tag{25}$$

For the last term of (23) the following inequality can be written.

$$\begin{aligned} &2\hat{e}^T P\hat{b}(1-K_1^{*T}H(s))Ff(t) \\ &\le \left|e(t)\right|\left|\rho^*\right|\left\|2(1-K_1^{*T}H(s))F\right\| f^* = 2\eta_2^*\left|\rho^*\right|\left|e(t)\right| \end{aligned} \tag{26}$$

According to the known inequality

$$\pm 2x^T y \le x^T Sx + y^T S^{-1} y$$

that is true for any vectors $x, y$ and any positive definite matrix $S$, the following expressions can be written for the fifth, sixth and seventh terms

$$\begin{aligned} &-2\sum_{l=1}^{M} \hat{e}^T p\hat{b}K_{dl}{}^{*T} L^T \hat{e}(t-d_l(t)) \\ &\le \hat{e}^T(t)P\hat{b}\Phi_1\hat{b}^T P\hat{e}(t) + \sum_{l=1}^{M} \hat{e}^T(t-d_l(t))S\hat{e}(t-d_l(t)), \\ &-2\sum_{l=1}^{M} \hat{e}^T p\hat{b}K_{zl}{}^{*T} c_e\hat{z}_{el} \\ &\quad \le \hat{e}^T(t)P\hat{b}\Phi_2\hat{b}^T P\hat{e}(t) + \sum_{l=1}^{M} \hat{z}_e^T S\hat{z}_e, \\ &2\sum_{l=1}^{M} \hat{z}_{el}^T P_{zl}B_e L^T \hat{e}(t-d_l(t)) \\ &\quad \le \sum_{l=1}^{M} \hat{z}_{el}^T\Phi_3\hat{z}_{el} + \sum_{l=1}^{M} \hat{e}^T(t-d_l(t))S\hat{e}(t-d_l(t)), \end{aligned} \tag{27}$$

where

$$\begin{aligned} \Phi_1 &= \sum_{l=1}^{M} K_{dl}{}^{*T} L^T S^{-1} LK_{dl}^* \\ \Phi_2 &= \sum_{l=1}^{M} K_{zl}{}^{*T} c_e S^{-1} c_e^T K_{zl}^* \\ \Phi_3 &= P_{zl}B_e L^T S^{-1} LB_e^T P_{zl}^T. \end{aligned}$$

Using (25), (26) and (27), choosing the coefficient $K_I$ from (21) and defining $\eta^* = \eta_1^* + \eta_2^*$, the inequality

$$\begin{aligned} \dot{V}(t) &\le -\hat{e}^T(t)qq^T\hat{e}(t) \\ &-\sum_{l=1}^{M} \hat{e}^T(t-d_l)(v\,d^* L_c - 2S)\hat{e}(t-d_l(t)) \\ &-\hat{e}^T P\hat{b}(r-\Phi_1-\Phi_2)\hat{b}^T P\hat{e} \\ &-\sum_{l=1}^{M} \hat{z}_{el}^T(t)(Q_{zl}-\Phi_3-S)\hat{z}_{el}(t) \end{aligned} \tag{28}$$

with $d^* = 1-\bar{d}$ is obtained. By choosing the arbitrary values $L_c$, $r$ and $Q_{zl}$ as

$$\begin{aligned} &\lambda_{\min}(vL_c) > \lambda_{\max}(2S), \\ &r > \lambda_{\max}(\Phi_1+\Phi_2), \\ &\lambda_{\min}(Q_{zl}) > \lambda_{\max}(\Phi_3+S) \end{aligned}$$

we have $\dot{V}(t) \le 0$ for $t \in (T_i, T_{i+1})$ , $i = 0,1,\ldots,m_0$.

Since there are only a finite number of failures in system, $V(T_{m_0})$ is finite and from

$$\dot{V}(t) \le 0, \quad t \in (T_{m_0}, \infty) \tag{29}$$

we have $V(t) \in L_\infty$ and therefore $\hat{e}(t)$ , $e(t)$ , $\hat{z}_e(t)$ , $K(t)$ , $\tilde{K}(t) \in L_\infty$. Using directly the analysis method in [21], it can be proved that all the closed loop signals are bounded and $\lim_{t\to\infty} e(t) = 0$. ■

## V. Simulation Example

Consider system (1) with parameters

$$A = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & -2 \end{bmatrix}, \; B = \begin{bmatrix} 1 & 2 & 1 \\ 1 & 1 & 3 \\ 0 & 0 & 0 \end{bmatrix},$$

$$c = \begin{bmatrix} 1 & 1 & 0.5 \end{bmatrix}, \quad b_f = \begin{bmatrix} 1 & 2 & 0 \end{bmatrix}^T$$

$$A_{d1} = \begin{bmatrix} 0.1 & -0.4 & 0.4 \\ 0.1 & -0.7 & 0.3 \\ 0 & 0 & 0 \end{bmatrix}, \; d_1(t) = 4 + 0.5\sin(t)$$

$$A_{d2} = \begin{bmatrix} -0.2 & 0.1 & 0 \\ 0.4 & 0.1 & -0.4 \\ 0 & 0 & 0 \end{bmatrix}, d_2(t) = 6 + 0.4\sin(t)$$

$$x(0) = \begin{bmatrix} 0.5 & 0 & 0 \end{bmatrix}^T$$

and the external disturbance $f(t) = 0.2\sin(0.5t) + 0.3$ and let the transfer function of the reference model be given by

$$W_m(s) = \frac{1}{s+1}.$$

All parameters of system and the delay value are assumed to be unknown to the controller. Simulation results are obtained with the adaptive control (16) with coefficients (21), reference input $r(t) = \sin(0.2t)$ and by choosing $\Gamma = 10I$, $\gamma = .001$ and $\Lambda(s) = (s+1)^2$, for one and two actuator failure cases, respectively. The results for the case of one actuator failure: $u_1(t) = .5,\ t \geq 50$, are shown in Fig. 1. In Fig. 2, the simulation results are shown for the case when two actuators fail: $u_1(t) = 1,\ t \geq 30, \quad u_2(t) = 0,\ t \geq 70$.

## VI. Conclusion

A robust output-feedback adaptive controller is presented for state delayed systems with unknown parameters and unknown actuator failures. The controller structure is composed of two terms. The first term is defined for compensating actuator failures. The second integral term is used to achieve robustness against unknown delay values and an external disturbance with unknown bounds. The results of this paper can be extended to higher relative degrees using normalized MRAC scheme.

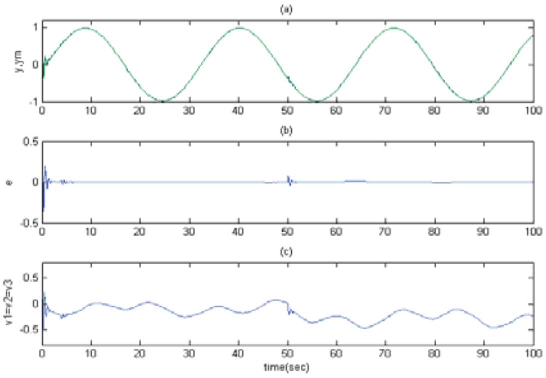


Figure 1. Simulation results for one actuator failure: (a) The plant and reference model outputs; (b) Tracking error; (c) control input.

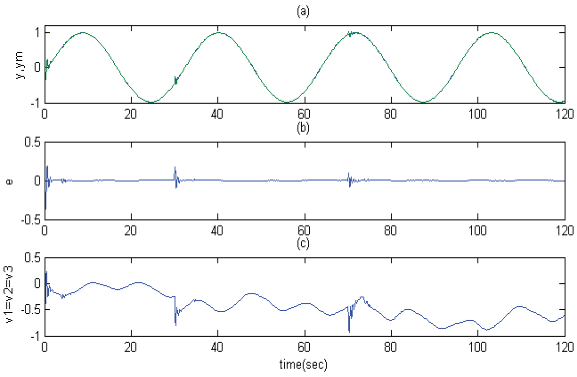


Figure 2. Simulation results for two actuator failures: (a) The plant and reference model outputs; (b) Tracking error; (c) control input.

## References

[1] J. D. Boskovic and R. K. Mehra, “An adaptive retrofit reconfigurable flight controller”, Proc. of IEEE Conf. Decision Contr., pp. 1257-1262 , 2002.

[2] M. L. Corradini and G. Orlando, ‘A sliding mode controller for actuator failure compensation’, Proc. 42nd IEEE conf. on decision and control, Denver, Colorado, pp. 4255-4260 , 2003.

[3] H. N. Wu and H. Y. Zhang, “Reliable $H_\infty$ fuzzy control for continuos-time nonlinear systems with actuator failures”, IEEE Trans. on Fuzzy Systems, vol.14, no. 5, pp. 609-618, 2006.

[4] W. Chen and M. Saif, “Adaptive actuator fault detection, isolation and accommodation (FDIA) in uncertain systems”, Int. Journal of Control, vol. 80, no.1, pp. 45-63, 2007.

[5] H. Yang, B. Jiang and M. Staroswiecki, “Observer-based fault-tolerant control for a class of switched nonlinear systems”, IET Control Theory and Applications, vol.1, no. 5, pp. 1523-1532 , 2007.

[6] M. Benosman and K.- Y. Lum, “Application of absolute stability theory to robust control against loss of actuator effectiveness”, IET Control Theory and applications, vol. 3, no. 6, pp. 772-786, 2009.

[7] Z. Zhang, S. Xu and B. Wang, “Adaptive actuator failure compensation with unknown control gain signs”, IET Control Theory and applications, vol. 5, no. 16, pp. 1859-1867, 2011.

[8] G. Tao, S. M. Joshi and X. Ma, “Adaptive state feedback and tracking control of systems with actuator failures”, IEEE Trans. Automat. Control, vol. 46, pp. 78–95, 2001.

[9] G. Tao, S. Chen and S. M. Joshi, “An adaptive actuator failure compensation using output feedback”, IEEE Trans. Automat. Control, vol. 47, no. 3, pp. 506-511, 2002.

[10] G. Tao, S. Chen, X. D. Tang and S. Joshi, Adaptive control of systems with actuator failures, Springer- Verlag, London, 2004.

[11] X. D. Tang, G. Tao and S. M. Joshi, “Adaptive actuator failure compensation for nonlinear MIMO systems with an aircraft control application”, Automatica, vol. 43, no. 11, pp. 1869-1883, 2007.

[12] W. Chen and M. Saif, “Fault detection and accommodation in nonlinear time-delay systems”, Proc. American Control Conference, Maui, Hawaii U.S.A., pp. 4291-4296, 2003.

[13] Q. Zhao and C. Cheng, “State-feedback control for time-delayed systems with actuator failures”, Proc. American Control Conference, Denver, Colorado, pp. 827-832, 2003.

[14] C. Cheng and Q. Zhao, “Reliable control of uncertain delayed systems with integral quadratic constrains”, IEE Proc. Control Theory and Applications, vol. 151, no. 6, pp. 790-796, 2004.

[15] B. M. Mirkin and P.-O. Gutman, “Model reference adaptive control of state delayed system with actuator failures”, International Journal of Control, vol. 78, no. 3, pp. 186-195, 2005.

[16] B. M. Mirkin and P.-O. Gutman, “Adaptive coordinated decentralized control of state delayed systems with actuator failures”, Asian Journal of Control, vol. 8, no. 4, pp. 441-448, 2006.

[17] D. Ye and G.-H. Yang, “Delay-dependent adaptive reliable $H_\infty$ control of linear time-varying delay systems”, Int. Journal of Robust and Adaptive Control, vol. 19, pp. 462-479, 2009.

[18] M. Kamali, J. Askari and F. Sheikholeslam, “An Output-feedback adaptive actuator failure compensation controller for systems with unknown state delays”, Nonlinear Dynamics, vol. 67, no. 4, pp. 2397-2410, 2012.

[19] M. Kamali and J. Askari, “utput-feedback model reference adaptive control of linear continuous state delayed systems in the presence of actuator failures”, 8th IEEE international conference on Control and Automation, Xiamen, China, June 9-11, 2010.

[20] B. M. Mirkin and P. O. Gutman, “Robust output-feedback model reference adaptive control of SISO plants with multiple uncertain, time-varying state delays”, IEEE Trans. automatic cont., vol. 53, no. 10, pp. 2414-2419, 2008.

[21] P. A., Ioannou and J. Sun, Robust Adaptive Control, Prentice-Hall, New Jersey , 1996.

# Invariant Control of Non-Linear Elements in a Stacked High Redundancy Actuator

T. Steffen, R. Dixon, R. Goodall
Department of Aeronautical and Automotive Engineering,
Department of Electronic and Electrical Engineering,
Loughborough University, LE11 3TU, UK

***Abstract*—The High Redundancy Actuator (HRA) concept aims to provide a single actuator comprising many cooperating actuation elements. The potential benefits of this include improved overall reliability, availability, and reduced need for oversizing of actuators in safety critical applications. This paper deals with the question of distributing the load evenly between a stack of elements despite non-linear characteristics.**

**The approach is to separate the state space into a high dimensional internal and a low dimensional invariant (or external) subspace. If the internal states can be decoupled and damped, the input-output behaviour only depends on the few states of the invariant subspace. In other words: the high redundancy actuator with many redundant elements behaves just like a conventional single actuator, and classical control strategies can then be applied.**

**This approach is demonstrated here for an HRA constructed from simple spring-damper-actuator elements. The non-linear behaviour is required to simulate the element behaviour at the end of the available travel. Without equalisation, excessive accelerations are caused by individual elements hitting the end stop, and this can be avoided by applying the proposed element tuning.**

***Index Terms*—electromagnetic actuation, fault tolerance, multi-variable control, non-linear control**

## I. High Redundancy Actuation

High Redundancy Actuation (HRA) is a novel approach for designing a fault tolerant actuator that comprises a relatively large number of actuation elements (see Figure 1). As a result, faults in the individual elements can be inherently accommodated without resulting in a failure of the complete actuator system.[1]

The concept of the High Redundancy Actuation (HRA) is inspired by the human musculature. A muscle is composed of many individual muscle cells, each of which provides only a minute contribution to the force and the travel of the muscle. These properties allow the muscle as a whole to be highly resilient to damage of individual cells. The aim of this project is not to replicate muscles, but to use the same principle of co-operation with existing technology to provide intrinsic fault tolerance.

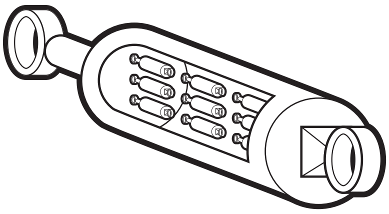


Figure 1. High Redundancy Actuator

An important feature of the High Redundancy Actuator is that the elements are connected both in parallel and in series. While the parallel arrangement is commonly used, the serial configuration is rarely employed, because it is perceived to be less efficient. However, the use of elements in series is the only configuration that can deal with the lock-up of an element. In a parallel configuration, this would immediately render all elements useless, but in the series configuration it only leads to a slight reduction of available travel (see Steffen et al. 2007, 2008a for details).

The paper is organised as follows: it starts with the motivation and background in Section 2, followed by the non-linear model of an actuator stack consisting of elements with a non-linear characteristic in Section 3. The control goal is presented in Section 4, followed by the solution in Section 5. Sections 6 presents a simulation example, and Section 7 proposes an extension for more complex HRA configurations.

## II. Motivation

While the parallel configuration of actuation elements is well established and researched, the dynamic behaviour of elements in series is very different. The

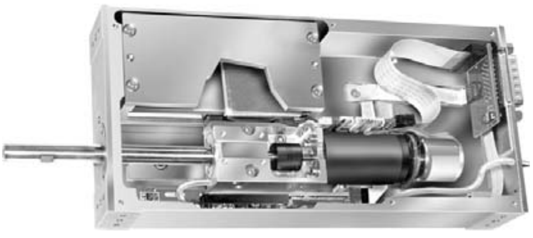


Figure 2. Electromechanical actuator

[1] This project is a cooperation of the Control Systems group at Loughborough University, the Systems Engineering and Innovation Centre (SEIC), and the actuator supplier SMAC Europe limited. The project was funded by the Engineering and Physical Sciences Research Council (EPSRC) of the UK under reference EP/D078350/1.

reason is that between each element in series, there is a moving mass. This creates a high number of mechanical degrees of freedom, in turn leading to a high order dynamic model. The model needs to describe the position and speed of each mass separately, and there may also be further states internal to each element.

For the envisioned number of elements (10x10 or more), this may lead to a model with hundreds of states. Dealing with this complexity constructively is crucial for the success of the high redundancy actuator, because the standard approach of using sophisticated instrumentation may not be suitable. The goal of this paper is to reduce the model and instrumentation complexity to a level comparable to a conventional actuator.

### A. Approach

The basic idea is to split the travel equally between all actuation elements. If this is achieved, the states of the elements are no longer individual variables, and they can all be reduced into a single simple model. In other words: because the whole system behaves like a single conventional actuator, a simple conventional actuator model is sufficient to describe it.

This approach is not trivial, because the elements experience different effective loads. The element at the bottom of the assembly for example experiences a higher load, because it needs to move all the other elements in addition to the load. This paper will present a number of ways to address this problem using active and passive, feedforward and feedback approaches. The relative advantages and disadvantages are discussed, and two approaches are considered in more detail. Based on this set of options, a specific solution or combination of solutions can be selected as appropriate according to the practical requirements.

### B. Literature Overview

High Redundancy Actuation is a novel approach to fault tolerance, and consequently the specific problem formulated in this paper has not been previously considered.

Previous work on High Redundancy Actuation did look into the possibility of aligning dynamics using different methods (see Steffen et al., 2008b, 2010), but without using the geometric approach. This leads to results that are much less general than the approach presented here.

A similar approach is used in rotary actuation with torque summing and velocity summing gears, see for example Ting et al. [1994], Bennett et al. [2004]. The approach has a number of characteristic differences to the systems studied in this paper: it only works for rotary motion, and the problem of complexity is much less prominent because, due to the symmetrical structure, all elements act and behave in exactly the same way. Thus the parameter tuning described in this paper is not applicable.

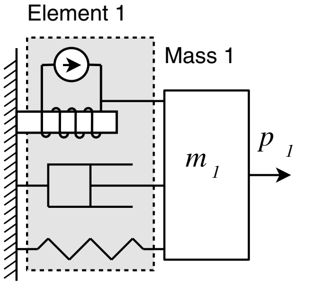


Figure 3. Dynamic components of a single element

A related problem has been studied in the dynamic behaviour of axial stacks of piezoelectric actuators by Jalili [2009]. The author models the stack as a distributed system with partial differential equations (compared to the lumped elements in this paper), which leads to similar results concerning the internal mode. However, the author makes no apttempt to control or decouple these modes internally.

The basis for decoupling internal modes is the geometric approach, because it creates a connection between the dynamics of the system and constraints formulated in terms of the states of the system. This approach was introduced by Wonham [1985] and later extended by Basile and Marro [1992]. It provides the standard solution for the disturbance decoupling problem (see Commault et al., 1997), which is the class of control problems at the heart of this paper. The geometric approach has previously been used for adaptive control of a High Redundancy Actuator in Steffen et al. [2009].

### C. Symbols

| | |
|---|---|
| diag{} | diagonal matrix |
| $f_j$ | force produced by element $j$ |
| $g_j$ | strength factor for element $j$ |
| $F_i$ | force total for mass $m_i$ |
| $f$ | characteristic element function (non-linear model) |
| $k_d\, k_r$ | damping and force constant (linear model) |
| $m_i$ | mass of moving mass number $i$ |
| $n_i\ n_j$ | number of masses and elements |
| $\mathbf{Q}$ | the connection matrix $\in \{-1, 0, -1\}^{j \times i}$ |
| $\mathbb{R}$ | set of real numbers |
| $t$ | time |
| $u_j$ | input of element $j$ |
| $x_i\ \dot{x}_i\ \ddot{x}_i$ | position, speed and acceleration of mass $m_i$ |

## III. System Model

The basic components of an electromechanical actuation element are shown in Figure 3. From a modelling perspective, it is a typical actuated spring-mass-damper system, which can be described by NEWTONian mechanics. Three forces act upon the mass: the electromagnetic force $F_{el}$, the damping force $F_d$, and the

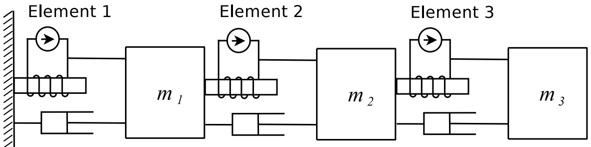


Figure 4. 3 Elements in Series

spring force $F_s = rx$ (see Davies et al. 2008 for more details). Damping and spring for are often assumed to be linear, but in reality that is rarely the case, and there are always limits to the linear region which may be relevant. Therefore, a nonlinear model is used here, using a characteristic function $f(x,\dot{x})$ to describe the behaviour of the spring and the damper:

$$m\ddot{x} = f(x,\dot{x},u) \quad .$$

Choosing $x$ and $\dot{x}$ as states leads to a full state space model:

$$\frac{d}{dt}\mathbf{x} = \mathbf{f}(\mathbf{x},\mathbf{u})$$

with

$$\mathbf{x} = \begin{pmatrix}\dot{x}\\ x\end{pmatrix}$$

$$\mathbf{f}(\mathbf{x},\mathbf{u}) = \begin{pmatrix}f(x,\dot{x},u)\\ \dot{x}\end{pmatrix}$$

$$\mathbf{u} = \begin{pmatrix}u\end{pmatrix} \quad .$$

Once several actuator elements are stacked, it is worth distinguishing between the mass (inertia) and the connection part of an element (generating the there forces). The connection part is attached to the element (or ground) below, so each element is subject to forces from both sides (except for the top one, as shown in Figure 4). The resulting model for actuation elements in series is:

$$\mathrm{diag}\{\mathbf{m}\}\ddot{\mathbf{x}} \;= -\mathbf{Q}f(\mathbf{Q}^T\mathbf{x},\mathbf{Q}^T\dot{\mathbf{x}},\mathbf{u}) \qquad (1)$$

where

$$\mathbf{Q} = \begin{pmatrix}1 & -1 & 0\\ 0 & 1 & -1\\ 0 & 0 & 1\end{pmatrix}$$

is the connection matrix defining the connections between the actuators and the masses. The notation $f(\mathbf{Q}^T\mathbf{x},\mathbf{Q}^T\dot{\mathbf{x}})$ indicates that the same function $f$ is applied element wise to the elements of $\mathbf{Q}^T\mathbf{x}$ and the derivative. More complicate configuration including parallel configurations can be modelled by supplying the corresponding connection matrix $\mathbf{Q} \in \mathbb{R}^{n_j \times n_i}$, but only stacks will be considered here with $n_i = n_j = n$, and $\mathbf{Q}$ a band diagonal matrix of the same structure as above. The time variables $\mathbf{u}$, $\mathbf{x}$, and $\mathbf{m}$ are vectors describing the input and the position of the elements, while the parameter vector $\mathbf{m}$ specifies the mass of the elements. This model is in state space form, and the state variable consists of the velocity and the position components

$$\begin{pmatrix}\dot{\mathbf{x}}\\ \mathbf{x}\end{pmatrix} \in \mathbb{R}^{2n} \quad .$$

Internal states of the elements (e.g. current in the coil) can be modelled in the same way if necessary.

The scalar equations behind this vector equation are

$$m_i\ddot{x}_i = -\mathbf{q}_i\mathbf{f}$$

and

$$f_j = f(\mathbf{q}_j^T\mathbf{x},\mathbf{q}_j^T\dot{\mathbf{x}},\mathbf{u})$$

where $\mathbf{q}_i$ is the $i$th row of $\mathbf{Q}$, $\mathbf{q}_j^T$ is the transpose of the $j$th column of $\mathbf{Q}$, and $\mathbf{f}$ is a column vector of all $f_j$.

In fact one further term is required to solve this problem, because using equal elements will not generally be sufficient. Because the elements close to the ground have a higher load, they need to be stronger, and this is modelling using a strength factor $\mathbf{l}$. The characteristic function is scaled with this factor:

$$f_j = g_j f(\mathbf{q}_j^T\mathbf{x},\mathbf{q}_j^T\dot{\mathbf{x}})$$

or

$$\mathrm{diag}\{\mathbf{m}\}\ddot{\mathbf{x}} \;= -\mathbf{Q}\mathrm{diag}\{\mathbf{g}\}f(\mathbf{Q}^T\mathbf{x},\mathbf{Q}^T\dot{\mathbf{x}},\mathbf{u}) \qquad (2)$$

## IV. Alignment Goal

One of the goals of high redundancy actuation is that all elements work together to perform the actuation task. This requires that they move in a coordinated and generally synchronous motion. This is called the alignment goal.

The key obstacle that the elements are not equal, because they at different positions along the stack and therefore subject to different accelerations. The naive approach of applying the same input signal to all elements does not achieve the alignment of motion. Instead, the force created by two elements cancels out for all masses except the load mass. The top element ($x_3$) responds first, then the middle element ($x_2$) begins to move, and finally the element on the base ($x_1$) will respond. So the input responses ripples through the system like a longitudinal wave. This phenomenon is already known from stacks of piezoelectric actuators.

An example is shown in Figure 5 for a nominal system with $f(x,\dot{x},u) = u - 2\dot{x} - 0.5x$, $m_1 = m_2 = 0.5$ and $m_3 = 1$. A simple single input/single output (SISO) proportional controller with a phase lead compensator is used $K(s) = 2\frac{0.4s+1}{4s+1}$, and a reference step of 30 mm (10 mm per element) is used in this simulation.

This kind of wave propagation complicates the control of the actuator, because it leads to situations where the elements work against each other, and not in cooperation. It can also cause elements to run into mechanical limits, which may create further vibrations and reduce the lifetime of the system significantly.

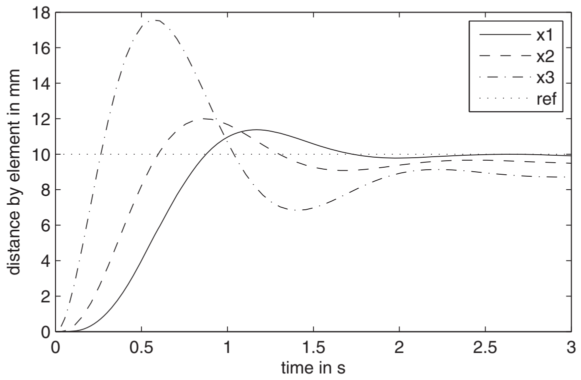


Figure 5. Delay between elements

To avoid this, the alignment goal is introduced. It states that all elements should follow a synchronous motion. Let $\Delta x_i = x_i - x_{i-1}$ denote the extensions of element number $i$, then the alignment goal is

$$\Delta x_1 = \Delta x_2 = \ldots = \Delta x_{n_j} \tag{3}$$

In the case of stacked elements, this can be reformulated as follows.

**Definition 1.** The alignment goal is to distribute the extension of the stack equally between all elements

$$\forall i : n\Delta x_i = x_n \quad . \tag{4}$$

Error signals for this goal can be defined either on the extension

$$e_i = n\Delta x_i - x_n$$

or on the position

$$e_i = x_i - \frac{i}{n} x_n \quad .$$

The following section will present a number of solutions to achieve this goal.

## V. Solution

There are several approaches that can be used either individually or in combination to address the alignment goal. For an overview see [to be published], it lists a number of active and passive approaches. This paper focuses on element tuning with the goal of creating an invariant subspace. Only the strength of the elements is changed here, while the characteristics are assumed to be identical across all elements.

The feedforward approach presented here is not about controlling the deviations, but making sure that they do not occur in the first place. So the goal becomes to equalise the derivatives:

$$\Delta \ddot{x}_1 = \Delta \ddot{x}_2 = \ldots = \Delta \ddot{x}_{n_j} \tag{5}$$

assuming that all $\Delta x_j$ and $\Delta \dot{x}_j$ are equal. This is the basic idea of invariant control: once the invariant condition is satisfied, invariant control can ensure that it remains valid.

In a stack, the travel of an element $\Delta x_j$ can be calculated as $\mathbf{q}_j^T \mathbf{x}$. Inserting the model equation results in

$$\Delta x_j = -\mathbf{q}_j^T \mathrm{diag}^{-1}\{\mathbf{m}\}\mathbf{Q}\mathrm{diag}\{\mathbf{g}\}\mathbf{f}$$

with

$$f_j = f(\mathbf{q}_j^T \mathbf{x}, \mathbf{q}_j^T \dot{\mathbf{x}}, \mathbf{u}) \quad .$$

Because of the assumption, all $f_j$ are equal, so the goal can be achieved if all elements of

$$\mathbf{Q}^T \mathrm{diag}^{-1}\{\mathbf{m}\}\mathbf{Q}_i \mathbf{g}$$

are equal, or

$$\mathbf{g} = \left(\mathbf{Q}^T \mathbf{Q}\right)^{-1} (1\ 1\ \cdots\ 1)^T \quad .$$

Since $\mathbf{Q}$ is invertible for the stack configuration considered here, this can be rewritten as:

$$\mathbf{g} = \mathbf{Q}^{-1} \mathbf{diag}\{\mathbf{m}\} \mathbf{Q}^{-1,T} (1\ 1\ \cdots\ 1)^T \quad .$$

With

$$\mathbf{Q}^{-1} = \begin{pmatrix} 1 & & \\ \vdots & 1 & \\ 1 & \cdots & 1 \end{pmatrix}$$

this can be rewritten as

$$g_j = \sum_{k=j}^{n} k m_k \quad .$$

In the example with three elements, this means

$$\begin{aligned} g_1 &= m_1 + 2m_2 + 3m_3 \\ g_2 &= 2m_2 + 3m_3 \\ g_3 &= 3m_3 \quad . \end{aligned}$$

The result of this tuning is shown in Figure 6. The parameters of the third element are equal to the step response in Figure **??**, and the other elements are tuned accordingly. Clearly the delay between the elements has been eliminated, and they all respond at the same time. The disturbance response (at $t = 5$) still deviates slightly between the elements, but the difference is small and not significant for most practical purposes.

## VI. Simulation Results

As an example, a stack of three elements is simulated. The non-linear characteristic of each element is used to include a linear damping term, a linear spring term, and a non-linear spring term representing end stops:

$$f(\dot{x}, x, u) = u - 2\dot{x} - 0.1x - 10000h(x)$$

where

$$h(x) = \begin{cases} x & \text{if } x < 0 \\ 0 & \text{if } x \in [0, 0.35] \\ x - 0.35 & \text{if } x > 0.35 \end{cases} \quad .$$

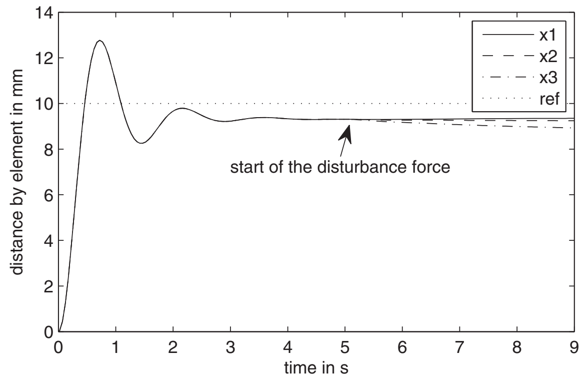


Figure 6. Step response after parameter tuning

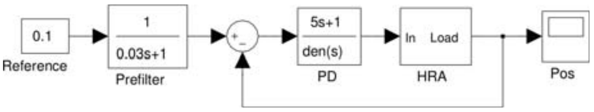


Figure 7. Simulation Control Structure

A simple PD type controller with a slight prefilter is used to control the stack as shown in Figure 7. This may not be the controller of choice for practical applications (for example because it does not provide reference tracking), but the simplicity makes the analysis of the system behaviour easier. With equal elements (without tuning), the simulation in Figure 8 shows a strong chattering effect caused by the end stops becoming active at an extension of $\Delta x = 0.35$. The state trace in Figure 9 also demonstrates that the travel is by no means equally distributed between the elements. This kind of chatter has also been observed in an experimental demonstrator of the HRA. The forces caused by the end stops and acting on the individual masses can be significant, so much so that they destroy the assembly.

Adjusting the strength of the three elements according to the methodology shown above changes the situation dramatically. The three weights are $m_1 = m_2 = 0.1\,\mathrm{kg}$ and $m_3 = 1\,\mathrm{kg}$, therefore the strength coefficients are

$$\begin{aligned} g_1 &= 3.3 \\ g_2 &= 3.2 \\ g_3 &= 3 \quad . \end{aligned}$$

In the implementation, these are scaled by $g_3$ such that $g_3$ becomes equal to 1. As can be seen in Figure 10, all elements of the adjusted HRA move in synchronose motion. Because the overall travel is well within the reach of the HRA, the end stops do not become relevant.

## VII. Generalised Configurations

This approach is not limited to stacks of actuators - it can be applied to any network of actuators. For example the system in Figure 12 has a connection matrix of the form

$$\mathbf{Q} = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 1 \end{pmatrix} \quad .$$

The equalisation can be calculated in the same way, but since $\mathbf{Q}$ is not square, the left side inverse is not unique. This agrees with the intuitive interpretation: both stacks can have the same or different strength, and the alignment goal can be achieved in either case. If the pseudo-inverse is used, they will both have the

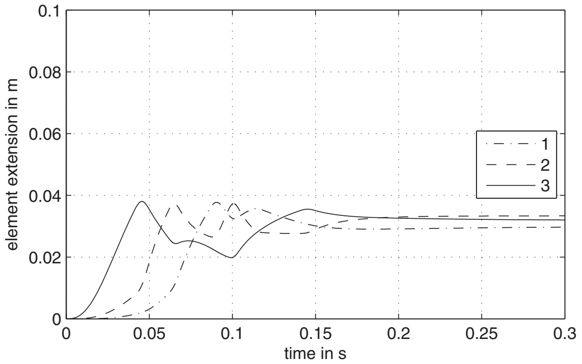


Figure 8. Simulation Results, unmatched

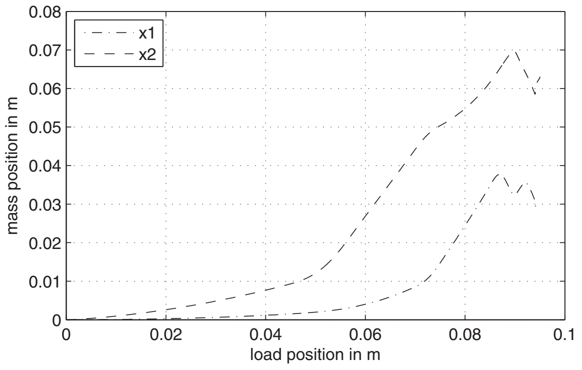


Figure 9. State trace, unmatched

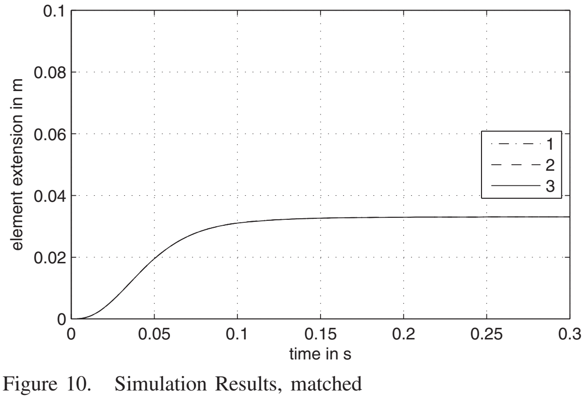


Figure 10. Simulation Results, matched

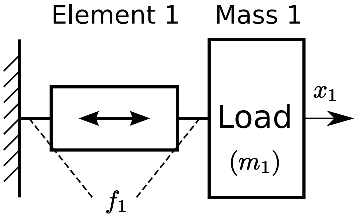


Figure 11. Abstracted Model of a Single Element

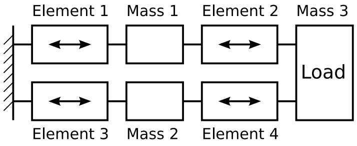


Figure 12. Example 2×2 Configuration

same strength:

$$\mathbf{Q}^* = \frac{1}{4}\begin{pmatrix} 3 & 1 & 2 \\ 1 & 3 & 2 \\ -1 & 1 & 2 \\ 1 & -1 & 2 \end{pmatrix}$$

and

$$\mathbf{g} = \lambda (m_{1,2} + m_3 \;\; m_{1,2} + m_3 \;\; m_3 \;\; m_3)^T \quad .$$

## VIII. Conclusion

Starting from a model of an actuator stack, it is shown how a desired equal distribution of the speed and travel between the elements can be formulated as an invariance condition called alignment goal. The invariance can be achieved by tuning the strength of the actuation elements according to their position in the stack. While this approach is obvious within a linear framework, it is equally applicable to elements with a non-linear characteristic, as long as the characteristic is the same for all elements.

Once the alignment goal is achieved, all elements move synchronously “as one”, and the stack of actuators behaves exactly the same way as a comparable single (larger) actuator. Further modes still exist in the system, but they are stable and decoupled from the inputs. Therefore it is possible to use a simple PID class controller to control an HRA.

The presented methods can be extended to more complex configurations by relaxing some of the assumptions. A simple 2x2 configuration is demonstrated, but it also covers much more extensive configuration.

## References

G. Basile and G. Marro. *Controlled and Conditioned Invariants in Linear System Theory*. Prentice Hall, 1992. ISBN 0-13-172974-8.

J. W. Bennett, A. G. Jack, B. C. Mecrow, D. J. Atkinson, C. Sewell, and G. Mason. Fault-tolerant control architecture for an electrical actuator. volume 6, pages 4371–4377 Vol.6, 2004. ISBN 0275-9306. ID: 115.

C. Commault, J. M Dion, and V. Hovelaque. A geometric approach for structured systems: Application to disturbance decoupling. *Automatica*, 33(3):403–409, 1997.

J. Davies, T. Steffen, R. Dixon, R. M. Goodall, A. C. Zolotas, and J. Pearson. Modelling of high redundancy actuation utilising multiple moving coil actuators. In *Proceedings of the IFAC World Congress 2008*, Jul 6-11 2008.

Nader Jalili. *Piezoelectric-based vibration-control: from macro to micro/nano scale systems (Google eBook)*, volume 2009. Springer, 2009. ISBN 1441900691.

T. Steffen, R. Dixon, R. M. Goodall, and A. C. Zolotas. Requirements analysis for high redundancy actuation. Technical Report CSG-HRA-2007-TR-4, 2007.

T. Steffen, J. Davies, R. Dixon, R. M. Goodall, J. Pearson, and A. C. Zolotas. Failure modes and probabilities of a high redundancy actuator. In *Proceedings of the IFAC World Congress 2008*, Jul 6-11 2008a. accepted.

T. Steffen, R. Dixon, R. M. Goodall, and A. Zolotas. Multi-variable control of a high redundancy actuator. In *Actuator 2008 – International Conference and Exhibition on New Actuators and Drive Systems – Conference Proceedings*, pages 473—476. HVG, 2008b. ISBN 3-933339-10-3.

T Steffen, AC Zolotas, RM Goodall, and R Dixon. Adaptive control of a high redundancy actuator using the geometric approach. In *7th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes*, pages . 1581–1586, Barcelona, Spain, 2009.

T Steffen, R Dixon, JT Pearson, and RM Goodall. Experimental verification of high redundancy actuation. In *UKACC International Conference on Control*, pages .1–6, Coventry, UK, Sep 2010.

Y. Ting, S. Tosunoglu, and R. Freeman. Torque redistribution methods for fault recovery in redundant serial manipulators. In *Proceedings of the 1994 IEEE International Conference on Robotics and Automation, May 8-13 1994*, Proceedings - IEEE International Conference on Robotics and Automation, pages 1396–1401, San Diego, CA, USA, 1994. Univ of Texas at Austin, Austin, TX, USA, Publ by IEEE, Piscataway, NJ, USA. ISBN 1050-4729; 0-8186-5332-9. ID: 132; Compilation and indexing terms, Copyright 2006 Elsevier Inc. All rights reserved.

W. M. Wonham. *Linear Multivariable Control–A Geometric Approach*. Springer, 1985.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Control of chaotic systems with uncertain parameters and stochastic disturbance by LMPC

Shuang Li*†, Guo-Ping Liu†
*School of Statistics, Xi'an University of Finance and Economics, Xi'an, PR China, 710100
Email: lish2006onetwo@163.com
†Faculty of Advanced Technology, University of Glamorgan, Pontypridd, UK, CF37 1DL
Email: gpliu@glam.ac.uk

***Abstract*—For the chaotic systems with uncertain parameters and stochastic disturbance, in order to satisfy some optimal performance index when chaos control is achieved, the Lyapunov-based model predictive control (LMPC) is introduced. The LMPC scheme is concerned with an auxiliary controller which is constructed in advance. Based on the auxiliary controller and stochastic stability theory, it is shown that the chaotic systems with uncertain parameters and stochastic disturbance are practical stable. With the help of the auxiliary controller, the stability of LMPC can be guaranteed as well as some optimality property. As an example, the unified chaotic system with uncertain parameter and stochastic disturbance is considered and simulation results show the effectiveness of the proposed method.**

***Index Terms*—chaos, predictive control, uncertain parameters, stochastic disturbance, optimization.**

## I. Introduction

Chaos is a complex nonlinear phenomenon that the behavior of dynamical systems is highly sensitive to initial conditions, which is also referred to as the butterfly effect. It exists in many practical fields such as biology, economics, engineering, finance, physics, oscillating chemical reactions, fluid and so on. Because of the sensitiveness on initial conditions, chaos is unpredictable in the long time and may be undesired in some applications. In order to suppress this undesirable behavior, Ott, Grebogi and Yorke first presented a kind of controlling chaos method, which is so called OGY control method [1]. Since then, chaos control has been a hot issue and many techniques have been presented [2], for example, delayed feedback control method [3], adaptive control method [4], linear and nonlinear control methods [5]–[7], active control method [8], etc.

In real systems or experimental situations, it is difficult to obtain the exact model. Instead, people usually only know the approximate system model, and uncertain or/and stochastic disturbance exists inevitably. It is important and necessary to discuss the control problem when the systems include uncertain or/and stochastic disturbance. On the other hand, in many practical control fields such as economics, engineering, finance and so on, the control problem is often connected with some optimal performance index for example saving costs (money, energy), gaining the most profits, and even some constraints. Obviously, the above question is concerned with the optimal control problem of the systems with uncertain or/and stochastic disturbance. In many application fields, model predictive control [9] is adopted widely when the control problem includes optimal performance, constraints, uncertainty, etc. MPC is a kind of optimal control technique, but where it differs from the traditional optimal control method is that it solves the standard optimal control problem on-line in a finite horizon, rather than determining off-line a feedback law. Also, when the on-line solution is obtained, MPC typically sends out the first control action to be implemented, and repeats the calculation at the next instant. The advantages of MPC are that it can handle constraints, owns the ability of prediction especially when there exists time delay, and to some extent overcomes the effect of uncertainty on the systems.

Motivated by the above discussions, in this paper, we consider the optimization and control problem of chaotic systems with uncertain parameters and stochastic disturbance by the Lyapunov-based model predictive control [10]–[16]. By using a previously designed Lyapunov controller, the stability is discussed and proved by stochastic Lyapunov stability theory. As a typical example, the unified chaotic system with the uncertain parameter and stochastic disturbance is considered and simulation results show the effectiveness of the proposed method.

## II. Control of chaotic systems with uncertain parameters and stochastic disturbance by LMPC

In many MPC formulations there often exist two important issues that how to guarantee the closed-loop stability and initial conditions starting from where the control is feasible. In order to solve these problems, the Lyapunov-based model predictive control(LMPC) is presented [10]–[16]. The idea of LMPC is that a Lyapunov-based controller $h(x)$ is designed previously, and the closed-loop stability and the optimization are based on the controller $h(x)$. With the help of the controller $h(x)$, the stability of LMPC can be inherited and the region of optimization feasibility can be explicitly characterized.

Let us consider the control problem of the following chaotic systems

$$\dot{x}(t) = f_0(x(t)) + f_1(x(t))\theta(x(t)) + l(x(t))\xi(t) + g(x(t))u(t) \tag{1}$$

where $f_0(0) = f_1(0) = l(0) = g(0) = 0$, namely the origin is the equilibrium point; $\theta(x(t))$ is the uncertain parameter and there exists $\theta_b > 0$ such that $\|\theta(x(t))\| \leq \theta_b$; $\xi(t)$ denotes

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

the standard Gaussian white noise which can be expressed as the formal derivative of Wiener process $w(t)$; $u(t)$ is the controller. We suppose that all the states of the system are available, $f_0(x), f_1(x), \theta(x), l(x), g(x)$ are continuous functions, and compared to the originally deterministic system $f_1(x(t))\theta(x(t)) + l(x(t))\xi(t)$ can be viewed as the small disturbance.

For a candidate Lyapunov function $V(x)$, the following feedback control law $u(t) = h(x(t))$ can be constructed

$$h(x) = \begin{cases} 0, & if\ L_gV(x) = 0 \\ -\frac{\omega+\sqrt{\omega^2+(L_gV(L_gV)^T)^2}}{L_gV(L_gV)^T}(L_gV)^T, & else \end{cases} \tag{2}$$

where $\omega = L_{f_0}V(x) + ||L_{f_1}V(x)||\theta_b + \frac{1}{2}Trace\left\{l^T\frac{\partial V^2}{\partial x^2}l\right\} + \rho V(x)$, $\rho > 0$, $L_{f_i}V(x) = \frac{\partial V(x)}{\partial x}f_i(x)(i = 0, 1)$ and $L_gV(x) = \frac{\partial V(x)}{\partial x}g(x)$. It can be investigated that if $V(x)$ is a control Lyapunov function, then $h(x)$ is optimal in some sense [16]–[18].

The LMPC can be designed as follows

$$\min_{u(\tau)\in S(\Delta)} \int_{t_k}^{t_k+T} (\hat{x}^T(\tau)Q\hat{x}(\tau) + u^T(\tau)Ru(\tau))d\tau \tag{3}$$

$$\dot{\hat{x}}(\tau) = f_0(\hat{x}(\tau)) + g(\hat{x}(\tau))u(\tau) \tag{4}$$

$$\hat{x}(t_k) = x(t_k) \tag{5}$$

$$\frac{\partial V(x(t_k))}{\partial x}g(x(t_k))u(t_k) \le \frac{\partial V(x(t_k))}{\partial x}g(x(t_k))h(x(t_k)) \tag{6}$$

where $S(\Delta)$ is the family of piece-wise constant functions with sampling period $\Delta$, which means that the controller is applied in a sample-and-hold fashion. $Q, R$ are positive definite weight matrices. LMPC is unnecessary to use a terminal penalty term, but needs an auxiliary Lyapunov-based control law $h(x)$ that gives the contractive constrains of the Lyapunov-based model predictive controller. With the help of the controller $h(x)$, the initial feasibility of the optimization is satisfied automatically. In the following section, we will prove the stability of the system (1) under the control action $u(t_k)$, which further implies that the optimization is consecutively feasible.

In order to make the theory analysis easily, we give the following assumption:

Assumption 1:

(1) $\|\Psi(y, \theta(y), u) - \Psi(x, \theta(x), u)\| \le \eta_\Psi \|y - x\|$, $\eta_\Psi > 0$. where $\Psi(x, \theta(x), u) = L_{f_0}V(x) + L_{f_1}V(x)\theta(x) + \frac{1}{2}Trace\left\{l(x)^T\frac{\partial V^2}{\partial x^2}l(x)\right\} + L_gV(x)u(t)$.

(2) $\|F(x, \theta(x), u)\| \le M_1$,$\|l(x(t))\| \le M_2$, $(M_1 > 0, M_2 > 0)$. where $F(x, \theta(x), u) = f_0(x) + f_1(x)\theta(x) + g(x)u(t)$.

Remark 1: For the chaotic system, its attractor is a bounded compact set. When the control input is added and not too large, usually the boundedness can be kept. Based on the properties that the functions $f_0(x), f_1(x), \theta(x), l(x), g(x)$ are continuous, then we can conclude the assumption (2) holds. Finally,if the function $\Psi(x, \theta(x), u)$ is further differential in $x$ for each $u$, we can obtain $\Psi(x, \theta(x), u)$ satisfies the Lipschitz property in the assumption(1). Thus, through the above discussion we know that Assumption 1 is not difficult to be satisfied.

*Lemma 1:* If $\dot{x}(t) = F(x(t), \theta(x(t)), u(t_k)) + l(x(t))\xi(t), t \in [t_k, t_{k+1})$, where $F(x(t), \theta(x(t)), u(t_k)) = f_0(x(t)) + f_1(x(t))\theta(x(t)) + g(x(t))u(t_k))$, then there exists a constant $\gamma > 0$ such that the following inequality holds $E\|x(t) - x(t_k)\| \le \gamma\sqrt{\Delta}$.

Proof.

$$\dot{x}(t) = F(x(t), \theta(x(t)), u(t_k)) + l(x(t))\xi(t),\ t \in [t_k, t_{k+1})$$

Integrating the above formula, we get $x(t) - x(t_k) = \int_{t_k}^{t} F(x(\tau), \theta(x(\tau)), u(t_k))d\tau + \int_{t_k}^{t} l(x(\tau))dw$

Applying the inequality $\|a + b\|^2 \le 2\|a\|^2 + 2\|b\|^2$, then

$$\begin{aligned} &E\left(\|x(t) - x(t_k)\|^2\right) \\ &= E\left\|\int_{t_k}^{t} F(x(\tau), \theta(x(\tau)), u(t_k))d\tau + \int_{t_k}^{t} l(x(\tau))dw\right\|^2 \\ &\le 2E\left\|\int_{t_k}^{t} F(x(\tau), \theta(x(\tau)), u(t_k))d\tau\right\|^2 \\ &\quad +2E\left\|\int_{t_k}^{t} l(x(\tau))dw\right\|^2 \end{aligned}$$

Using Cauchy-Schwarz inequality and Assumption 1, we can obtain

$$\begin{aligned} &E\left(\left\|\int_{t_k}^{t} F(x(\tau), \theta(x(\tau)), u(t_k))d\tau\right\|^2\right) \\ &\le (t - t_k)E\left(\int_{t_k}^{t} \|F(x(\tau), \theta(x(\tau)), u(t_k))\|^2 d\tau\right) \\ &= (t - t_k)\int_{t_k}^{t} E\left(\|F(x(\tau), \theta(x(\tau)), u(t_k))\|^2\right)d\tau \\ &\le M_1^2(t - t_k)^2 \le M_1^2\Delta^2 \end{aligned}$$

$$\begin{aligned} &E\left(\left\|\int_{t_k}^{t} l(x(\tau))dw\right\|^2\right) = E\left(\int_{t_k}^{t} \|l(x(\tau))\|^2 d\tau\right) \\ &\le M_2^2(t - t_k) \le M_2^2\Delta \end{aligned}$$

Thus

$$E\left(\|x(t) - x(t_k)\|^2\right) \le 2(M_1^2\Delta + M_2^2)\Delta$$

If we let $\gamma^2 = 2(M_1^2\Delta + M_2^2)$, then by Jensen's inequality we get

$$\begin{aligned} &E\|x(t) - x(t_k)\| = E\sqrt{\|x(t) - x(t_k)\|^2} \\ &\le \sqrt{E\left(\|x(t) - x(t_k)\|^2\right)} \le \gamma\sqrt{\Delta}. \end{aligned}$$

*Theorem 1:* Consider the trajectory $x(t)$ of the system (1) under the control law $u(t)$, which satisfies the conditions of Assumption 1 and is implemented in a sample-and-hold fashion:

$$\dot{x}(t) = f_0(x(t)) + f_1(x(t))\theta(x(t)) + l(x(t))dw + g(x(t))u(t_k) \tag{7}$$

where $t \in [t_k, t_{k+1})$ and $t_k = t_0 + k\Delta, k = 1, 2, ...$

Then the origin of the system (1) is practically stable in some mean sense.

Proof. The time derivative of the Lyapunov function $V(x)$ along the trajectory $x(t)$ of the system (1) in $t \in [t_k, t_{k+1})$ is given by

$$\dot{V}(x(t)) = \Psi(x(t), \theta(x(t)), u(t_k)) + \frac{\partial V(x(t))}{\partial x} l(x(t))dw$$

Adding and subtracting $\Psi(x(t_k), \theta(x(t_k)), u(t_k))$, and taking into account Assumption 1 and the constraint (6), we obtain

$$\begin{aligned}
&\dot{V}(x(t)) \leq -\rho V(x(t_k)) + \Psi(x(t), \theta(x(t)), u(t_k)) \\
&-\Psi(x(t_k), \theta(x(t_k)), u(t_k)) + \tfrac{\partial V(x(t))}{\partial x} l(x(t))dw \\
&\leq -\rho V(x(t_k)) + \eta_\Psi ||x(t) - x(t_k)|| + \tfrac{\partial V(x(t))}{\partial x} l(x(t))dw
\end{aligned}$$

Taking the expectation of the above inequality and using Lemma 1, it leads to

$$\begin{aligned}
&E\dot{V}(x(t)) \\
&\leq -\rho EV(x(t_k)) + \eta_\Psi E||x(t) - x(t_k)|| \\
&\leq -\rho EV(x(t_k)) + \gamma\eta_\phi \sqrt{\Delta}
\end{aligned}$$

When the right side of the above inequality is less than zero, we know that the value of $EV(x(t))$ will decrease. Therefore, if we take $r_1 > r_2 > 0$ and a small constant $\varepsilon > 0$ such that $r_2 = (\varepsilon + \gamma\eta_\Psi\sqrt{\Delta})/\rho$, $\Omega_{r_1} = \{x : EV(x(t)) \leq r_1\}$, $\Omega_{r_2} = \{x : EV(x(t)) \leq r_2\}$, then for $x(t) \in \Omega_{r_1}/\Omega_{r_2}$, the state will converge to $\Omega_{r_2}$ in some mean sense. If we further take $r_{\min} < r_1$ and $r_{\min} = \max\limits_{\Delta_1 \in [0,\Delta]} \{EV(x(t + \Delta_1)) : EV(x(t)) \leq r_2\}$, then once the state converge to $\Omega_{r_2} \subseteq \Omega_{r_{\min}}$, it will remains inside $\Omega_{r_{\min}}$ for all times. That is to say $\lim\limits_{t\to\infty} supEV(x(t)) \leq r_{\min}$.

## III. Example

Let us consider the control problem of the unified chaotic system with the uncertain parameter $\theta$ and stochastic disturbance $l(x)\xi(t)$.The system (8) is called the unified chaotic system when the right of (8) only includes $f(x(t), \theta)$ and $\theta \in [0, 1]$. As shown in [19], if $\theta \in [0, 0.8)$, it is called the generalized Lorenz chaotic system; if $\theta = 0.8$, it is called Lu chaotic system; if $\theta \in (0.8, 1]$, it becomes the generalized Chen system. Therefore, the parameter $\theta$ plays an important role in the unified chaotic system, and it's natural to discuss the control problem when the parameter $\theta$ is uncertain [20], [21]. Furthermore, stochastic disturbance exists inevitably in practice. It is necessary to discuss the control problem of the unified chaotic system with uncertain parameter and stochastic disturbance.

$$\dot{x}(t) = f(x(t), \theta) + l(x(t))\xi(t) + Bu(t) \tag{8}$$

where

$$f(x, \theta) = \begin{pmatrix} (25\theta + 10)(x_2 - x_1) \\ (28 - 35\theta)x_1 + (29\theta - 1)x_2 - x_1x_3 \\ -(8 + \theta)x_3/3 + x_1x_2 \end{pmatrix} \tag{9}$$

$$l(x(t))dw = \begin{pmatrix} \sigma_1 x_1 & 0 & 0 \\ 0 & \sigma_2 x_2 & 0 \\ 0 & 0 & \sigma_3 x_3 \end{pmatrix} \begin{pmatrix} dw_1 \\ dw_2 \\ dw_3 \end{pmatrix} \tag{10}$$

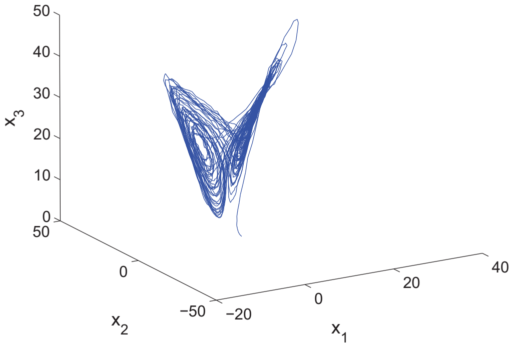


Fig. 1. Phase portrait of chaotic attractor in $(x_1, x_2, x_3)$ space.

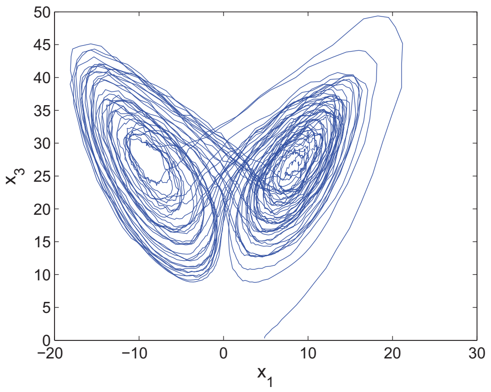


Fig. 2. Projective portrait of chaotic attractor in $(x_1, x_3)$ plane.

$$B = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \tag{11}$$

In this example, we take $l(x)$ and $B$ as (10) and (11), respectively. Choosing $V(x) = \frac{1}{2}(x_1^2 + x_2^2 + x_3^2)$, we get $\Psi(x, \theta, u) = a(x, \theta) + x_2 u$, where

$$\begin{aligned}
&a(x, \theta) = (x_1, x_2, x_3)f(x, \theta) + \tfrac{1}{2}(\sigma_1^2 x_1^2 + \sigma_2^2 x_2^2 + \sigma_3^2 x_3^2) \\
&= (38 - 10\theta)x_1x_2 + (29\theta - 1)x_2^2 - (10 + 25\theta)x_1^2 - \tfrac{8+\theta}{3}x_3^2 \\
&+ \tfrac{1}{2}(\sigma_1^2 x_1^2 + \sigma_2^2 x_2^2 + \sigma_3^2 x_3^2)
\end{aligned}$$

when $x_2 = 0, x \neq 0$, and $\theta \in [0, 1], \sigma_1 < 2\sqrt{5}, \sigma_3 < 4/\sqrt{3}$,

$$\alpha(x, \theta) = -(25\theta + 10 - \frac{1}{2}\sigma_1^2)x_1^2 - (\frac{8 + \theta}{3} - \frac{1}{2}\sigma_3^2)x_3^2 < 0$$

Thus, $V(x) = \frac{1}{2}(x_1^2 + x_2^2 + x_3^2)$ is a control lyapunov function [17], and the auxiliary controller $h(x)$ is taken as (2).

In numerical simulations, we take the initial value $x(0) = (2, -1, 1)^T$ and the parameters $\sigma_1 = \sigma_2 = \sigma_3 = 0.1, \theta(x) = 0.1\,|\sin(x_1)|\,, \rho = 0.001, Q = I, R = 1.0, \Delta = 0.02, T = 100\Delta$, to investigate the proposed method. Figs. 1-2 displays the phase portrait of the uncertain and stochastic chaotic attractor when the system (8) is without control input. When

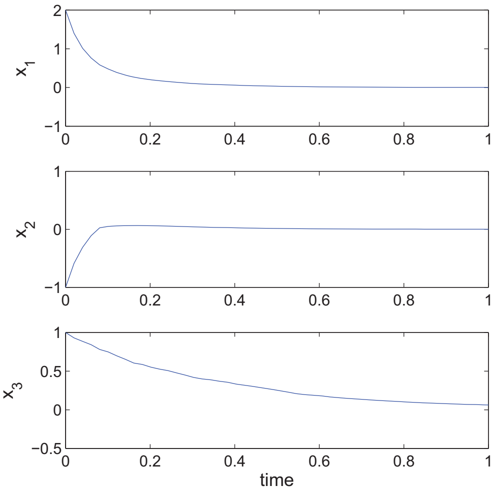


Fig. 3. The evolution of system states under LMPC.

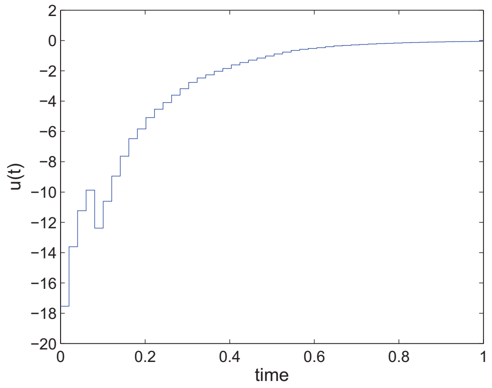


Fig. 4. The varying curve of control input by LMPC.

the control law of LMPC is applied to the chaotic system, from Figs. 3-4 one can see that the state trajectories of the chaotic system and the control action will approach to the neighborhoods of zero points as time increases. These results investigate the effectiveness of the proposed method for the chaotic systems with uncertain parameters and stochastic disturbance.

## IV. Conclusion

Chaos is usually undesirable in real systems, therefore many methods are proposed to control it. In this paper, for the chaotic systems with uncertain parameters and stochastic disturbance, we discussed the chaotic control problem by using the Lyapunov-based model predictive control (LMPC). Through introducing the auxiliary control law $h(x)$, the stability of LMPC is discussed and proved by stochastic Lyapunov stability theory. Compared to other chaotic control schemes, the advantage of the proposed method is that the optimality can be considered and guaranteed as well as the close-loop stability. Simulation results show the effectiveness of the proposed method.

## References

[1] E. Ott, C. Grebogi, and J. A. Yorke, “Controlling chaos,” *Phys. Rev. Lett.*, vol. 64, no. 11, pp.1196-1199, 1990.

[2] S. Boccaletti, C. Grebogi, Y. C. Lai, H. Mancini, and D. Maza, “The control of chaos: theory and applications,” *Physics Report*, vol. 329, no. 3, pp. 103-197, 2000.

[3] K. Pyragas, “Continuous control of chaos by self-controlling feedback,” *Phys. Lett. A*, vol. 170, pp. 421-428, 1992.

[4] S. Sinha, R. Ramaswamy, and J. Rao, “Adaptive control in nonlinear dynamics,” *Phys. D*, vol.43, pp.118-128,1990.

[5] G. Chen and X. Dong, “From Chaos to order: perspectives and methodologies in controlling chaotic nonlinear dynamical system,” *Int. J. Bifur. Chaos*, vol. 3, no. 6, pp. 1363-1369, 1993.

[6] C. C. Hwang, J. Y. Hsieh, and R. S. Lin, “A linear continuous feedback control of Chua’s circuit,” *Chaos Solit. Fract.*, vol. 8, no. 9, pp. 1507-1515, 1997.

[7] H. Ren and D. Liu, “Nonlinear feedback control of chaos in permanent magnet synchronous motor,” *IEEE Trans. Circuits Syst. II*, vol. 53, no. 1, pp. 45-50, 2006.

[8] S. C. Sinha, J. T. Henrichs, and B. Ravindra, “A general approach in the design of active controllers for nonlinear systems exhibiting chaos,” *Int. J. Bifur. Chaos*, vol. 10, no. 1, pp. 165-178, 2000.

[9] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. M. Scokaert, “Constrained model predictive control: Stability and optimality,” *Automatica*, vol. 36, pp. 789-814, 2000.

[10] P. Mhaskar, N. H. El-Farra, and P. D. Christofides, “Predictive control of switched nonlinear systems with scheduled mode transitions,” *IEEE Trans. Automat. Control*, vol. 50, no. 11, pp. 1670-1680, Nov. 2005.

[11] P. Mhaskar, N. H. El-Farra, and P. D. Christofides, “Stabilization of nonlinear systems with state and control constraints using Lyapunovbased predictive control,” *Syst. Control Lett.*, vol. 55, pp. 650-659, 2006.

[12] P. Mhaskar, N. H. El-Farra, and P. D. Christofides. “Robust predictive control of switched systems: Satisfying uncertain schedules subject to state and control constraints,” *Int. J. Adapt. Control Signal Process.*, vol. 22, pp.161-179, 2008.

[13] D. Munoz de la Pena and P. D. Christofides, “Lyapunov-Based Model Predictive Control of Nonlinear Systems Subject to Data Losses,” *IEEE Trans. Automat. Control*, vol. 53, no. 9, pp. 2076-2089, Oct 2008.

[14] J. F. Liu, X. Z. Chen, D. Munoz de la Pena, P. D. Christofides,“ Iterative Distributed Model Predictive Control of Nonlinear Systems: Handling Asynchronous, Delayed Measurements,” *IEEE Trans. Automat. Control*, vol.57, no.2, pp.528-534, 2012.

[15] X. Z. Chen, M. Heidarinejad, J.F. Liu,P. D. Christofides,“Distributed economic MPC: Application to a nonlinear chemical process network,” *Journal of Process Control*, vol.22, pp.689-699, 2012.

[16] J. A. Primbs, V. Nevistic, and J. C. Doyle, “A receding horizon generalization of pointwise min-norm controllers,” *IEEE Trans. Automat. Control*, vol. 45, no. 5, pp. 898-909, May 2000.

[17] E. Sontag, “A ’universal’ construction of arstein’s theorem on nonlinear stabilization,” *Syst. Control Lett.*, vol. 13, pp. 117-123, 1989.

[18] R. Freeman, P. Kokotovic, “Inverse optimality in robust stabilization,” *SIAM J. Contr. Optimiz.*, vol. 34, pp. 1365-1391, July 1996.

[19] J. H. Lu, G. R. Chen, D. Z. Cheng, S. Celikovsky, “Bridge the gap between the Lorenz system and the Chen system,” *Int. J. Bifur. Chaos*,vol.12, no.11, pp. 2917-2926, 2002.

[20] G. Cai, W. Tu, “Adaptive Backstepping Control of the Uncertain Unified Chaotic System,” *Int. J. Nonlinear Science*, vol.4, no.1, pp.17-24, 2007.

[21] W. Zhang,Y. Liang, H. S. Su, Z. Q. C. Michael, Z. Z. Han, “Robust stabilization of a class of nonlinear systems with uncertain parameters based on CLFs,” in *Proceedings of the 30th Chinese Control Conference*, Yantai, China, 2011, pp. 2275-2279.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Hybrid Geno-Fuzzy Controller For Seismic Vibration Control

Monica Patrascu
Dept. of Automatic Control and Systems Engineering
University Politehnica of Bucharest
Bucharest, Romania
e-mail: monica.patrascu@acse.pub.ro

Ioan Dumitrache
Dept. of Automatic Control and Systems Engineering
University Politehnica of Bucharest
Bucharest, Romania
e-mail: ioan.dumitrache@acse.pub.ro

*Abstract*— **This paper evaluates the possibility of applying a geno-fuzzy control strategy to a magnetorheological semi-active damper for seismic vibration control. The proposed control starategy is designed and then tested and validated in a simulated environment. The control strategy is validated by considering a more destructive seismic disturbance as input for the damper-structure system. The proposed geno-fuzzy hybrid controller offers improved performance and implementability for real-time applications.**

***Keywords:*** *semi-active control, fuzzy controller, magnetorheological fluid damper, seismic response, genetic algorithms, hybrid geno-fuzzy controller.*

## I. Introduction

Earthquake induced vibration of civil structures is a current topic of high interest. As buildings rise and materials are sought to be lighter and cheaper, the limitations of structural design are visible in buildings subjected to seismic and wind disturbance. A common approach is using additional structural vibration damping devices, be they passive, active, or semi-active. The technological advancements in computer and information technology make possible the real-time implementation of complex, advanced or non-linear control systems.

This article focuses on the use of magnetorheological (MR) dampers for semi-active seismic vibration control, by means of a hybrid geno-fuzzy non-linear controller. Using MR actuators for their dissipative properties is not a new idea. However, the applied control strategies vary in structure and performances. MR and ER (electrorheologic) bracing systems' analysis of seismic response with a linear controller based on ground acceleration is performed in [1], the authors showing showed that placing dampers near the base of a structure, as opposed to the upper levels, gives a better response reduction. Instead, in [2] a non-linear fuzzy control system is successfully employed for a base isolation MR damper.

Due to the non-linear intrinsic character of the MR dampers and of the fuzzy controllers, researchers have attempted evolutionary optimization of the fuzzy control systems for active and semi-active dampers. A base isolation system is designed in [3], combining a MR damper with a passive friction pendulum system. The authors have obtained a neuro-fuzzy model of the MR damper and have performed a genetic algorithm optimization of the fuzzy rules and the associated membership functions' parameters. A decrease in base displacements was observed, without increasing the acceleration, as seen in fully passive systems.

Another base isolation system was developed for a benchmark structure using MR dampers and a fuzzy controller, in [4]. Numerically simulated genetic algorithms were used to optimize the control system's performance. The genetic algorithm chromosomes have been used to code the membership functions' parameters, as well as a set of weighting factors for the fuzzy rules, the authors obtaining structural response reduction in both the base and the supra-structure. The evolutionary optimization of fuzzy controllers is also applied to an active tuned mass damper (ATMD) system in [5], where a genetic algorithm was used to optimize membership functions, rule weighting coefficients and three of the ATMD parameters.

In recent years, development of optimized fuzzy control systems for structural vibration control has taken a turn towards implementation of these algorithms. Thus, in [6], genetic algorithms are used to optimize model parameters of a benchmark building and a fuzzy model of a MR damper is obtained. Genetic algorithms are also used to optimize the fuzzy controller rules, that are considered to be described by membership functions and their parameters, reaching a chromosome length of 80 to 100 genes.

Computation time was still an issue in [7], where genetic algorithms and particle swarm optimization are used for the optimization of a fuzzy rule base and membership functions' parameters. The considered damping system was a semi-active MR actuator, obtaining a computing time between 2 and 60 hours for the tuning of the fuzzy controller.

Genetic algorithms also are used for determining the rule base of a fuzzy control system for seismic vibration control with an MR semi-active damper in [8]. In comparison with an adaptive controllers, the authors have obtained a better response reduction of the structural system from the optimized fuzzy controller.

This paper is organized as follows: in section 2, the authors present the models for the magnetorheological damper and

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

structure, along with the principle of semi-active vibration control. In section 3, the proposed control strategy is presented, while section 4 contains a case study along with its results. Finally, section 5 contains the conclusions.

## II. MAGNETORHEOLOGICAL SEMI-ACTIVE DAMPER

### *A. Magnetorheological Damper Model*

A magnetorheological damper is a hydraulic-class actuator used in seismic protection design [9]. This device is used to generate the necessary control forces using as an input a command current and the velocity of the story on which it is mounted. The damper is a hydraulic cylinder filled with MR fluid – a magnetically polarizable micron size suspension in oil or other fluids [2]. The damping coefficient of this device is controlled by varying a magnetic field, thus changing the fluid from viscous to semi-solid in milliseconds. In this specific case, the command signal is a current between *0-2 A*.

In order to obtain a satisfactory vibration control of a given structure, it is necessary that the actuators perform inside a set of strict performance criteria, such as response time and robustness versus uncertainties. The output force generated by these dampers is required to be maintained into specific limits, as not to induce instability into the structure. Therefore, a control loop for the actuator is required. In this paper, a magnetorheological damper (figure 1) is controlled via a geno-fuzzy hybrid controller.

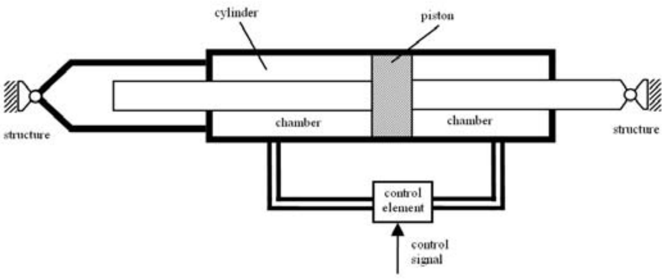


Figure 1. Magnetorheological damper.

A MR (magnetorheological) damper is analyzed. Due to uncertainties in modeling of the damper, a two velocity domain approximation is performed. The behavior of the semi-active MR damper used in this paper has is given by [10]:

$$\left\{\begin{array}{c} f = c_0\dot{x} + \alpha z \\ \dot{z} = -\gamma|\dot{x}||z||z|^{n\text{-}1} - \beta\dot{x}|z|^{n} + A\dot{x} \\ \alpha = \alpha_a + \alpha_b u \\ c_0 = c_{0a} + c_{0b}u \\ \dot{u} = -\eta(u - v) \end{array}\right. \tag{1}$$

where $f$ is the force generated by the damper, $u$ is the command signal and $v$ is the command voltage. The control element is modeled by the fourth equation with first order dynamic. This research makes use of the following parameter values [11]: $c_{0a}$ = 0.0064 Ns/cm, $c_{0b}$ = 0.0052 Ns/cmV, $a_a$ = 8.66 N/cm, $a_b$ = 8.86 N/cmV, $\gamma$ = 300cm$^{-2}$, $\beta$ = 300cm$^{-2}$, A = 120, n = 2 and $\eta$ = 80s$^{-1}$.

### *B. Structure Model*

The damper is considered to be mounted in the base of a building. Thus, the equation used to describe a structural system with a damper that is subject to earthquake disturbances is:

$$M\ddot{x} + C\dot{x} + Kx = F_d + F_u \tag{2}$$

where $x$ is a vector containing the displacements $x_i$ of each story, $F_d$ is the force induced by the earthquake and $F_u$ is a vector containing the control forces for each story. Due to physical connection of each damper, this vector consists of pairs of control forces, equal in value, but in opposing directions, each pair corresponding to one damper. *M*, *C* and *K* are mass, damping and stiffness matrices, respectively, as follows: *M* is a diagonal matrix that contains the stories' masses, *C* and *K* are tri-diagonal matrices describing the codependency between adjacent stories [12].

### *C. Semi-Active Vibration Control*

Semi-active vibration control allows the damping system behaves as a passive one while the structural vibration remains in specified constraints, otherwise the control is active. Advantages of these type of actuators show low costs and little need for auxiliary power sources. Semi-active dampers can efficiently respond with precision, to strong wind or damaging earthquakes. The necessary control forces are generated based on the information received from the sensor distribution throughout the structure. The performance levels are comparable to the ones offered by active control strategies, without their major drawbacks and with minimal risk to generate unstable behaviour. The principle of the semi-active control is presented in the figure 2. Earthquake induced ground motion information is submitted to a switch module which disables or enables the passive and active modes accordingly.

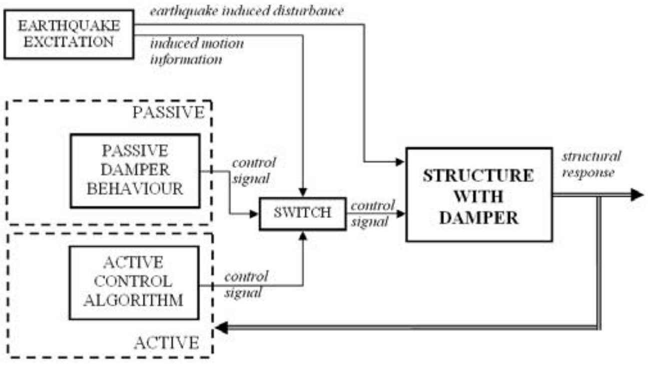


Figure 2. Semi-active control principle.

## III. PROPOSED CONTROL STRATEGY

Fuzzy sets are generally used to reflect vague information of the real world environment. Due to model nonlinearities and uncertainties, it is difficult to accurately describe the structural response or even the damper behavior. Nevertheless, fuzzy logic presents a viable solution, its main advantage being high performance when dealing with non-linear systems and vague information.

To this extent, the fuzzy controller implemented in this paper is constructed with a set of rules that would have been extracted from human experience. The authors propose an evolutionary learning strategy for the fuzzy controller. Because of the non-linear nature of these controllers, genetic algorithms (GAs) are used for tuning the existing "knowledge" of the fuzzy controller. The general strategy is presented in [13]. Figure 3 illustrates the control architecture proposed in this paper, where: $x$ is the displacement of the structure; $\dot{x}$ and $\ddot{x}$ are the velocity and acceleration of the structure, respectively; $F$ is the control force; $i$ is the command current, , $i_P$ is the passive command current, $i_A$ is the active command voltage; $I$ is the performance index for the genetic algorithm; $\Phi$ represents a vector containing the fuzzy controller parameters that are optimized via genetic algorithm; $a$ and $v$ are the earthquake induced ground acceleration and velocity, respectively.

The proposed strategy is structured on two levels. One level deals with damper control and building displacement reduction, while the other level deals with tuning the implemented fuzzy controller. The scheme functions online only through the first level, GA tuning being an offline procedure. The output is a generated control force. The main loop implements a hybrid geno-fuzzy controller.

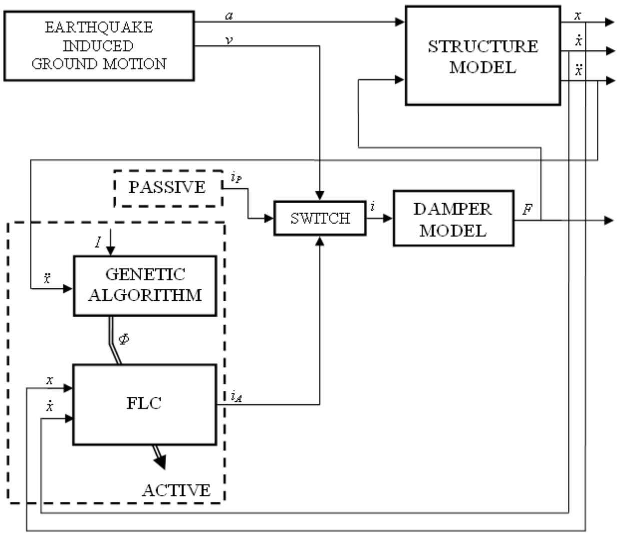


Figure 3. The proposed control architecture.

Fuzzy logic controllers (FLCs) make use of linguistic terms to describe process variables. The main characteristic of fuzzy systems is the use of vague notions instead of crisp numerical values. The linguistic information is represented through words like *small*, *medium*, *large* etc, and makes for a valuable tool in control systems design when the numerical information is incomplete. A FLC is comprised of a fuzzification module, a rule base, an inference mechanism, and a defuzzification module. The fuzzification module transforms the crisp variable into their vague counterparts, while the defuzzifier performs the inverse operation, allowing the controller to generate commands that are compatible with the real life plant. The inference mechanism retrieves the rules from the rule base according to the current inputs of the controller.

Genetic algorithms (GAs) are evolutionary optimization algorithms, also suited for searching problems, specifically when the solution pool is vast and the information is little. GAs are able to perform multi-dimensional searches, even in conditions of uncertainty, and impartial to the strictness of the constraints. The main steps of a GA [14] require first an initialization of the solution pool, knows as a population. Then, for each generation, each individual is evaluated based on a problem-specific fitness function and thus is either selected or eliminated from the current population, followed by a recombination of the selected individuals. The loop repeats until a termination condition is met, either considering a maximum number of generations, or a specified population fitness. The fitness function models the objective of the algorithm, its purpose being the rejection of unfit/unwanted solutions from the gene pool.

The genetic learning of fuzzy controllers is a fine tuning procedure for the parameters of the latter. The procedure requires two steps. First, an algebraic model (AM) for the fuzzy controller is obtained, by fitting a non-linear function over the fuzzy input-output dependency using GAs. Next, the AM's coefficients are tuned in closed loop in order to obtain a better system response, again by means of GAs. The structure of a hybrid geno-fuzzy controller is presented in figure 4, pointing out the inputs and outputs of the closed loop tuning genetic algorithm.

The algebraic model (AM) of the FLC has the following general structure:

$$\delta = \sum_{i=1}^{N} \frac{f_{i0} + f_{i1} \cdot \varepsilon + f_{i2} \cdot d\varepsilon}{f_{i3} + f_{i4} \cdot \varepsilon + f_{i5} \cdot d\varepsilon} \tag{3}$$

where $\delta$ is the controller output (command), $\varepsilon$ is the control deviation between the setpoint and the system's output, $d\varepsilon$ is the derivative of $\varepsilon$, and $N$ is the number of terms in the AM, dependant on the complexity of the designed FLC.

The parameters $f$ in the model represent the algebraic model coefficients. In order to find the AM coefficients, a GA was implemented that minimizes a performance index $I$ as the sum of the square errors between the AM output $\delta$ and the actual fuzzy controller output.

Further tuning of the AM is required, by means of GA. This time around, the GA searches the adjiacent space of the previously found solutions in order to obtain a better performance from the control system. Since the AM has numerical coefficients, the tuning procedure is much faster than if human operator would apply modifications to the fuzzy membership functions or rulebase.

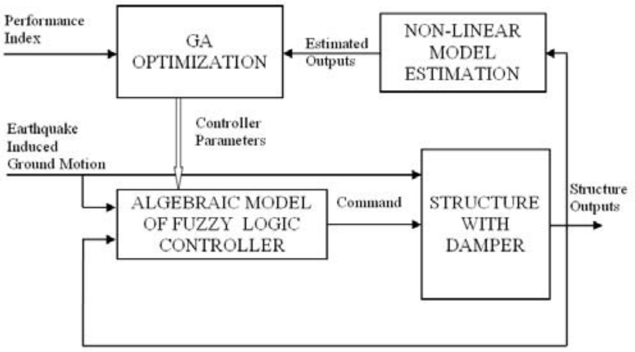
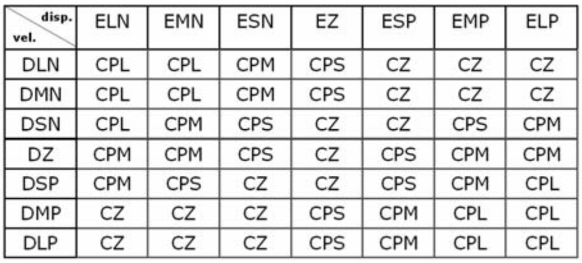


Figure 4. Geno-fuzzy controller.

## IV. CASE STUDY

The case study presented here implements the proposed control architecture and analyzes the hybrid geno-fuzzy controller's performance in conditions of model uncertainties and external disturbances. The authors considered a base mounting of the MR damper. Since the destructive effect of the earthquake is a result of its horizontal vibrational components, the vertical load on the ground story is not taken into account, and all spring and damper structural elements are considered in a horizontal direction. The masses cover both the floors and the associated walls.

This paper implements a geno-fuzzy control strategy for seismic vibration mitigation. Considering non-linearities and model uncertainties, the authors approached a hybrid geno-fuzzy controller, which was included in a semi-active control strategy. The fuzzy controller input variables are the displacement and velocity of the structure and the output is the command voltage used for the elecrohydraulic damper. The discourse universes for each input variable are normalized to [-1, 1], while the output is generated in the interval [0, 5]. The scaling factors used were obtained by analyzing the structure output. The input and output variables and the rulebase are presented in figure 5, in which all membership functions are triangular, with a 50% overlap. The linguistic terms are coded as follows: E - displacement (deviation from zero), D - velocity (derivative of first input), C - command (output), N - negative, P- positive, L - large, M - medium, S - small, Z - zero.



| disp. / vel. | ELN | EMN | ESN | EZ | ESP | EMP | ELP |
|---|---|---|---|---|---|---|---|
| DLN | CPL | CPL | CPM | CPS | CZ | CZ | CZ |
| DMN | CPL | CPL | CPM | CPS | CZ | CZ | CZ |
| DSN | CPL | CPM | CPS | CZ | CZ | CPS | CPM |
| DZ | CPM | CPM | CPS | CZ | CPS | CPM | CPM |
| DSP | CPM | CPS | CZ | CZ | CPS | CPM | CPL |
| DMP | CZ | CZ | CZ | CPS | CPM | CPL | CPL |
| DLP | CZ | CZ | CZ | CPS | CPM | CPL | CPL |

Figure 5. Rulebase of the fuzzy controller.

The control methods described above have been simulated together with a 3-story structure. For the design of the hybrid geno-fuzzy controller, the structure is modeled as a system which integrates the damper in its base (figure 6).

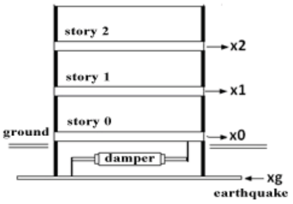


Figure 6. Base isolation of the structure.

The three story building has the following parameters:

$$\begin{cases} M = \begin{bmatrix} 98.3 & 0 & 0 \\ 0 & 98.3 & 0 \\ 0 & 0 & 98.3 \end{bmatrix} [kg] \\ C = \begin{bmatrix} 175 & -50 & 0 \\ -50 & 100 & -50 \\ 0 & -50 & 50 \end{bmatrix} [Ns/m] \\ K = 10^5 \begin{bmatrix} 12 & -6.84 & 0 \\ -6.84 & 13.7 & -6.84 \\ 0 & -6.84 & 6.84 \end{bmatrix} [N/m] \end{cases} \tag{4}$$

Details of this model are given in [15].

A set of evaluation criteria [8] has been chosen:

$$\begin{cases} J_1 = \dfrac{\max|x_i(t)|}{x_{open}} \\ J_2 = \dfrac{\max|d_i(t)|}{d_{open}} \\ J_3 = \dfrac{\max|\ddot{x}_i(t)|}{\ddot{x}_{open}} \end{cases} \tag{5}$$

where $x_i(t), \ddot{x}_i(t)$ are the relative displacement, acceleration of the *i*-th story, while $d_i(t)$ is the interstory drift and the notation *open* designates the overall maximum absolute displacements, accelerations and drifts of the uncontrolled structure.

The control system of the MR damper is simulated by means Matlab (Simulink). Figure 8 presents the response of the MR damper-structure system when excited with an earthquake input signal (Northridge 1994 - the earthquake accelerogram is presented in figure 7), controlled with the geno-fuzzy controller vs. the uncontrolled response of the ground story. The performance criteria is to reduce ground story movement.

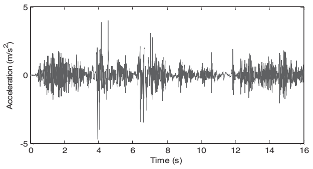


Figure 7. Northridge earthquake accelerogram.

For the fuzzy controller designed, the maximum controlled displacement and acceleration was observed to be 15.22% lower than for the open loop simulation.

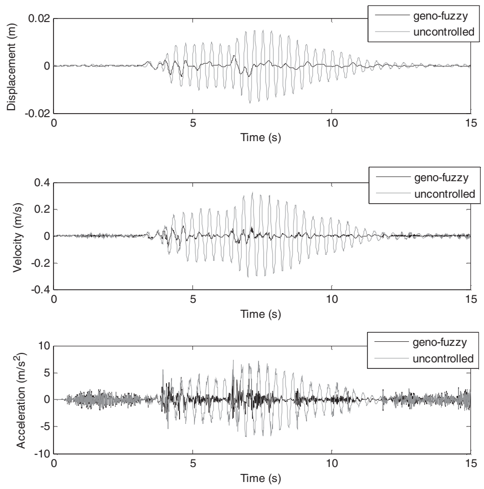


Figure 8. Comparative base structural responses: geno-fuzzy.

The fuzzy logic controller (FLC) designed in this paper has a set of rules, with a rule base created based on human experience. The AM obtained in this paper is:

$$\delta = \frac{0.113+0.415\cdot d+0.743\cdot v}{-0.4+0.843\cdot d+0.167\cdot v} + \frac{0.005+0.86\cdot d+0.434\cdot v}{0.02+0.054\cdot d+0.7\cdot v} + \frac{-0.88\cdot d\cdot v}{0.04-0.4\cdot d-0.9\cdot v} \quad (6)$$

where $v$ is the velocity, $\varepsilon$ is the deviation and $d\varepsilon$ is the difference of the deviation at two consecutive time samples as inputs of the controller, $\delta$ is the output, a current between 0 and 5 V, with a performance index $I = 0.0039$.

Equation (6) presents an approximate model for the FLC. Further tuning of the AM parameters is necessary. Using GA optimization once more, the model for the fuzzy controller is:

$$\delta_{tuned} = \frac{0.113-0.28\cdot d+0.74\cdot v}{-0.4+0.83\cdot d+0.167\cdot v} + \frac{0.764+0.4\cdot d+0.464\cdot v}{0.02+0.054\cdot d+0.7\cdot v} + \frac{0.75\cdot d\cdot v}{0.04-1.03\cdot d+0.02\cdot v} \quad (7)$$

with a performance index of $J = 0.0087$ ($J$ is computed as the normalized sum of square acceleration deviations in closed loop).

The responses of the system using both the AM and the tuned geno-fuzzy controllers, versus the uncontrolled structural responsem show an acceleration reduction of the ground story of 39.25% for the tuned controller relative to the uncontrolled system and 37.5% for the AM controller relative to the uncontrolled system.

Further validation of the hybrid geno-fuzzy controller has been obtained by using a different set of external disturbances: the Vrancea 1977 earthquake accelerogram is presented in figure 9.

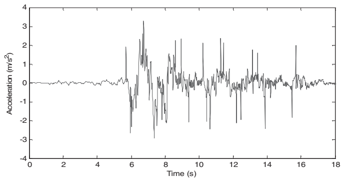


Figure 9. Vrancea earthquake accelerogram.

Results for each case are presented in Table I, for the following cases: using the fuzzy controller, using the algebraic model of the fuzzy controller and finally using the tuned geno-fuzzy controller. A comparative analysis of the proposed control strategy has been performed, using bang-bang and skyhook controllers [16]. Although these two controllers offer a good displacement reduction, their effect on the acceleration of the building is massive, adding up to 63% acceleration. In opposition, the geno-fuzzy controller reduces both acceleration and displacements, as well as inter-story drifts.

TABLE I. PERFORMANCE EVALUATION

| Disturbance | Controller | $J_1$ | $J_2$ | $J_3$ |
|---|---|---|---|---|
| Northridge | Fuzzy | 1.2419 | 1.2385 | 1.3227 |
| | AM | 0.9032 | 0.8979 | 1.0433 |
| | Geno-Fuzzy | 0.4729 | 0.4373 | 0.6317 |
| | Bang-bang | 0.1770 | 0.1570 | 1.5442 |
| | Skyhook | 0.7772 | 0.8101 | 1.6285 |
| Vrancea | Fuzzy | 1.3022 | 1.3541 | 1.2934 |
| | AM | 0.7802 | 0.7995 | 0.7563 |
| | Geno-Fuzzy | 0.5989 | 0.5959 | 0.7470 |
| | Bang-bang | 0.1760 | 0.1455 | 1.3737 |
| | Skyhook | 0.9076 | 0.9540 | 1.4363 |

For the Northridge earthquake, the tuned hybrid geno-fuzzy controller offers the best performances: 39.25% acceleration reduction, 75.32% velocity reduction and 70.1% displacement reduction. For the Vrancea earthquake, the hybrid geno-fuzzy controller offers the best performances: 12.6% acceleration reduction, 33.3% velocity reduction and 7.8% displacement reduction.

Figures 10 and 11 show the displacements, the inter-story drifts and the accelerations on each story, for the Northridge and Vrancea earthquakes, respectively. The designed geno-fuzzy controller offers a good performace for an earthquake with a maximum displacement amplitude approximately 4.5 times greater than the one it was originally designed for: maximum displacement for the Northridge earthquake is 27.9mm, while the maximum displacement for the Vrancea earthquake is 124.6mm. It was thus shown that the designed geno-fuzzy controller can still offer acceptable performances in conditions of high uncertainties regarding the seismic magnitude of the region for which the controller is meant to be implemented.

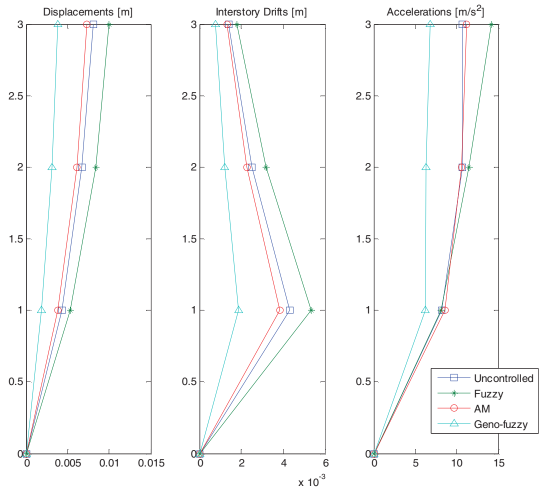


Figure 10. Maximum displacements, inter-story drifts and accelerations for the Northridge earthquake.

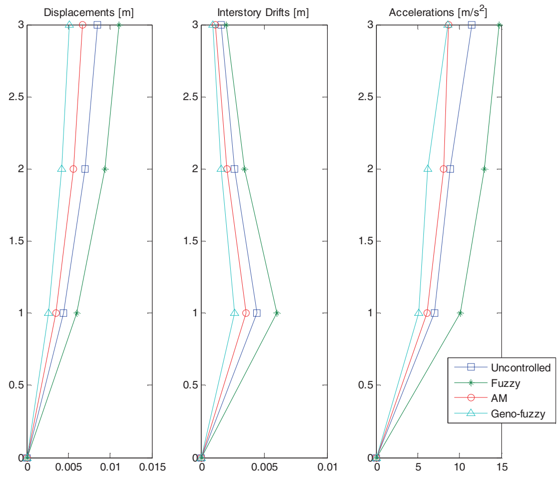


Figure 11. Maximum displacements, inter-story drifts and accelerations for the Vrancea earthquake.

The total computing time, including generating the algebraic model and the subsequent tuning procedure, is between 50 and 60 seconds, using an Intel Pentium Dual CPU 2.16 GHz processor. Through optimization of the GA code on a dedicated processor, the total computing time can be significantly reduced, thus making possible the real-time implementation of the geno-fuzzy tuning procedure.

## V. CONCLUSIONS

By means of genetic algorithms, the learning capabilities of a fuzzy logic controller have been enhanced. The designed hybrid geno-fuzzy controller offered better performance than the equivalent fuzzy controller.

Moreover, by changing the basic implementation form of the fuzzy controller from a heterogenic entity to an algebraic model, the computing time decreases exponentially, making the obtained geno-fuzzy controller easy to implement without removing its non-linear essence.

The proposed geno-fuzzy hybrid controller can offer improved performance and increased implementability in real life buildings.

## REFERENCES

[1] Hiemenz,G.J., Wereley, N.M. (1999). Seismic Response of Civil Structures Utilizing Semi-Active MR and ER Bracing Systems. *Journal of Intelligent Material Systems and Structures*, p. 646-651

[2] Dumitrache I., Catana I., Panduru V., Patrascu M. (2009) Fuzzy control strategies for magnetorheological dampers. *Proceedings of the 17th Intl. Conference on Control Systems and Computer Science*, Bucharest, Romania, p. 215-220

[3] Kim, H.-S., Roschke, P.N. (2005). Design of fuzzy logic controller for smart base isolation system using genetic algorithm. *Engineering Structures*, vol. 28, p. 84-96.

[4] Kim, H.-S., Roschke, P.N. (2006). GA-fuzzy control of smart base isolated benchmark building using supervisory control technique. *Advances in Engineering Software*, vol. 38, p. 453-465.

[5] Pourzeynali, S., Lavasani, H.H., Modarayi, A.H. (2006). Active control of high rise building structures using fuzzy logic and genetic algorithms. *Engineering Structures*, vol. 29, p. 346-357.

[6] Shook, D.A., Roschke, P.N., Lin, P.-Y., Loh, C.-H., (2007). GA-optimized fuzzy logic control of a large-scale building for seismic loads. *Engineering Structures,* vol. 30, p. 436-449.

[7] Ali, Sk.F., Ramaswamy, A. (2008). Optimal fuzzy logic control for MDOF structural systems using evolutionary algorithms. *Engineering Applications of Artificial Intelligence*, vol. 22, p. 407-419.

[8] Bitaraf, M., Ozbulut, O.E., Hurlebaus, S., Barroso, L. (2010). Application of semi-active control strategies for seismic protection of buildings with MR dampers. *Engineering Structures,* vol. 32, p. 3040-3047.

[9] Sims, N.D., Stanway, R., Johnson, A. (1999). Vibration control using smart fluids: a state-of-the-art review. *The Schock and Vibration Digest*, p. 195-203.

[10] Spencer Jr. B.F., Carlson, J.D., Sain, M.K., and Yang, G. (1997). On the Current Status of Magnetorheological Dampers: Seismic Protection of Full-Scale Structures, *Proc. of the Amer.Control Conf.*, pp. 458–62.

[11] Yan, G., Zhou L.L. (2006). Integrated fuzzy logic and genetic algorithms for multi-objective control of structures using MR dampers. *Journal of Sound and Vibration,* vol. 296 p. 368–382

[12] Chong, K.P., Liu, S.C., Li, J.C. (1990). Intelligent Structures, *Elsevier Publishers*, p. 249-250.

[13] Dumitrache, I., Buiu, C. (1999). Genetic learning of fuzzy controllers. *Mathematics and Computers in Simulation,* vol. 49, p. 13-26.

[14] Dumitrache, I., Buiu, C. (1999). *Algoritmi genetici*. Editura Nemira, Cluj-Napoca, Romania.

[15] Dyke S.J., Spencer B.F. Jr., Sain M.K., Carlson J.D., Modeling and control of magnetorheological dampers for seismic response reduction, *Smart Materials and Structures 5* (5) (1996) 565–575.

[16] Patrascu M., Dumitrache I., Patrut P. (2012). A comparative study for advanced seismic vibration control algorithms. *Scientific Bulletin of University Politehnica of Bucharest, Series C. Article in press.*

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Optimal Control Scheme for Nonlinear Systems with Saturating Actuator Using ε-Iterative Adaptive Dynamic Programming

Xiaofeng Lin, Yuanjun Huang and Nuyun Cao
School of Electrical Engineering, Guangxi University
Nanning, China
gxulinxf@163.com

Yuzhang Lin
Department of Electrical Engineering, Tsinghua University
Beijing, China
90lyz@163.com

*Abstract*— **In this paper, a finite-horizon optimal control scheme for a class of nonlinear systems with saturating actuator is proposed by an improved iterative adaptive dynamic programming (ADP) algorithm. The Hamilton-Jacobi-Bellman (HJB) equation corresponding to constrained control is formulated using a suitable nonquadratic function. Then mathematical analysis of the convergence is presented, by proving that the performance index function can reach the optimum using the adaptive iteration. Finally the finite-horizon optimal control law can be obtained by the ε-iterative adaptive algorithm. The examples are given to demonstrate the effectiveness of the above methods.**

***Keywords-Adaptive dynamic programming(ADP); Saturating actuators; nonlinear system; Finite time optimal control***

## I. INTRODUCTION

In a practical control system, the saturating actuators will reduce the system's dynamic performance, and even affect the stability of the system. Therefore seeking a better way to design control systems with saturating actuator has attached considerable attention by many researchers in recent years. The stability of discrete-time linear systems subject to actuator saturation was analyzed using a saturation-dependent Lyapunov function based on the solution of an LMI optimization problem[1]. Saberi (1996) and Sussmann(1994) proposed several processes to control saturation problems, but they did not consider non-linear systems and optimal problems (see [2],[3]). A gain-scheduled output control design for systems coping with nonlinear time-varying parameter dependent systems subject to saturated actuators was proposed in [4]. Pontryagin's Minimum Principle is a way to solve optimal control problem with Saturating Actuator. However, this needs to solve differential equations with boundary, and the result we get is an open-loop control by this way. Lyshevski designed an optimal control for a closed-loop feedback system using a non-quadratic performance index function problem to deal with the control constraints based on dynamic programming principle in [5], but the difficulty lies in the HJB equations.

Adaptive Dynamic Programming (ADP) is a powerful tool proposed by the idea of adaptive critic and reinforcement learning with dynamic programming[6]. It is solved by iterative algorithm to get an approximate solution of HJB equation, ADP has become an effective tool for optimal control problems and has achieved many results[7],[8-12]. A greedy iterative adaptive algorithm was proposed to solve the nonlinear discrete-time systems HJB equation in [8]. Iterative ADP was used to get an infinite horizon optimal control scheme for nonlinear systems with saturating actuator in [9].

However, for practical systems, a limited period of time is required to achieve control. Finite time optimal control problem could be backward solved by dynamic programming, when facing the multi-dimensional nonlinear characteristics of complex systems, the calculation will be very large, that is the problem of the "curse of dimensionality". ε-error bounds of adaptive algorithm was proposed to deal with finite time optimal control in [10],[11]. To the best of our knowledge, quite few research has been presented to deal with finite-horizon optimal control with saturating actuator. This motivates our research.

This paper aims to solve finite time optimal control problem for nonlinear systems with saturating actuator. It is organized as follows. In Section II, the problem is introduced and HJB equation corresponding to constrained control is presented. In Section III, the iterative ADP algorithm and its convergence for finite-horizon optimal control problem are derived. In Section IV, ε-optimal control algorithm is developed with the definition of finite iteration steps and ε-optimal control. In Section V, an example is given to demonstrate the effectiveness of the algorithm. In Section VI, the conclusion is drawn.

## II. HJB EQUATION CORRESPONDING TO CONSTRAINED CONTROL

### A. *Deal with saturated problem*

Consider the following class of discrete-time nonlinear systems

$$\boldsymbol{x}(k+1)=\boldsymbol{F}\left(\boldsymbol{x}(k),\boldsymbol{u}(k)\right) \qquad k=0,1,2,\ldots \tag{1}$$

Where $\boldsymbol{x}(k)\in\mathbb{R}^n$ is the state, assume the system $\boldsymbol{F}\left(\boldsymbol{x}(k),\boldsymbol{u}(k)\right)$ is Lipschitz continuous controllable on a set $\boldsymbol{\Omega}$

This work is supported by National Natural Science Foundation of China (60964002, 61034002)

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

containing the origin. The control $\boldsymbol{u}(k)\in\boldsymbol{\Omega}_{\boldsymbol{u}}$, and $\boldsymbol{\Omega}_{\boldsymbol{u}}=\left\{\boldsymbol{u}(k)=[u_1(k),u_2(k),\cdots,u_m(k)]^{\mathrm{T}}\in\mathbf{R}^{\mathrm{m}}:\left|u_i(k)\right|\le\bar{u}_i\right\}$, where $\overline{u}_i$ is the $i$-th execution controller saturation boundary, $i=1,\dots,m$. On the other hand, the constant diagonal matrix is $\mathbf{A}=diag\left[\bar{u}_1,\bar{u}_2,\cdots,\bar{u}_m\right],\mathbf{A}\in\mathbb{R}^{\mathrm{m\times m}}$ . $\boldsymbol{F}(\mathbf{0,0})=\mathbf{0}$, hence $\boldsymbol{x}=\mathbf{0}$ is an equilibrium state of system (1) under the control $\boldsymbol{u}=\mathbf{0}$.

For the initial state $\boldsymbol{x}(0)$, to minimize the performance index function defined in (2) with finite control sequences as $\underline{\boldsymbol{u}}_0^{N-1}=(\boldsymbol{u}(0),\boldsymbol{u}(1),\cdots,\boldsymbol{u}(N-1))$ .

$$\boldsymbol{J}\left(\boldsymbol{x}(0),\underline{\boldsymbol{u}}_0^{N-1}\right)=\sum_{k=0}^{N-1}\left\{\boldsymbol{x}(k)^{\mathrm{T}}Q\boldsymbol{x}(k)+\boldsymbol{W}(\boldsymbol{u}(k))\right\} \tag{2}$$

Where $\boldsymbol{W}(\boldsymbol{u}(k))=2\int_0^{\boldsymbol{u}(k)}(\mathbf{A}\tanh^{-1}(\boldsymbol{s}/\mathbf{A}))^{\mathrm{T}}\mathbf{R}d\boldsymbol{s}$. The length $N$ is determined with terminal time, this kind of optimal control problems has been called finite-horizon problems with unspecified terminal time [10].

If we set $\boldsymbol{W}(\boldsymbol{u}(k))=\boldsymbol{u}(k)^{\mathrm{T}}\boldsymbol{R}\boldsymbol{u}(k)$ in (2), the quadratic performance index function with unconstrained control is used to design the optimal control for control system with saturated actuator, however, this system could not ensure the optimal system performance, even may lead to system instability. According to [12], let

$$\boldsymbol{W}(\boldsymbol{u}(k))=2\int_0^{\boldsymbol{u}(k)}\mathbf{A}\boldsymbol{\phi}^{-T}({}^{\boldsymbol{s}}\!/\!_{\mathbf{A}})\mathbf{R}ds \tag{3}$$

Where

$$\boldsymbol{\phi}^{-1}(u(k))=[\phi^{-1}(u_1(k)),\phi^{-1}(u_2(k)),\cdots\phi^{-1}(u_m(k))]^T$$

$|\boldsymbol{u}|\le\mathbf{A}$, $\boldsymbol{s}\in\mathbb{R}^m$, $\boldsymbol{\phi}\in\mathbb{R}^m$, let $\boldsymbol{R}$ be a diagonal positive definite, $\phi(\bullet)$ is a bounded monotonically increasing odd function belongs to $\mathbb{C}^p(p\ge1)$ and $\boldsymbol{L}_2(\Omega)$ with $|\phi(\bullet)|\le1$ , the first derivative is a bounded constant $\boldsymbol{M}$ , such function as $\phi(\bullet)=\tanh(\bullet)$. Figure 1 shows a well approaching saturation when $|u|\le0.5$ with function $\boldsymbol{\phi}(\bullet)$. Therefore, we can guarantee the control output signal within the range of actuator saturation using the performance index function in (2).

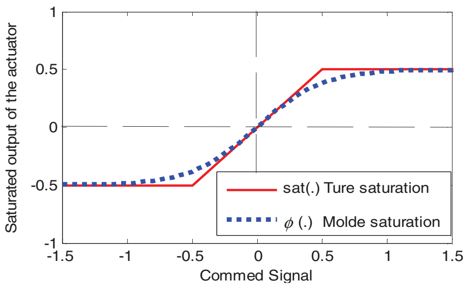


Figure 1. The model of saturation when $|u|\le0.5$

### B. HJB equation and solution

*Definition 1 :* The corresponding finite-horizon admissible control sequence of performance index function in (2) $\underline{\boldsymbol{u}}_k^{k+N-1}$ is defined as follows: For $\forall\boldsymbol{x}(k)\in\mathbb{R}^n$, there exists a control sequence $\underline{\boldsymbol{u}}_k^{k+N-1}$ satisfy $\boldsymbol{x}^f(\boldsymbol{x}(k),\underline{\boldsymbol{u}}_k^{k+N\text{-}1})=\boldsymbol{0}$ and $\boldsymbol{J}(\boldsymbol{x}(k),\underline{\boldsymbol{u}}_k^{N\text{-}1})$ is finite. Where $N>0$ is a positive integer, $\boldsymbol{x}^f(\boldsymbol{x}(k),\underline{\boldsymbol{u}}_k^{k+N\text{-}1})$ is the terminal state.

Let $\mathbf{C}_{\boldsymbol{x}_k}^N=\left\{\underline{\boldsymbol{u}}_k^{k+N-1}:\ \boldsymbol{x}^f(\boldsymbol{x}(k),\underline{\boldsymbol{u}}_k^{k+N\text{-}1})=\boldsymbol{0},\left|\underline{\boldsymbol{u}}_k^{k+N\text{-}1}\right|=N\right\}$ be all the admissible control sets with length $N$ .Assume a state $\boldsymbol{x}(k)$ is a finite-horizon admissible control. By Definition 1, the optimal performance index function at the finite-horizon admissible control could be written as

$$\boldsymbol{J}^*(\boldsymbol{x}(k))=\underset{\underline{\boldsymbol{u}}_k^{N\text{-}1}}{\boldsymbol{inf}}\left\{\boldsymbol{J}(\boldsymbol{x}(k),\underline{\boldsymbol{u}}_k^{N\text{-}1}):\underline{\boldsymbol{u}}_k^{N\text{-}1}\in\mathbf{C}_{\boldsymbol{x}_k}^N\right\} \tag{4}$$

According to the performance index function defined by equation (2) and Bellman principle of optimality, $\boldsymbol{J}^*(\boldsymbol{x}(k))$ under discrete time HJB equation can be written as

$$\boldsymbol{J}^*(\boldsymbol{x}(k))=\min_{\boldsymbol{u}(k)}\left\{\boldsymbol{x}(k)^{\mathrm{T}}Q\boldsymbol{x}(k)+\boldsymbol{W}(\boldsymbol{u}(k))+\boldsymbol{J}^*(\boldsymbol{F}(\boldsymbol{x}(k),\boldsymbol{u}(k)))\right\} \tag{5}$$

Define the optimal control sequence starting at $k$ with length of $N$ by

$$\underline{\boldsymbol{u}}_k^*(\boldsymbol{x}(k))=\underset{\underline{\boldsymbol{u}}_k^{k+N-1}}{\boldsymbol{inf}}\left\{\boldsymbol{J}(\boldsymbol{x}(k),\underline{\boldsymbol{u}}_k^{k+N-1}):\ \underline{\boldsymbol{u}}_k^{k+N-1}\in\mathbf{C}_{\boldsymbol{x}_k}^N\right\} \tag{6}$$

and define the one step optimal control vector by

$$\boldsymbol{u}^*(\boldsymbol{x}(k))=\mathbf{arg}\min_{\boldsymbol{u}(k)}\left\{\boldsymbol{x}(k)^{\mathrm{T}}Q\boldsymbol{x}(k)+\boldsymbol{W}(\boldsymbol{u}(k))+\boldsymbol{J}^*(\boldsymbol{F}(\boldsymbol{x}(k),\boldsymbol{u}(k)))\right\} \tag{7}$$

Dynamic programming is used to solving optimal control sequence in equation (6), while the first step is to determine $\boldsymbol{u}^*(\boldsymbol{x}(k+N-1))$ by equation(6) in terminal state $\boldsymbol{x}(k+N-1)$.

$$\boldsymbol{u}^*(\boldsymbol{x}(k+N-1))=\underset{\boldsymbol{u}(N\text{-}1)}{\mathbf{argmin}}\left\{\boldsymbol{x}(k+N-1)^{\mathrm{T}}Q\boldsymbol{x}(k+N-1)+\boldsymbol{W}(\boldsymbol{u}(k+N-1))\right\}$$
$$\boldsymbol{s.t.}\ \boldsymbol{F}(\boldsymbol{x}(k+N-1),\boldsymbol{u}(k+N-1))=\mathbf{0} \tag{8}$$

Putting $\boldsymbol{u}^*(\boldsymbol{x}(k+N-1))$ into (5), the optimal performance index function

$$\boldsymbol{J}^*(\boldsymbol{x}(k+N-1))=\boldsymbol{x}(k+N-1)^{\mathrm{T}}Q\boldsymbol{x}(k+N-1)+\boldsymbol{W}(\boldsymbol{u}^*(k+N-1)) \tag{9}$$

After $\boldsymbol{u}^*(\boldsymbol{x}(k+N-1))$ and $\boldsymbol{J}^*(\boldsymbol{x}(k+N-1))$ is obtained, $\boldsymbol{u}^*(\boldsymbol{x}(k+N-2))$ and $\boldsymbol{J}^*(\boldsymbol{x}(k+N-2))$,…, could be determined by equation (5) and (7), finally $\boldsymbol{u}^*(\boldsymbol{x}(k))$ and $\boldsymbol{J}^*(\boldsymbol{x}(k))$ is solved. Then solving $\boldsymbol{u}^*(k)=\boldsymbol{u}^*(\boldsymbol{x}(k))$ with $\boldsymbol{x}(k)$, and putting $\boldsymbol{u}^*(k)$ into equation (1) so as to get $\boldsymbol{x}(k+1)=\boldsymbol{F}(\boldsymbol{x}(k),\boldsymbol{u}^*(k))$. By the same way, solving $\boldsymbol{u}^*(k+1)$ by $\boldsymbol{x}(k+1)$ and $\boldsymbol{u}^*(\boldsymbol{x}(k+1))$ , putting $\boldsymbol{u}^*(\boldsymbol{x}(k+1))$ into equation (1) , then we can easily get $\boldsymbol{x}(k+2)=\boldsymbol{F}(\boldsymbol{x}(k+1),\boldsymbol{u}^*(k+1))$, repeating this process [9], the optimal control sequence will be obtained as $\underline{\boldsymbol{u}}_k^*(\bullet)=\{\boldsymbol{u}^*(k),\boldsymbol{u}^*(k+1),\dots,\boldsymbol{u}^*(k+N-1)\}$.

However, the performance index function in this paper is non- quadratic, and the system is non-linear. Therefore, it is hard to get an analytical solution of optimal control law by solving HJB equation. On the other hand, when dealing with discrete time dynamic programming by backward method, one has to calculate and save all $\boldsymbol{J}^*(\boldsymbol{x}(k))$ and $\boldsymbol{u}^*(\boldsymbol{x}(k))$ of the sequence. Hence, it will meet difficulties when calculating finite-horizon optimal control problems by dynamic programming.

## III. FINITE TIME ITERATIVE ADP ALGORITHM

### *A. Formula derived for iterative ADP*

For any state $\boldsymbol{x}(k)$, the performance index function $\{\boldsymbol{V}_i\}$ and control policy $\{\upsilon_i\}$ in the iterative ADP algorithm are updated by recursive iterations. The iterative starts with $i=0$ and the initial performance index function $\boldsymbol{V}_0(\boldsymbol{x}(k))=0$, the performance index function for $i=1$ is computed as

$$\begin{aligned}\boldsymbol{V}_1(\boldsymbol{x}(k)) &= \min_{\boldsymbol{u}(k)}\left\{\boldsymbol{x}(k)^{\mathrm{T}}Q\boldsymbol{x}(k)+\boldsymbol{W}(\boldsymbol{u}(k))+\boldsymbol{V}_0(\boldsymbol{F}(\boldsymbol{x}(k),\boldsymbol{u}(k)))\right\}\\ &\quad s.t.\ \boldsymbol{F}(\boldsymbol{x}(k),\boldsymbol{u}(k))=\boldsymbol{0}\\ &=\min_{\boldsymbol{u}(k)}\left\{\boldsymbol{x}(k)^{\mathrm{T}}Q\boldsymbol{x}(k)+\boldsymbol{W}(\boldsymbol{u}(k))\right\}\\ &\quad s.t.\ \boldsymbol{F}(\boldsymbol{x}(k),\boldsymbol{u}(k))=\boldsymbol{0}\\ &=\boldsymbol{x}(k)^{\mathrm{T}}Q\boldsymbol{x}(k)+\boldsymbol{W}(\boldsymbol{u}^*(\boldsymbol{x}(k)))\end{aligned} \tag{10}$$

Where $\boldsymbol{F}\left(\boldsymbol{x}(k),\boldsymbol{u}^*(k)\right)=\boldsymbol{0}$, let $\boldsymbol{u}^*(k)=\upsilon_1(\boldsymbol{x}(k))$, then (10) can be written as

$$\begin{aligned}\boldsymbol{V}_1(\boldsymbol{x}(k)) &= \min_{\boldsymbol{u}(k)}\left\{\boldsymbol{x}(k)^{\mathrm{T}}Q\boldsymbol{x}(k)+\boldsymbol{W}(\boldsymbol{u}(k))+\boldsymbol{V}_0(\boldsymbol{F}(\boldsymbol{x}(k),\boldsymbol{u}(k)))\right\}\\ &\quad s.t.\ \boldsymbol{F}(\boldsymbol{x}(k),\boldsymbol{u}(k))=\boldsymbol{0}\\ &=\boldsymbol{x}(k)^{\mathrm{T}}Q\boldsymbol{x}(k)+\boldsymbol{W}(\upsilon_1(\boldsymbol{x}(k)))\end{aligned} \tag{11}$$

where

$$\begin{aligned}\upsilon_1(\boldsymbol{x}(k)) &= \mathbf{arg}\min_{\boldsymbol{u}(k)}\{\boldsymbol{x}(k)^{\mathrm{T}}Q\boldsymbol{x}(k)+\boldsymbol{W}(\boldsymbol{u}(k))\}\\ &\quad s.t.\ \boldsymbol{F}(\boldsymbol{x}(k),\boldsymbol{u}(k))=\boldsymbol{0}\end{aligned} \tag{12}$$

For $i=2,3,4\ldots$, the iterative ADP algorithm is updated as follows

$$\begin{aligned}\boldsymbol{V}_i(\boldsymbol{x}(k)) &= \min_{\boldsymbol{u}(k)}\left\{\boldsymbol{x}(k)^{\mathrm{T}}Q\boldsymbol{x}(k)+\boldsymbol{W}(\boldsymbol{u}(k))+\boldsymbol{V}_{i-1}(\boldsymbol{F}(\boldsymbol{x}(k),\boldsymbol{u}(k)))\right\}\\ &=\boldsymbol{x}(k)^{\mathrm{T}}Q\boldsymbol{x}(k)+\boldsymbol{W}(\upsilon_i(\boldsymbol{x}(k)))+\boldsymbol{V}_{i-1}(\boldsymbol{F}(\boldsymbol{x}(k),\upsilon_i(\boldsymbol{x}(k))))\end{aligned} \tag{13}$$

where

$$\upsilon_i(\boldsymbol{x}(k))=\mathbf{arg}\min_{\boldsymbol{u}(k)}\{\boldsymbol{x}(k)^{\mathrm{T}}Q\boldsymbol{x}(k)+\boldsymbol{W}(\boldsymbol{u}(k))+\boldsymbol{V}_{i-1}(\boldsymbol{F}(\boldsymbol{x}(k),\boldsymbol{u}(k)))\} \tag{14}$$

After $i$ steps iteration, the performance index function sequence is obtained as $\{\boldsymbol{V}_i\}=(\boldsymbol{V}_i,\boldsymbol{V}_{i-1},\cdots,\boldsymbol{V}_1)$ and control policy sequence as $\{\upsilon_i\}=(\upsilon_i,\upsilon_{i-1},\cdots\upsilon_1)$. Note that in this case, each control sequence $\boldsymbol{u}(\cdot)$ will obey with different control policy, that is, for $i=0,1,\cdots N-1$, control sequence $\boldsymbol{u}(i)$ is obtained by $\upsilon_i$ respectively. However, one could prove that $\boldsymbol{V}_i(\boldsymbol{x}(k))$ is limit to $\boldsymbol{J}^*(\boldsymbol{x}(k))$ when $i\to\infty$. Therefore, the performance index function $\boldsymbol{J}^*(\boldsymbol{x}(k))$ in HJB equation will be replaced by the iterative performance index function $\boldsymbol{V}_i(\boldsymbol{x}(k))$ while control policy $\boldsymbol{u}(\boldsymbol{x})$ will be replaced by the iterative control policy.

### *B. Convergence of iteration ADP*

*Remark 1:* According to (2) and (13), we have

$$\boldsymbol{V}_{i+1}(\boldsymbol{x}(k))=\min_{\underline{\boldsymbol{u}}_k^{k+i}}\{\boldsymbol{J}(\boldsymbol{x}(k),\underline{\boldsymbol{u}}_k^{k+i}):\underline{\boldsymbol{u}}_k^{k+i}\in\mathbf{C}_{\boldsymbol{x}_k}^{(i+1)}\} \tag{15}$$

*Prove:* According to (13), we have

$$\begin{aligned}\boldsymbol{V}_i(\boldsymbol{x}(k+i)) &= \min_{\boldsymbol{u}(k+i)}\{\boldsymbol{x}(k+i)^{\mathrm{T}}Q\boldsymbol{x}(k+i)\ +\boldsymbol{W}(\boldsymbol{u}(k+i)))\}\\ &\quad \boldsymbol{s.t.}\ \boldsymbol{F}(\boldsymbol{x}(k+i),\boldsymbol{u}(k+i))=\boldsymbol{0}\end{aligned} \tag{16}$$

Thus,

$$\begin{aligned}\boldsymbol{V}_{i+1}(\boldsymbol{x}(k)) &= \min_{\boldsymbol{u}(k)}\{\boldsymbol{x}(k)^{\mathrm{T}}Q\boldsymbol{x}(k)+\boldsymbol{W}(\boldsymbol{u}(k))+\boldsymbol{V}_i(\boldsymbol{x}(\mathrm{k}+1)\}\\ &=\min_{\boldsymbol{u}(k)}\left\{\ \boldsymbol{x}(k)^{\mathrm{T}}\mathrm{Q}\boldsymbol{x}(k)+\boldsymbol{W}(\boldsymbol{u}(k))\right.\\ &\quad+\min_{\boldsymbol{u}(k+1)}\left\{\ \{\boldsymbol{x}(k+1)^{\mathrm{T}}Q\boldsymbol{x}(k+1)+\boldsymbol{W}(\boldsymbol{u}(k+1))\right.\\ &\quad+\min_{\boldsymbol{u}(k+2)}\{\ \boldsymbol{x}(k+2)^{\mathrm{T}}Q\boldsymbol{x}(k+2)+\boldsymbol{W}(\boldsymbol{u}(k+2))+\cdots\\ &\quad+\min_{\boldsymbol{u}(k+i-1)}\{\boldsymbol{x}(k+i\text{-}1)^{\mathrm{T}}Q\boldsymbol{x}(k+i\text{-}1)+\boldsymbol{W}(\boldsymbol{u}(k+i\text{-}1))\\ &\quad\left.\left.+\boldsymbol{V}_1(\boldsymbol{x}(k+i))\}\cdots\ \}\ \right\}\ \right\}\end{aligned} \tag{17}$$

Then (17) could be written as

$$\begin{aligned}\boldsymbol{V}_{i+1}(\boldsymbol{x}(k)) &= \min_{\underline{\boldsymbol{u}}_k^{k+i}}\{\boldsymbol{x}(k)^{\mathrm{T}}Q\boldsymbol{x}(k)+\boldsymbol{W}(\boldsymbol{u}(k))\\ &\quad+\boldsymbol{x}(k+1)^{\mathrm{T}}Q\boldsymbol{x}(k+1)\ +\boldsymbol{W}(\boldsymbol{u}(k+1))+\cdots\\ &\quad+\boldsymbol{x}(k+i)^{\mathrm{T}}Q\boldsymbol{x}(k+i)+\boldsymbol{W}(\boldsymbol{u}(k+i))\}\\ &\quad\boldsymbol{s.t.}\ \boldsymbol{F}(\boldsymbol{x}(k+i),\boldsymbol{u}(k+i))=\boldsymbol{0}\\ &=\min_{\underline{\boldsymbol{u}}_k^{k+i}}\{\boldsymbol{J}(\boldsymbol{x}(k),\underline{\boldsymbol{u}}_k^{k+i}):\underline{\boldsymbol{u}}_k^{k+i}\in\mathbf{C}_{\boldsymbol{x}_k}^{(i+1)}\}\end{aligned} \tag{18}$$

Thus (15) holds.

Three theorems as well as a corollary are given in the following. For the proof in detail, see [10],[11].

*Theorem 1:* Giving an arbitrary state vector $\boldsymbol{x}(k)$, assume there exist an integral $i$ such that $\mathbf{C}_{\boldsymbol{x}_k}^{(i)}\neq\varnothing$, then $\mathbf{C}_{\boldsymbol{x}_k}^{(i+1)}\neq\varnothing$, the performance index function $\boldsymbol{V}_i(\boldsymbol{x}(k))$ is a monotonically nonincreasing sequence. i.e. $\forall i\geq 1$, $\boldsymbol{V}_{i+1}(\boldsymbol{x}(k))\leq\boldsymbol{V}_i(\boldsymbol{x}(k))$.

*Theorem 2:* Giving an arbitrary state vector $\boldsymbol{x}(k)$,Define the performance index function $\boldsymbol{V}_\infty(\boldsymbol{x}(k))$ as the limit of the iterative function $V_i(\boldsymbol{x}(k))$, i.e.,

$$\boldsymbol{V}_\infty(\boldsymbol{x}(k)) = \lim_{i\to\infty} \boldsymbol{V}_i(\boldsymbol{x}(k)) \tag{19}$$

Thus,

$$V_\infty(\boldsymbol{x}(k)) = \min_{\boldsymbol{u}(k)} \{\boldsymbol{x}(k)^{\mathrm{T}} Q\boldsymbol{x}(k) + \boldsymbol{W}(\boldsymbol{u}(k)) + V_\infty(\boldsymbol{x}(k+1))\} \tag{20}$$

*Theorem 3:* Define the performance index function $\boldsymbol{V}_\infty(\boldsymbol{x}(k))$ as (13), assume the state $\boldsymbol{x}(k)$ is controllable, then the performance index function $\boldsymbol{V}_\infty(\boldsymbol{x}(k))$ equals the optimal performance index function $\boldsymbol{J}^*(\boldsymbol{x}(k))$, i.e.

$$\lim_{i\to\infty} \boldsymbol{V}_i(\boldsymbol{x}(k)) = \boldsymbol{J}^*(\boldsymbol{x}(k)) \tag{21}$$

*Corollary 1:* Assume the system state $\boldsymbol{x}(k)$ is admissible controllable and the performance index function is defined in (13), and Theorem 3 holds, then the iterative control law $\upsilon_i(\boldsymbol{x}(k))$ converges to the optimal control law $\boldsymbol{u}^*(\boldsymbol{x}(k))$.

## IV. ε-Optimal Control Algorithm

### *A. ε-Optimal Control*

In order to get the limit of the performance index function $\boldsymbol{V}_i(\boldsymbol{x}(k))$, iterative ADP algorithm (10) is need to perform, optimal control law $\boldsymbol{u}^*(\boldsymbol{x}(k))$ won't obtain until $i \to \infty$. However, $i$ is usually numbered in practical, $\boldsymbol{J}^*(\boldsymbol{x}(k)) = \boldsymbol{V}_i(\boldsymbol{x}(k))$ could not hold for any finite $i$. Therefore, error bound ε is introduced for $\boldsymbol{V}_i(\boldsymbol{x}(k))$ and $\boldsymbol{J}^*(\boldsymbol{x}(k))$ in iterative ADP algorithm, so that the performance index function $\boldsymbol{V}_i(\boldsymbol{x}(k))$ approximates to the optimal performance index function $\boldsymbol{J}^*(\boldsymbol{x}(k))$ in finite number of steps.

*Definition 2:* let $\boldsymbol{\Gamma}_\infty$ be a controllable state set, $\boldsymbol{x}(k) \in \boldsymbol{\Gamma}_\infty, \varepsilon > 0$, then the number of iteration steps $K_{\boldsymbol{\varepsilon}}(\boldsymbol{x}(k))$ for optimal control is defined as

$$K_{\boldsymbol{\varepsilon}}(\boldsymbol{x}(k)) = \min\left\{ i : \left| \boldsymbol{V}_i(\boldsymbol{x}(k)) - \boldsymbol{J}^*(\boldsymbol{x}(k)) \right| \le \boldsymbol{\varepsilon} \right\} \tag{22}$$

$K_{\boldsymbol{\varepsilon}}(\boldsymbol{x}(k))$ refers to the length of control sequence reaching equilibrium point from $\boldsymbol{x}(k)$, since $\boldsymbol{x}(k) \in \boldsymbol{\Gamma}_\infty$, thus $\lim_{i\to\infty} \boldsymbol{V}_i(\boldsymbol{x}(k)) = \boldsymbol{J}^*(\boldsymbol{x}(k))$. Therefore, there exists a finite $i$ such that

$$\left| \boldsymbol{V}_i(\boldsymbol{x}(k)) - \boldsymbol{J}^*(\boldsymbol{x}(k)) \right| \le \boldsymbol{\varepsilon} \tag{23}$$

holds. Thus $\left\{ i : \left| \boldsymbol{V}_i(\boldsymbol{x}(k)) - \boldsymbol{J}^*(\boldsymbol{x}(k)) \right| \le \boldsymbol{\varepsilon} \right\} \ne \varnothing$, so that $K_{\boldsymbol{\varepsilon}}(\boldsymbol{x}(k))$ is defined.

*Definition 3:* let $\boldsymbol{x}(k) \in \boldsymbol{\Gamma}_\infty$ be a controllable state vector, for any $\boldsymbol{\varepsilon} > 0$, if $\left| \boldsymbol{V}_i(\boldsymbol{x}(k)) - \boldsymbol{J}^*(\boldsymbol{x}(k)) \right| \le \boldsymbol{\varepsilon}$ holds for the iterative control law $\upsilon_i(\boldsymbol{x}(k))$, then the $\upsilon_i(\boldsymbol{x}(k))$ is defined as a ε-Optimal Control $\boldsymbol{\mu}_\varepsilon^*(\boldsymbol{x}(k))$, i.e.

$$\boldsymbol{\mu}_\varepsilon^*(\boldsymbol{x}(k)) = \upsilon_i(\boldsymbol{x}(k)) = \arg\min_{\boldsymbol{u}(k)} \left\{ \begin{array}{l} \boldsymbol{x}(k)^{\mathrm{T}} Q\boldsymbol{x}(k) + \boldsymbol{W}(\boldsymbol{u}(k)) \\ + \boldsymbol{V}_{i-1}(\boldsymbol{F}(\boldsymbol{x}(k), \boldsymbol{u}(\boldsymbol{x}(k)))) \end{array} \right\} \tag{24}$$

### *B. Summary of the ε -Optimal Control Algorithm*

Step A1. Giving an initial state $\boldsymbol{x}(k)$ and an error bound ε .

Step A2. Set $i = 0$, $V_0(\boldsymbol{x}(k)) = 0$ and $K_{\boldsymbol{\varepsilon}}(\boldsymbol{x}(k)) = 0$.

Step A3. Calculate $\upsilon_1(\boldsymbol{x}(k)) = \boldsymbol{u}^*(k)$ by (12).

Step A4. For $i = 1$, calculate $\boldsymbol{V}_1(\boldsymbol{x}(k))$ by (11).

Step A5. Set $i = i + 1$ and $K_{\boldsymbol{\varepsilon}}(\boldsymbol{x}(k)) = i$

Step A6. For $i = 2, 3, \cdots$, calculate $\upsilon_i(\boldsymbol{x}(k))$ by (14), calculate $\boldsymbol{V}_i(\boldsymbol{x}(k))$ by (13).

Step A7. If $\left| \boldsymbol{V}_{i+1}(\boldsymbol{x}(k)) - \boldsymbol{V}_i(\boldsymbol{x}(k)) \right| \le \boldsymbol{\varepsilon}$, go to step A8; then $K_{\boldsymbol{\varepsilon}}(\boldsymbol{x}(k)) = i$ is the number of optimal control steps, ε-optimal control law is $\boldsymbol{\mu}_\varepsilon^*(\boldsymbol{x}(k)) = \upsilon_i(\boldsymbol{x}(k))$ .otherwise, go to step A5.

Step A8. Stop.

## V. Simulation Study

### *A. ε-Iterative ADP Algorithm for Saturating Actuator*

Consider the following nonlinear system

$$\boldsymbol{x}(k+1) = f(\boldsymbol{x}(k)) + g(\boldsymbol{x}(k))\boldsymbol{u}(k) \tag{25}$$

where

$$f(\boldsymbol{x}(k)) = \begin{bmatrix} -0.5x_2(k) \\ sin(0.8x_1(k) - x_2(k)) + 1.8x_2(k) \end{bmatrix}$$

$$g(\boldsymbol{x}(k)) = \begin{bmatrix} -0.1x_1(k) & 0 \\ 0 & -0.8x_2(k) \end{bmatrix}, \mathbf{A} = \begin{bmatrix} \bar{u}_1 & 0 \\ 0 & \bar{u}_2 \end{bmatrix} = \begin{bmatrix} 0.5 & 0 \\ 0 & 1 \end{bmatrix}, \boldsymbol{Q} = \boldsymbol{R} = \boldsymbol{I}_{2\times 2},$$

The performance index function is

$$\boldsymbol{J}\left(\boldsymbol{x}(k), \boldsymbol{u}(\cdot)\right) = \sum_{i=k}^{k+N-1} \left\{ \boldsymbol{x}(i)^{\mathrm{T}} Q\boldsymbol{x}(i) + \boldsymbol{W}(\boldsymbol{u}(i)) \right\} \tag{26}$$

Neural networks are used to implement the iterative ADP algorithm in [7],[9],[10] with its good function approximating characteristics. The structure diagram of the iterative ADP algorithm using Neural-Network approximate function is shown in figure 2. In the diagram, critic neural network is used to approximate function $\boldsymbol{V}_i(\boldsymbol{x})$, action neural network is used to control law $\upsilon_i(\boldsymbol{x}(k))$, gradient descent algorithm is used to adjust the weight by neural network training rule, the approximate proof and formula derivation can be checked in [7],[9],[12]. The critic network and the action network are chosen as three-layer back-propagation (BP) neural networks with the structures of 2–10–1 and 2–10–2. According to ε-optimal control algorithm, let $\boldsymbol{\varepsilon} = 10^{-3}$, the initial state is chosen as $\boldsymbol{x}(k) = [1, -1]^{\mathrm{T}}$, learning rate $\alpha = 0.05$ . For each iterative step, the critic network and the action network are trained for 100 iteration steps so as to guarantee the neural network training error is less than $10^{-6}$.

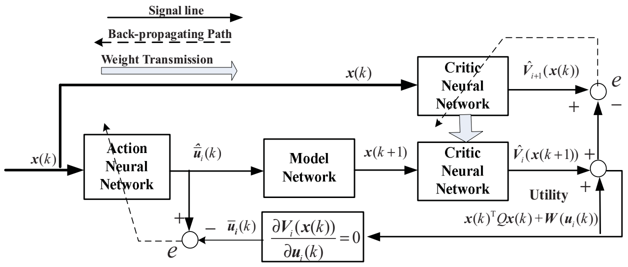


Figure 2. The structure diagram of the iterative ADP algorithm using Neural-Network approximate $V_i(x)$ and $\upsilon_i(x)$

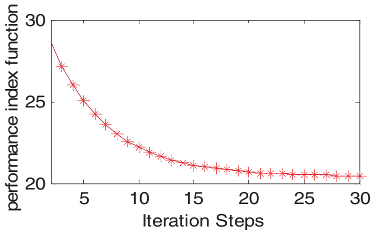


Figure 3. The convergence process of the cost function

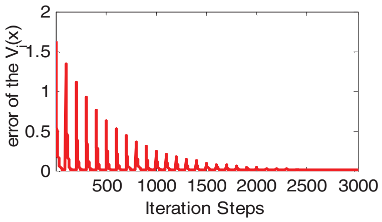


Figure 4. The error of neural-network approximate cost function

Simulation result in Figure 3 shows the iteration convergence process of the cost function $V_i(x(k))$. When $i=30$, $|V_i(x(k))-J^*(x(k))|\le\varepsilon$ holds, thus ε−optimal control law $\mu_\varepsilon^*$ is obtained in finite step $K_\varepsilon(x(k))=i=30$. Besides, Figure 3 also shows that $V_i(x(k))$ satisfy the monotonically nonincreasing and constringency property in Theorem 1 since $V_{i+1}(x(k))\le V_i(x(k))$. Figure 4 shows the error of neural-network approximate cost function. In order to verify the control law $\mu_\varepsilon^*$ obtained by ε-iteration algorithm, the state trajectories and the optimal control trajectories are shows in Figure 5 and Figure 6 for the state $x(k)=[1,-1]^{\mathrm{T}}$, Figure 7 shows the state trajectories in 3d space performance. Notice that the state has been control in stable within finite step $T_f=30$ showed in Figure 6, which satisfies the theory, on the other hand, the control output $|u_1|\le\bar{u}_1=0.5$, $|u_2|\le\bar{u}_2=1$ in Figure 6 shows that control output signal actuator keeps under the constraints $|u|\le\mathbf{A}$.

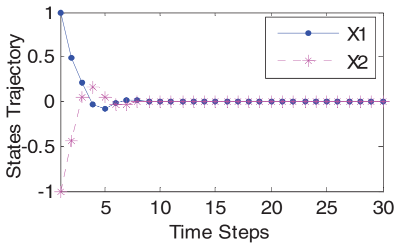


Figure 5. The state trajectories

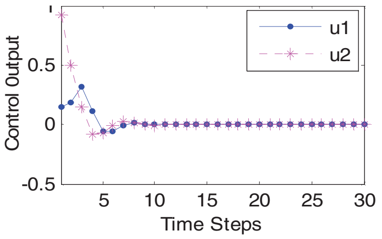


Figure 6. The optimal control trajectories

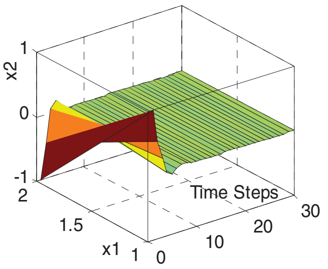


Figure 7. The state trajectories in 3d space performance

## *B. Comparison of simulation*

In order to contrast with the controller with saturating actuator, the state trajectories without actuators saturation is showed in Figure 8 and the optimal control trajectories without actuators saturation is showed in Figure 9. In Figure 9, control output satisfy $|u_1|\ge 0.5$, $|u_2|\ge 1$, thus the control signal will be distorted with saturating actuator, while it keeps under the constraints in Figure 6. Comparing results with Figure 6 and Figure 9, result shows that ε-Iterative adaptive dynamic programming works effective for finite-horizon optimal control scheme for nonlinear systems with saturating actuator.

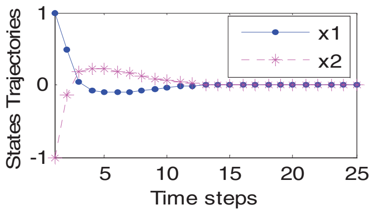


Figure 8. The state trajectories without actuators saturation

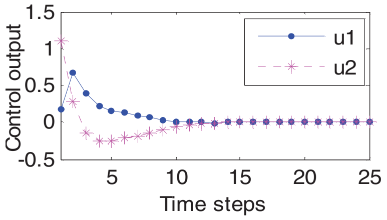


Figure 9. The optimal control trajectories without actuators saturation

## VI. CONCLUSION

In this paper, a functional performance index function was used to deal with saturating actuator control with constraints effectively. Furthermore, discrete HJB equation of nonlinear systems was derived. Finite horizon iteration ADP algorithm for control with saturation was developed via mathematical analysis. Finally the finite-horizon optimal control was obtained by the ε−iterative adaptive algorithm.

## REFERENCES

[1] Yong-Yan Cao, Zongli Lin, "Stability analysis of discrete-time systems with actuator saturation by a saturation-dependent Lyapunov function," Automatica,Vol.39, 2003, pp. 1235─1241.

[2] Saberi, A Z,Lin,A Teel, "Control of Linear Systems with Saturating Actuators," IEEE Transactions on Automatic Control, Vol 41, NO.3, March 1996, pp. 368─378.

[3] Sussmann H,E D Sontag,Y Yang, "A General Result on the Stabilization of Linear Systems Using Bounded Controls," IEEE Trans. Automatic Control, Vol.39, No.12, December 1994, pp. 2411─2425.

[4] Marc Jungers a, Eugênio B. Castelan, "Gain-scheduled output control design for a class of discrete-time nonlinear systems with saturating actuators," Systems & Control Letters, Vol.60, 2011, 169─173.

[5] Lyshevski, SE, "Optimal Control of Nonlinear Continuou Time Systems: Design of Bounded Controllers ViaGen-ralized Nonquadratic Functionals," American Control Conference, June 1998, pp.205─209.

[6] Werbos P J, "Approximate dynamic programming for reall controlll and neural modeling．Handbook of Intelligent Control," Neural Fuzzy and Adaptive Approaches．New York：Van No strand Reinhold, 1992

[7] Wei Qing-lai, Zhang Huang-guang, Liu De-rong, Zhao Yan, "An Optimal Control Scheme for a Class of Discrete-time Nonlinear Systems with Time Delays Using Adaptive Dynamic Programming," Acta Automatica Sinica, Vol 36, NO.1, 2010, pp. 121─129.

[8] Al-Tamimi, F L. Lewis, and M Abu-Khalaf, "Model Free Q-learning designs for linear discrete-time zero-sum games with application to H-infinity control," Automatica, Vol. 43, NO.3, Mar 2007, pp473─481.

[9] Luo Yan-Hong, "Research on Adaptive and Optimal Control for Nonlinear Systems Based on Neural Networks," Northerestern University, 2008, pp.33─59.

[10] Wang Fei-Yue, Jin Ning, Liu Derong, Wei Qing-Lai, "Adaptive Dynamic Programming for Finite-Horizon Optimal Control of Discrete-Time Nonlinear Systems With ε-Error Bound," Transactions on Neural Networks, Vol. 22, 2011, pp. 24─36.

[11] Wang Fei-Yue, Jin Ning, Liu Derong, Hou Zeng-Guang, "adaptive Dynamic Programming with epsilon-Error Bound for Nonlinear Discrete-Time Systems Using Neural Networks," Transactions on Neural Networks, 2008.

[12] Abu-Khalaf M, Lewis F L, " Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach," Automatica 2005, Vol 41, NO.5, pp. 779–791.

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# Active control of speed fluctuations in rotating machines using feedback linearization

Mirjam Holm, Pablo Ballesteros, Stephan Beitler, Alex Tarasow, Christian Bohn
Institute of Electrical Information Technology
Clausthal University of Technology
Clausthal-Zellerfeld, Germany
{ballesteros, beitler, bohn, holm, tarasow}@iei.tu-clausthal.de

*Abstract* - **This paper presents a method for actively controlling torsional vibrations in rotating machines caused by angle-dependent parameters. The work is motivated by rotating machines with crank or cam gear mechanisms that cause fluctuations in the angular speed when the machine is driven by a constant load torque or when the speed is controlled with conventional controllers. A very general model for such a system is introduced and used to derive a control law by feedback linearization. With this control law, the speed fluctuations are completely eliminated and desired linear dynamics can be prescribed for the system. The method is tested in a simulation study with a model of a real industrial machine. Although the proposed method works well, the study is preliminary in the sense that the method has not been applied experimentally and its robustness has not been assessed.**

***Active vibration control; Feedback linearization; Nonlinear control; Speed control; Torsional vibrations***

## I. Motivation

There are different causes for torsional vibrations in drives. These vibrations can be the result of eccentric masses [1] or be excited by the propulsion itself, e.g. a combustion engine [2, 3, 4] or an electric drive [5, 6]. Another commonly occurring cause is the transformation of the rotational motion of the propulsion into a translational motion. This typically happens in a cam gear or a crank assembly and causes a periodically varying load torque.

As an example, a simple crank assembly, which only includes a single oscillating mass, is shown in Fig. 1. The torque equilibrium for this example leads to the equation of motion given as

$$\begin{aligned} T &= m\cdot\ddot{s}(\varphi)\frac{\mathrm{d}s}{\mathrm{d}\varphi} \\ &= m\left(\frac{\mathrm{d}^2s}{\mathrm{d}\varphi^2}\dot{\varphi}^2\cdot+\frac{\mathrm{d}s}{\mathrm{d}\varphi}\cdot\ddot{\varphi}\right)\frac{\mathrm{d}s}{\mathrm{d}\varphi} \qquad (1)\\ &= m\frac{\mathrm{d}^2s}{\mathrm{d}\varphi^2}\frac{\mathrm{d}s}{\mathrm{d}\varphi}\dot{\varphi}^2+m\left(\frac{\mathrm{d}s}{\mathrm{d}\varphi}\right)^2\ddot{\varphi}\ , \end{aligned}$$

where $T$ is the driving torque and all the other variables are as shown in Fig. 1. The displacement $s$ depends on the rotation angles $\alpha$ and $\varphi$ and is given as

$$\begin{aligned} s &= l\left(1-\cos\alpha\right)+r\left(1-\cos\varphi\right) \\ &= l\left(1-\sqrt{1-\frac{r^2}{l^2}\sin^2\varphi}\right)+r\left(1-\cos\varphi\right)=s\left(\varphi\right). \qquad (2) \end{aligned}$$

The introduction of $m\cdot\left(\mathrm{d}s/\mathrm{d}\varphi\right)^2$ as an angle depended parameter $J(\varphi)$ (which could be seen as an apparent, angle-dependent, moment of inertia) leads to

$$T=\frac{1}{2}\frac{\mathrm{d}J(\varphi)}{\mathrm{d}\varphi}\dot{\varphi}^2+J(\varphi)\ddot{\varphi}\,. \qquad (3)$$

Even if a constant driving torque is applied to the system, the internal dynamics lead to nonlinear vibrations. This is due to the dependency of the parameters in the model describing the system on the rotational angle.

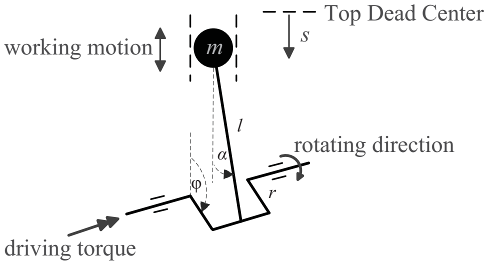


Fig. 1 Schematic diagram of the crankshaft

Such torsional vibrations are a well-known problem in industrial drives like rolling mill systems [7, 8], paper machines [9, 10] or turbo generators [11] and also in vehicles [2]. They lead to product quality problems like damage of machinery and a shorter lifecycle of the plant.

Possible approaches for the attenuation of these vibrations are to change the inertia or to use damping elements, but this often has a negative effect on the dynamics of the machines. Because of these disadvantages several active control approaches to attenuate torsional vibrations have been proposed. In [12], a damping filter is implemented and in [13] a digital PI/PID controller is used to achieve active damping of the vibrations. In [14], an adaptive sliding neuro-fuzzy approach is suggested and in [15] sliding mode and force dynamics are considered. Active damping based on learning control is presented in [16]. In [4] the main target is to reduce the torsional vibrations of a crankshaft by adjusting the fuel injection duration by minimizing a cost function.

Another way to control the dynamic system and attenuate the nonlinear behavior and prescribe linear system dynamics is to use the method of feedback linearization. In [17] and [18], feedback linearization control is used for improved fuel consumption in combustion engines by using feedback linearization, but the afore-mentioned dynamical connections are only regarded marginally. In other fields, vibration control is the main aspect as in [19] where feedback linearization is used for active vehicle suspension control to reduce the vibrations. To the best of the authors' knowledge, however, no applications of feedback linearization to the problem considered in this paper have been reported in the literature to date.

As mentioned above, the torsional vibrations can result from angle-dependent parameters in the system dynamics. The angle itself is also a state of the system. In order to make the method presented in this paper applicable to a wide range of systems, a very universal system description is introduced for systems with state-dependent parameters. The resulting description is very compact. With this description, it is very easy to derive a control law (which is given here only for systems with a relative degree of two, but can easily be extended).

The remainder of this paper is organized as follows. The system description is introduced in Sec. II. The control law is derived in Sec. III. The control approach has been tested in simulation studies for a model that corresponds to an industrial machine that is operated by an industrial partner with whom this study has been conducted. The obtained results are shown in Sec. IV. Some conclusions are given in Sec. V.

## II. System Description

Real industrial drives are more complex systems than the simple example discussed in the introduction. Often, they are composed of more than one crank assembly or cam gear or even combine both types. Furthermore, elasticity, damping and gear ratios have to be taken into account.

Nevertheless, the final system description of such systems will still contain constant and angle-dependent parameters. Due to the rotational mode of operation, the angle-dependent parameters are periodic in the rotational angle (see for example, the parameter $J_\alpha$ for the example considered in Sec. IV shown in Fig. 2).

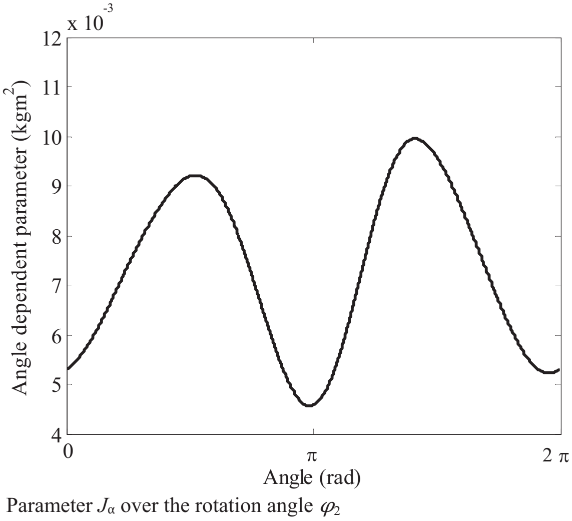


Fig. 2 Parameter $J_\alpha$ over the rotation angle $\varphi_2$

These angle-dependent parameters can then be expanded into a Fourier series. Here, the exponential Fourier series is used for an easier handling of the derivatives. Expanding system parameters into Fourier systems is quite common for describing rotational systems (see, for example, [2], [4] and [20]).

The system dynamics are described by the nonlinear state-space model

$$\dot{\boldsymbol{x}} = \boldsymbol{A}(\boldsymbol{x})\,\boldsymbol{x} + \boldsymbol{B}u\,, \tag{4}$$

where $\boldsymbol{x}$ is the state vector, $u$ the input of the system and $y$ the output. The "system matrix" $\boldsymbol{A}(\boldsymbol{x})$ is assumed to depend on linear combinations of the state vectors according to

$$\boldsymbol{A}(\boldsymbol{x}) = \boldsymbol{A}_0 + \bar{\boldsymbol{A}}\left(\boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{x},\dots,\boldsymbol{v}_N^{\mathrm{T}}\boldsymbol{x}\right) + \tilde{\boldsymbol{A}}\left(\boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{x},\dots,\boldsymbol{v}_N^{\mathrm{T}}\boldsymbol{x}\right). \tag{5}$$

The matrix $\boldsymbol{A}_0$ is constant (and therefore, corresponds to the "linear part" of the system dynamics), the matrix $\bar{\boldsymbol{A}}$ depends periodically on linear combinations of the state vector and the matrix $\tilde{\boldsymbol{A}}$ depends periodically and linearly on the linear combinations of the state vector.

The matrices $\bar{\boldsymbol{A}}(\boldsymbol{x})$ and $\tilde{\boldsymbol{A}}(\boldsymbol{x})$ in the model are expressed as

$$\begin{aligned}\bar{\boldsymbol{A}}(\boldsymbol{x}) &= \bar{\boldsymbol{A}}\left(\boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{x},\dots,\boldsymbol{v}_N^{\mathrm{T}}\,\boldsymbol{x}\right)\\ &= \sum_{l=1}^{N}\sum_{\mu=-L}^{L}\bar{\boldsymbol{M}}_{l,\mu}\,\mathrm{e}^{\,\mathrm{j}\mu\,\boldsymbol{v}_l^{\mathrm{T}}\boldsymbol{x}}\\ &= \sum_{l=1}^{N}\left[\bar{\boldsymbol{M}}_{l,-L}\;\cdots\;\bar{\boldsymbol{M}}_{l,L}\right]\left(\begin{bmatrix}\mathrm{e}^{\left(-L\,\mathrm{j}\boldsymbol{v}_l^{\mathrm{T}}\boldsymbol{x}\right)}\\ \vdots\\ \mathrm{e}^{\left(L\mathrm{j}\boldsymbol{v}_l^{\mathrm{T}}\boldsymbol{x}\right)}\end{bmatrix}\otimes\mathbf{I}_n\right)\end{aligned} \tag{6}$$

and

$$\begin{aligned}&\tilde{\boldsymbol{A}}(\boldsymbol{x}) =\\ &= \tilde{\boldsymbol{A}}\left(\boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{x},\dots,\boldsymbol{v}_N^{\mathrm{T}}\boldsymbol{x}\right)\\ &= \sum_{h=1}^{N}\sum_{l=1}^{N}\sum_{\mu=-L}^{L}\tilde{\boldsymbol{M}}_{h,l,\mu}\;\mathrm{e}^{\,\mathrm{j}\mu\,\boldsymbol{v}_l^{\mathrm{T}}\boldsymbol{x}}\;\boldsymbol{v}_h^{\mathrm{T}}\boldsymbol{x}\\ &= \sum_{h=1}^{N}\sum_{l=1}^{N}\left[\tilde{\boldsymbol{M}}_{h,l,-L}\;\cdots\;\tilde{\boldsymbol{M}}_{h,l,L}\right]\left(\begin{bmatrix}\mathrm{e}^{\left(-L\mathrm{j}\boldsymbol{v}_l^{\mathrm{T}}\boldsymbol{x}\right)}\\ \vdots\\ \mathrm{e}^{\left(L\mathrm{j}\boldsymbol{v}_l^{\mathrm{T}}\boldsymbol{x}\right)}\end{bmatrix}\otimes\left(\boldsymbol{v}_h^{\mathrm{T}}\boldsymbol{x}\mathbf{I}_n\right)\right).\end{aligned} \tag{7}$$

Here, $\otimes$ stands for the Kronecker product [21], $L$ is the order of the Fourier series and $N$ the number of the linear combinations of the state vector that the matrices depend upon. The linear combinations are expressed as $\boldsymbol{v}_p^{\mathrm{T}}\boldsymbol{x}$ with $p = 1,\dots,N$.

The matrices $\bar{\boldsymbol{M}}_{l,\mu}$ and $\tilde{\boldsymbol{M}}_{h,l,\mu}$ result from expanding the periodic system parameters into Fourier series. Since the exponential Fourier series representation is used, it holds that

$$\operatorname{Re}\bar{\boldsymbol{M}}_{l,\mu} = \operatorname{Re}\bar{\boldsymbol{M}}_{l,-\mu}\,,\ \operatorname{Re}\tilde{\boldsymbol{M}}_{h,l,\mu} = \operatorname{Re}\tilde{\boldsymbol{M}}_{h,l,-\mu} \tag{8}$$

$$\operatorname{Im}\bar{\boldsymbol{M}}_{l,\mu} = -\operatorname{Im}\bar{\boldsymbol{M}}_{l,-\mu}\,,\ \operatorname{Im}\bar{\boldsymbol{M}}_{h,l,\mu} = -\operatorname{Im}\bar{\boldsymbol{M}}_{h,l,-\mu} \tag{9}$$

and

$$\operatorname{Im}\bar{\boldsymbol{M}}_{l,0} = \boldsymbol{0}\,,\ \operatorname{Im}\bar{\boldsymbol{M}}_{h,l,0} = \boldsymbol{0}\,. \tag{10}$$

## III. Derivation of the Control Law

Consider a completely controllable SISO system of the form already introduced above, namely

$$\dot{\boldsymbol{x}} = \boldsymbol{A}(\boldsymbol{x})\,\boldsymbol{x} + \boldsymbol{B}u \tag{11}$$

with the output equation

$$y = \boldsymbol{C}\boldsymbol{x}\,. \tag{12}$$

The first step in the derivation of the control law in feedback linearization consists of differentiating the output $y$ until the input signal $u$ shows up in the derivative. The amount of times the output has to be differentiated is known as the relative degree. In the following, it is assumed that the relative degree is two.
The first derivative of $y$ with respect to the time $t$ is

$$\dot{y} = \boldsymbol{C}\,\dot{\boldsymbol{x}} = \boldsymbol{C}\boldsymbol{A}(\boldsymbol{x})\boldsymbol{x} + \boldsymbol{C}\boldsymbol{B}u\,. \tag{13}$$

Since the relative degree is two, it holds that

$$\boldsymbol{C}\boldsymbol{B} = 0\,. \tag{14}$$

The second derivative is then given by

$$\begin{aligned}\ddot{y} &= \boldsymbol{C}\boldsymbol{A}(\boldsymbol{x})\dot{\boldsymbol{x}} + \boldsymbol{C}\dot{\boldsymbol{A}}(\boldsymbol{x})\boldsymbol{x} \\ &= \boldsymbol{C}\boldsymbol{A}(\boldsymbol{x})\big(\boldsymbol{A}(\boldsymbol{x})\boldsymbol{x} + \boldsymbol{B}u\big) + \boldsymbol{C}\dot{\boldsymbol{A}}(\boldsymbol{x})\boldsymbol{x}.\end{aligned} \tag{15}$$

To obtain the second derivative of $y$, the derivative of $\boldsymbol{A}$ with respect to the time $t$ is needed. Using standard rules from matrix differential calculus [21], the time derivative of $\boldsymbol{A}$ follows as

$$\begin{aligned}\dot{\boldsymbol{A}}(\boldsymbol{x}) &= \frac{\partial \boldsymbol{A}(\boldsymbol{x})}{\partial \boldsymbol{x}^{\mathrm{T}}}\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \frac{\partial \boldsymbol{A}(\boldsymbol{x})}{\partial \boldsymbol{x}^{\mathrm{T}}}\big(\dot{\boldsymbol{x}} \otimes \mathbf{I}_n\big) \\ &= \frac{\partial \boldsymbol{A}(\boldsymbol{x})}{\partial \boldsymbol{x}^{\mathrm{T}}}\big(\boldsymbol{A}(\boldsymbol{x})\boldsymbol{x} \otimes \mathbf{I}_n + \boldsymbol{B}u \otimes \mathbf{I}_n\big).\end{aligned} \tag{16}$$

From the system description (5), the derivative of $\boldsymbol{A}$ with respect to the transposed state vector $\boldsymbol{x}$ follows as

$$\frac{\partial \boldsymbol{A}(\boldsymbol{x})}{\partial \boldsymbol{x}^{\mathrm{T}}} = \frac{\partial \boldsymbol{A}_0}{\partial \boldsymbol{x}^{\mathrm{T}}} + \frac{\partial \bar{\boldsymbol{A}}(\boldsymbol{x})}{\partial \boldsymbol{x}^{\mathrm{T}}} + \frac{\partial \tilde{\boldsymbol{A}}(\boldsymbol{x})}{\partial \boldsymbol{x}^{\mathrm{T}}}\,. \tag{17}$$

Since $\boldsymbol{A}_0$ is constant, it follows that

$$\frac{\partial \boldsymbol{A}_0}{\partial \boldsymbol{x}^{\mathrm{T}}} = \boldsymbol{0}\,. \tag{18}$$

Using the matrix chain rule and the matrix product rule [21], the second and the third term on the right hand side of (13) become

$$\begin{aligned}\frac{\partial \bar{\boldsymbol{A}}(\boldsymbol{x})}{\partial \boldsymbol{x}^{\mathrm{T}}} &= \sum_{p=1}^{N} \frac{\partial \bar{\boldsymbol{A}}(\boldsymbol{x})}{\partial \boldsymbol{v}_p^{\mathrm{T}}\boldsymbol{x}} \frac{\mathrm{d}\,\boldsymbol{v}_p^{\mathrm{T}}\boldsymbol{x}}{\mathrm{d}\,\boldsymbol{x}^{\mathrm{T}}} \\ &= \sum_{p=1}^{N} \frac{\partial \bar{\boldsymbol{A}}(\boldsymbol{x})}{\partial \boldsymbol{v}_p^{\mathrm{T}}\boldsymbol{x}} (\boldsymbol{v}_p^{\mathrm{T}} \otimes \mathbf{I}_n)\end{aligned} \tag{19}$$

and

$$\begin{aligned}\frac{\partial \tilde{\boldsymbol{A}}(\boldsymbol{x})}{\partial \boldsymbol{x}^{\mathrm{T}}} &= \sum_{p=1}^{N} \frac{\partial \tilde{\boldsymbol{A}}(\boldsymbol{x})}{\partial \boldsymbol{v}_p^{\mathrm{T}}\boldsymbol{x}} \frac{\mathrm{d}\,\boldsymbol{v}_p^{\mathrm{T}}\boldsymbol{x}}{\mathrm{d}\,\boldsymbol{x}^{\mathrm{T}}} \\ &= \sum_{p=1}^{N} \frac{\partial \tilde{\boldsymbol{A}}(\boldsymbol{x})}{\partial \boldsymbol{v}_p^{\mathrm{T}}\boldsymbol{x}} (\boldsymbol{v}_p^{\mathrm{T}} \otimes \mathbf{I}_n)\,.\end{aligned} \tag{20}$$

To calculate these derivatives, the derivatives of the matrices $\bar{\boldsymbol{A}}(\boldsymbol{x})$ and $\tilde{\boldsymbol{A}}(\boldsymbol{x})$ with respect to $\boldsymbol{v}_p^{\mathrm{T}}\boldsymbol{x}$ are required. These derivatives can be easily calculated from (6) and (7) and are given as

$$\begin{aligned}\frac{\partial \bar{\boldsymbol{A}}(\boldsymbol{x})}{\partial \boldsymbol{v}_p^{\mathrm{T}}\boldsymbol{x}} &= \sum_{\mu=-L}^{L} \bar{\boldsymbol{M}}_{p,\mu}\; \mathrm{j}\mu\; \mathrm{e}^{\,\mathrm{j}\mu\, \boldsymbol{v}_p^{\mathrm{T}}\boldsymbol{x}} \\ &= \begin{bmatrix} \bar{\boldsymbol{M}}_{p,-L} & \cdots & \bar{\boldsymbol{M}}_{p,L} \end{bmatrix} \left( \begin{bmatrix} -L\mathrm{j}\cdot \mathrm{e}^{\left(-L\,\mathrm{j}\boldsymbol{v}_p^{\mathrm{T}}\boldsymbol{x}\right)} \\ \vdots \\ L\mathrm{j}\cdot \mathrm{e}^{\left(L\,\mathrm{j}\boldsymbol{v}_p^{\mathrm{T}}\boldsymbol{x}\right)} \end{bmatrix} \otimes \mathbf{I}_n \right)\end{aligned} \tag{21}$$

and

$$\begin{aligned}&\frac{\partial \tilde{\boldsymbol{A}}(\boldsymbol{x})}{\partial \boldsymbol{v}_p^{\mathrm{T}}\boldsymbol{x}} = \\ &= \sum_{h=1}^{N}\sum_{l=1}^{N}\sum_{\mu=-L}^{L} \tilde{\boldsymbol{M}}_{h,l,\mu} \left( \mathrm{e}^{\,\mathrm{j}\mu\, \boldsymbol{v}_l^{\mathrm{T}}\boldsymbol{x}} \frac{\partial\, \boldsymbol{v}_h^{\mathrm{T}}\boldsymbol{x}}{\partial\, \boldsymbol{v}_p^{\mathrm{T}}\boldsymbol{x}} + \frac{\partial\, \mathrm{e}^{\,\mathrm{j}\mu\, \boldsymbol{v}_l^{\mathrm{T}}\boldsymbol{x}}}{\partial\, \boldsymbol{v}_p^{\mathrm{T}}\boldsymbol{x}}\, \boldsymbol{v}_h^{\mathrm{T}}\boldsymbol{x} \right) \\ &= \sum_{l=1}^{N}\sum_{\mu=-L}^{L} \tilde{\boldsymbol{M}}_{p,l,\mu}\, \mathrm{e}^{\,\mathrm{j}\mu\, \boldsymbol{v}_l^{\mathrm{T}}\boldsymbol{x}} + \sum_{h=1}^{N}\sum_{\mu=-L}^{L} \tilde{\boldsymbol{M}}_{h,i,\mu}\, \mathrm{j}\mu\, \mathrm{e}^{\,\mathrm{j}\mu\, \boldsymbol{v}_p^{\mathrm{T}}\boldsymbol{x}}\; \boldsymbol{v}_h^{\mathrm{T}}\boldsymbol{x} \\ &= \sum_{l=1}^{N} \begin{bmatrix} \tilde{\boldsymbol{M}}_{p,l,-L} & \cdots & \tilde{\boldsymbol{M}}_{p,\,l,L} \end{bmatrix} \left( \begin{bmatrix} \mathrm{e}^{-\mathrm{j}\,L\boldsymbol{v}_l^{\mathrm{T}}\boldsymbol{x}} \\ \vdots \\ \mathrm{e}^{\,\mathrm{j}\,L\boldsymbol{v}_l^{\mathrm{T}}\boldsymbol{x}} \end{bmatrix} \otimes \mathbf{I}_n \right) + \\ &+ \sum_{h=1}^{N} \begin{bmatrix} \tilde{\boldsymbol{M}}_{h,\,p,-L} & \cdots & \tilde{\boldsymbol{M}}_{h,\,p,L} \end{bmatrix} \cdot \\ &\qquad\cdot \left( \begin{bmatrix} -\mathrm{j}L\,\mathrm{e}^{-\mathrm{j}\,L\boldsymbol{v}_p^{\mathrm{T}}\boldsymbol{x}} \\ \vdots \\ \mathrm{j}L\,\mathrm{e}^{\,\mathrm{j}\,L\boldsymbol{v}_p^{\mathrm{T}}\boldsymbol{x}} \end{bmatrix} \otimes \left(\boldsymbol{v}_h^{\mathrm{T}}\boldsymbol{x}\,\mathbf{I}_n\right) \right).\end{aligned} \tag{22}$$

From (21) and (22) the advantage of the description in the introduced matrix notation becomes obvious because the matrices remain unchanged since their elements are constant. Only the vectors including the exponential functions of the Fourier series and the second factor of the Kronecker product change due to the differentiation.

With (16) to (22) all necessary expressions are given to compute the control law

$$\begin{aligned}u = &\frac{-\boldsymbol{C}\left(\dfrac{\partial \boldsymbol{A}(\boldsymbol{x})}{\partial \boldsymbol{x}^{\mathrm{T}}}\big(\boldsymbol{A}(\boldsymbol{x})\boldsymbol{x} \otimes \boldsymbol{I}_n\big) + \boldsymbol{A}^2(\boldsymbol{x})\right)\boldsymbol{x}}{\boldsymbol{C}\left(\dfrac{\partial \boldsymbol{A}(\boldsymbol{x})}{\partial \boldsymbol{x}^{\mathrm{T}}}\big(\boldsymbol{B} \otimes \boldsymbol{I}_n\big)\boldsymbol{x} + \boldsymbol{A}(\boldsymbol{x})\boldsymbol{B}\right)} + \\ &+ \frac{-\gamma_1 \boldsymbol{C}\boldsymbol{A}(\boldsymbol{x})\boldsymbol{x} - \gamma_0 \boldsymbol{C}\boldsymbol{x} + \beta w}{\boldsymbol{C}\left(\dfrac{\partial \boldsymbol{A}(\boldsymbol{x})}{\partial \boldsymbol{x}^{\mathrm{T}}}\big(\boldsymbol{B} \otimes \boldsymbol{I}_n\big)\boldsymbol{x} + \boldsymbol{A}(\boldsymbol{x})\boldsymbol{B}\right)},\end{aligned} \tag{23}$$

which can be calculated from (15). Applying the control law (23) leads to the resulting closed loop dynamics described by

$$\ddot{y} = -\gamma_1\dot{y} - \gamma_0 y + \beta w \tag{24}$$

As mentioned above, the derivation shown here is for systems with a relative degree of two. An extension to systems with a higher relative degree is fairly straightforward.

## IV. Simulation Studies

The control design described in the previous section is applied to an industrial machine in a simulation study. A schematic diagram of the model is shown in Fig. 3.

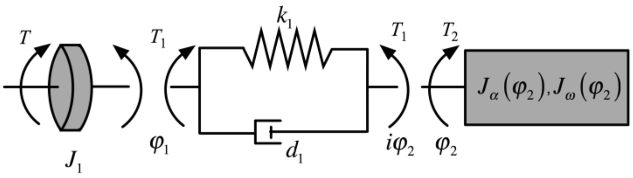


Fig. 3 Schematic diagram of the industrial machine (belt not shown)

The dynamics of this model are described by

$$J_1 \ddot{\varphi}_1 = T - T_1 , \tag{25}$$

$$J_\alpha(\varphi_2)\ddot{\varphi}_2 + J_\omega(\varphi_2)\dot{\varphi}_2^2 = T_2 \tag{26}$$

with

$$T_1 = k_1(\varphi_1 - i\varphi_2) + d_1(\dot{\varphi}_1 - i\dot{\varphi}_2), \tag{27}$$

and

$$T_2 = iT_1 \ . \tag{28}$$

In this model, $T$ is the driving torque generated from an electrical machine (not included in the model) and $T_1$ and $T_2$ are inner torques that are transmitted through a belt drive (not shown in the diagram) with gear ratio $i$; $\varphi_1$ and $\varphi_2$ are the rotation angles of the drive side and the load side; $J_1$ is the moment of inertia at the driving end, $k_1$ is the stiffness and $d_1$ the damping of the system; $J_\alpha$ and $J_\omega$ are angle-dependent parameters that can be calculated from the construction data of the machine and were provided by an industrial partner that manufactures this type of machine. The variations of the parameters $J_\alpha$ over the angle $\varphi_2$ and $J_\omega$ over the angle $\varphi_2$ are shown in Fig. 2 and Fig. 4, respectively.

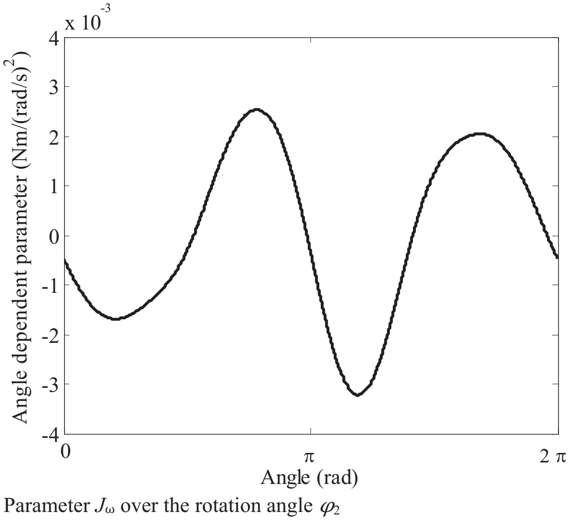


Fig. 4 Parameter $J_\omega$ over the rotation angle $\varphi_2$

If the system is driven by a constant input torque, the resulting angular speed fluctuates around a mean value. This undesirable behavior is shown in Fig. 5. The potential of standard cascaded PID/PI speed control for reducing these oscillations is limited. The tuning of PID/PI control for such systems is difficult since certain settings of the controller parameters can even amplify the oscillations in certain operating conditions.

The system equations can be written as the state space model given by (11) and (12) with

$$\boldsymbol{x} = \begin{bmatrix} \varphi_1 & \dot{\varphi}_1 & \varphi_2 & \dot{\varphi}_2 \end{bmatrix}^{\mathrm{T}}, \tag{29}$$

$$\boldsymbol{A}(\boldsymbol{x}) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ \frac{-k_1}{J_1} & \frac{-d_1}{J_1} & \frac{ik_1}{J_1} & \frac{id_1}{J_1} \\ 0 & 0 & 0 & 1 \\ \frac{ik_1}{J_\alpha(x_3)} & \frac{id_1}{J_\alpha(x_3)} & -\frac{i^2 k_1}{J_\alpha(x_3)} & -\frac{i^2 d_1 + J_\omega(x_3) x_4}{J_\alpha(x_3)} \end{bmatrix}, \tag{30}$$

$$\boldsymbol{B} = \begin{bmatrix} 0 & 1/J_1 & 0 & 0 \end{bmatrix}^{\mathrm{T}}, \tag{31}$$

and

$$\boldsymbol{C} = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}. \tag{32}$$

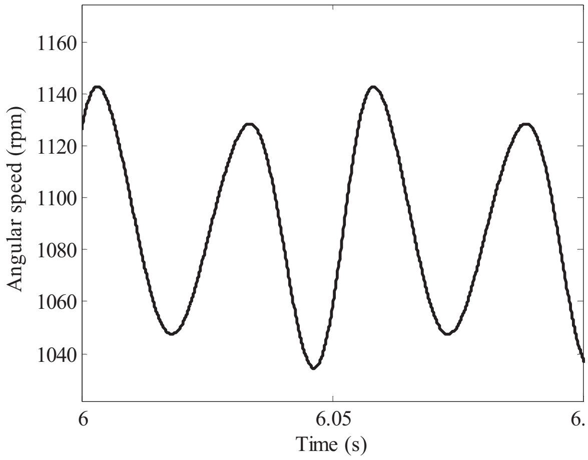


Fig. 5 Angular speed $\dot{\varphi}_2$ if the system is driven by constant torque

From (30) it can be seen that the behavior of the considered system depends on the rotation angle $\varphi_2$ which corresponds to the state $x_3$ and its derivative $\dot{\varphi}_2$ which corresponds to the state $x_4$.

For this system the state space representation of the form introduced in Sec. II becomes

$$\dot{\boldsymbol{x}} = \left(\boldsymbol{A}_0 + \bar{\boldsymbol{A}}(\boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{x}, \boldsymbol{v}_2^{\mathrm{T}}\boldsymbol{x}) + \tilde{\boldsymbol{A}}(\boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{x}, \boldsymbol{v}_2^{\mathrm{T}}\boldsymbol{x})\right)\boldsymbol{x} + \boldsymbol{B}u \tag{33}$$

$$y = \boldsymbol{C}\boldsymbol{x}, \tag{34}$$

with the vectors

$$\mathbf{v}_1 = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix}^{\mathrm{T}} \tag{35}$$

and

$$\mathbf{v}_2 = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}^{\mathrm{T}} \tag{36}$$

the relevant states $x_3$ and $x_4$ can be selected. Thus, $N = 2$ .

From the system matrix in (30) the linear part of the system matrix can be read off directly and gives

$$\boldsymbol{A}_0 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -k_1/J_1 & -d_1/J_1 & ik_1/J_1 & id_1/J_1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}. \tag{37}$$

As introduced in (5), the nonlinear parts of (30) are divided into a part $\bar{\boldsymbol{A}}$ with a harmonic dependence on the states and $\tilde{\boldsymbol{A}}$ with a harmonic and a linear dependence on the states. These nonlinear parts can be represented by a Fourier series.

The matrix coefficients $\bar{\boldsymbol{M}}_{l,\mu}$ for the matrix $\bar{\boldsymbol{A}}$ in (6) are obtained as

$$\bar{\boldsymbol{M}}_{1,\mu} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ ik_1\bar{m}_\mu & id_1\bar{m}_\mu & -i^2k_1\bar{m}_\mu & -d_1i^2\bar{m}_\mu \end{bmatrix} \tag{38}$$

and

$$\bar{\boldsymbol{M}}_{2,\mu} = \boldsymbol{0}\,, \tag{39}$$

where the scalars $\bar{m}_\mu$ are the coefficients of the Fourier expansion of $J_\alpha^{-1}$. Similarly, the matrix coefficients for $\tilde{\boldsymbol{A}}$ in (7) are obtained as

$$\tilde{\boldsymbol{M}}_{2,1,\mu} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \tilde{m}_\mu \end{bmatrix}, \tag{40}$$

$$\tilde{\boldsymbol{M}}_{1,1,\mu} = \tilde{\boldsymbol{M}}_{1,2,\mu} = \tilde{\boldsymbol{M}}_{2,2,\mu} = \boldsymbol{0}\,, \tag{41}$$

where the scalars $\tilde{m}_\mu$ are the coefficients of the Fourier expansion of $J_\omega / J_\alpha$. Fourier series of eight order are used, that is, $L = 8$ .

The simulated control result for the model of the industrial machine is shown in Fig. 6. The effects of the nonlinearities are attenuated by the calculated control law with feedback linearization. Also the desired linear behavior as a second order system is achieved.

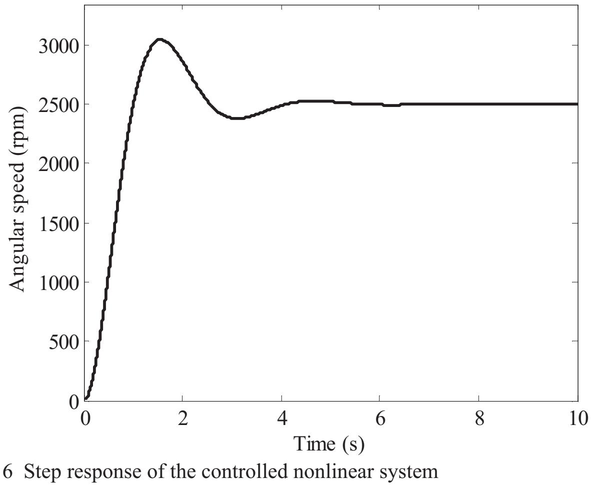


Fig. 6 Step response of the controlled nonlinear system

In Fig. 7 it can be seen that the angular speed at the drive side is still fluctuating while the angular speed at the load side is constant.

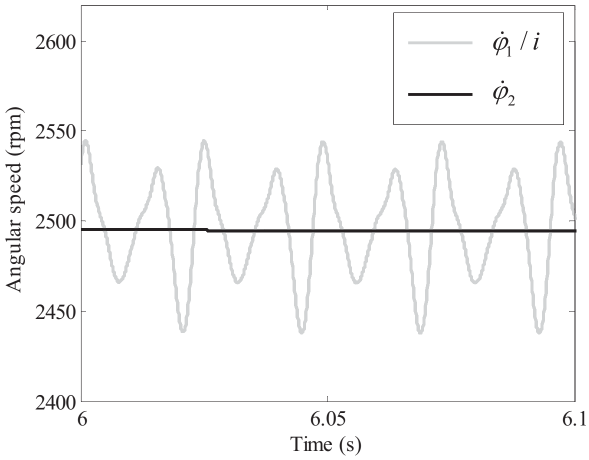


Fig. 7 Fluctuating angular speed $\dot{\varphi}_1$ at the drive side (related to the load side by including the gear ratio *i*) and constant angular speed $\dot{\varphi}_2$ at the load side

The control signal for the considered system to achieve the desired behavior is shown in Fig. 8. No calculations are needed to obtain the derivatives and the control law since the matrix coefficients can be simply plugged into the corresponding equations and the control law is readily obtained.

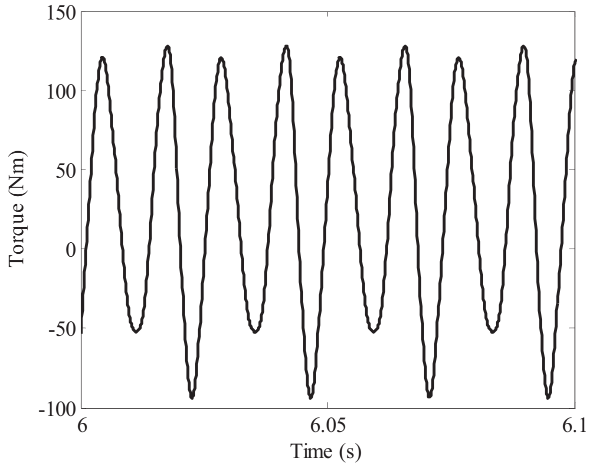


Fig. 8 Control signal to achieve the desired behavior on the output

## V. Summary and Conclusion

In this paper, the active control of speed fluctuations in rotating machines has been considered. After a simple motivating example, a very general state-space description of the considered systems was introduced. With this system description, it is very easy to derive a control law that eliminates the speed fluctuations and results in a second-order linear dynamics for the controlled system. The proposed control law has been applied to a model that represents the behavior of an industrial machine that shows undesired speed fluctuations. To the best of the authors knowledge, this is the first time that feedback linearization has been used for active vibration control for such systems.

The steps that need to be carried out in the control design process consist of transforming the model equations into the general form. This involves reading off entries from the model equations and expanding periodic model parameters into Fourier series. This can be easily done using the standard discrete Fourier transform.

Although the simulation results show that the approach works well (as predicted by theory), it is stressed that this is a preliminary study. The setup considered here corresponds to an ideal case (noise free measurements, no actuator and sensor dynamics included, all system parameters exactly known, all states available for feedback, continuous-time control possible). The applicability of this approach in more realistic cases will be investigated in future studies, where robustness issues (sensitivity to noise and model inaccuracies), implementation issues (discretization of the control algorithm) and extensions (including actuator and sensor dynamics, using an observer to reconstruct unmeasured state variables) will be investigated. The main contribution of this paper is the idea of using feedback linearization for the active control of torsional vibrations.

## REFERENCES

[1] Huang, C.-K., P.-Y. Yu and H.-C. Chen. 2007. Robust BDCM sensorless control with position dependent load torque. *Proceedings of the Power Elektronics Specialists Conference* Orlando, June 2007. 2739-44.

[2] Rizzoni, G. 1989. Estimate of indicated torque from crankshaft speed fluctuations: A model for the dynamics of the IC engine. *IEEE Transactions on Vehicular Technology* 38:168-179.

[3] Huang, Y. and T. Wang. 2010. The study about torsional vibration characteristics and its optimization of vehicle transmission system. *Proceedings of the 3rd International Conference on Advanced Computer Theory and Engineering)*. Chengdu, China, August 2010. V2-358-62.

[4] Östman, F. and H. T. Toivonen. 2008. Model based torsional vibration control of internal combustion engines. *IET Control Theory and Applications* 2:1024-32.

[5] Zhu, Z. Q. and J. H. Leong. 2011. Analysis and mitigation of torsional vibration of PM brushless DC drives with direct torque controller. *Proceedings of the Energy Conversion Congress and Exposition*. Phoenix, September 2011. 1502-9.

[6] Lefevre, Y., B. Davat and M. Lajoie-Mazenc. 1989. Determination of synchronous motor vibrations due to electromagnetic force harmonics. *IEEE Transactions of Magnetics* 24:2974-76.

[7] Zhang, R., Z. Chen, Y. Yang and C. Tong. 2007. Torsional vibration suppression control in the main drive system of rolling mill by state feedback speed controller based on extended state observer. *Proceedings of the IEEE International Conference on Control and Automation*. Guangzhou, China, May-June 2007. 2172-77.

[8] Xiaoyan, X., X. Shibo, C. Dongbing and W. Ranfeng. 2010. Analysis of the self-excited vibration and dynamics modification for rolling mills. *Proceedings of the 11th International Conference on Probabilistic Methods applied on Power Systems*. Singapore, June 2010. 304-7.

[9] Valenzuela, M. A., J. M. Bentley and R. D. Lorenz. 2005. Evaluation of torsional oscillations in paper machine sections. *IEEE Transactions on Industry Applications* 41:493-501.

[10] Michael, C. A. and A. N. Safacas. 2007. Dynamic and vibration analysis of a multimotor DC drive system with elastic shafts driving a tissue paper machine. *IEEE Transactions on Industrial Electronics* 54:2033-46.

[11] Xiang, L., X. Chen and G. Tang. 2009. The torsional vibration of turbo-generator groups in mechanically and electrically coupled influences. *Proceedings of the 2nd International Congress on Image and Signal Processing*. Tianjin, China, October 2009, 1-4.

[12] Dutka, A. and M. Orkisz. 2011. Analysis and remedies for torsional oscillations in rotating machinery. *Proceedings of the IEEE International Symposium on Diagnostics for Electric Machines, Power Electronics & Drives*. Bologna, September 2011. 474-481.

[13] Muszynski, R. and J. Deskur. 2010. Damping of torsional vibrations in high-dynamic industrial drives. *IEEE Transactions on Industrial Electronics* 57:544-52.

[14] Orlowska-Kowalska, T. and K. Szabat. 2008. Damping of torsional vibrations in two-mass system using adaptive sliding neuro-fuzzy approach. *IEEE Transactions on Industrial Informatics* 4:47-57.

[15] Vittek, J., P. Makys, M. Stulrajter, S. J. Dodds and R. Perryman. 2008. Comparison of sliding mode and forced dynamics control of electric drive with a flexible coupling employing PMSM. *Proceedings of the IEEE International Conference on Industrial Technology*. Chengdu, China, April 2008. 1-6.

[16] Zaremba, A. T., I. V. Burkov and R. M. Stuntz. 1998. Active damping of engine speed oscillations based on learning control. *Proceedings of the American Control Conference*. Philadelphia, June 1998. 2143-47.

[17] Guzzella, L. and A. M. Schmid.1995. Feedback linearization of spark-ignition engines with continuously variable transmissions. *IEEE Transactions on Control Systems Technology* 3:54-60.

[18] Shigehiro, S. and O. Hiromitsu. 2008. Design of starting controller for spark ignition engines based on adaptive feedback linearization. *Proceedings of the 27th Chinese Control Conference*. Kunming, Yunnan, China, July 2008. 566-71.

[19] Buckner, G. D., K. T. Schuetze and J. H. Beno. 2000. Active vehicle suspension control using intelligent feedback linearization. *Proceedings of the American Control Conference*. Chicago, June 2000. 4014-18.

[20] Burkov, I. V., and A. T. Zaremba. 1999. Adaptive control for angle speed oscillations generated by periodic disturbances. *Proceedings of the 6th St. Petersburg Symposium on Adaptive Systems Theory*. St. Petersburg, Russia, September 1999. 34–36.

[21] Weinmann, A. 1991. *Uncertain models and Robust Control*. Wien: Springer.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# An adaptive sliding mode approach to decentralized control of uncertain systems

Zheng Huang, Ron J Patton
Department of Engineering
University of Hull
Hull, UK
Z.Huang@2009.hull.ac.uk; r.j.patton@hull.ac.uk

*Abstract*—**This paper proposes a systematic adaptive sliding mode controller design for the decentralized system with nonlinear interactions and unmatched uncertainties. An adaptive tuning approach is developed to deal with unknown but bounded uncertainties/interactions. The sliding surface is designed which obviates the use of regular transformation, by solving a simple LMI-based optimization problem. The feasibility of the LMIs is also discussed in this paper. Finally, a numerical example is used to illustrate this method.**

***Keywords-Decentralized control, sliding mode, unmatched unccertainties/interactions, adaptive control***

## I. Introduction

Decentralized control has gained considerable attention in the literature for two decades. [1]-[9]. The main idea of decentralized control is to use only local information at the level of each subsystem in the controller for large-scale interconnected systems. This feature can overcome the limitations of the traditional centralized control or partial decentralized control [10] that requires large communication bandwidth to exchange information between subsystems and controller. The decentralized control has much simpler control structure and more practical approach than centralized controller.

On the other hand, sliding mode control (SMC), as a powerful robust control method, has been widely researched,[11], [12]. When using sliding mode, there are two steps i). sliding surface design and ii). control law design. The system states will be driven to the sliding surface and be maintained on it. Once the system is running in the sliding surface, the system is insensitive to the matched perturbations (perturbations coming from input channels). Edwards and Spurgeon [11] develop their approach to the classical sliding mode control design algorithm by introducing a so called *"regular form"* to set up a decomposition comprising matched/unmatched state space components. This approach can be considered too complex for a single SMC system and this motivates the use of an alternative approach obviating a need for transformations, whilst still satisfying matching condition properties.

Choi [12],[13] proposed another SMC approach in which the sliding surface can be designed by solving a simple LMI problem. Although no transformations are used Choi's work is focused on SMC of single rather than decentralized control systems. The proposal here is to further develop Choi's approach with application to uncertain decentralized systems.

When applying sliding mode in the decentralized system, researchers consider the interactions as perturbations and try to eliminate or at least attenuate these perturbations. [9], [14]-[15]. However, the sliding mode can only deal with matched perturbations and the upper bound of perturbations should also be known. In this case, researchers start trying to combine the sliding mode with other robust control methods to overcome the unmatched problem limitation [9]. Šiljak [3] proposes a feedback control using an LMI approach that can deal with the interactions no matter whether or not they are matched. Also the assumption of unknown interactions in [3] is more general which includes both linear and non-linear types of interactions. Kalsi [5], [6] uses the control method of [3] to develop a sliding approach to observer-based control. Although a lot of research has been done in this area [9], [14]-[16], the main contribution of this paper is to use the LMI-based work of Choi [12], [13] to construct a simpler and more general approach to decentralized control. This will also provide an efficient strategy for accounting for modeling uncertainty and subsystem interactions. Also the known upper bound limitation is removed by an adaptive mechanism in our work. Although this paper focuses on the state feedback strategy, output feedback control can be easily formulated as an extension based on the proposed method. Moreover, other robust improvements can be extended based on this approach.

This paper is organized as follows. The basic assumptions are given and the design objective is proposed in Section II. Then the main results are given in Section III, where both the sliding surface and control law designs are represented. Also in Section III, the feasibility of LMIs is discussed and more constraints are proposed to improve the methods. A numerical example is given in Section V which demonstrates the efficacy of the techniques developed in this paper. Finally, the concluding remarks and further work are given in Section V.

## II. System description and Problem formation

Consider a class of perturbed large-scale systems which are comprise N-linked subsystems with uncertainties in the interactions. The dynamic equation of each subsystem is represented as:

Zhen Huang's research study at the University of Hull [2009-2012] is supported by the China Scholarship Council (CSC) and the Hull-China scholarship.

$$\dot{x}_i(t) = A_i x_i(t) + B_i(u_i(x_i,t) + f_i(x_i,t)) + h_i(x,t),$$
$$i = 1,2,\dots,N \qquad (1)$$

where $x_i(t) \in R^{n_i}$ is the state variable of the *i*-th subsystem, $u_i(x_i,t) \in R^{m_i}$ is the control input vector of the *i*-th subsystem, the matrix $A_i \in R^{n\times n}$ is the system characteristic matrix, and the matrix $B_i \in R^{n\times m}$ is the input matrix with full rank $m < n$. And $f_i(x,t)$ is any uncertainty or disturbance in the input channel. The term $h_i(x,t)$ reflects the interaction of the *i*-th subsystem with other subsystems and the uncertainty dynamics associated with the *i*-th subsystem itself.

We also assume the followings to be valid:

*1)* All the pair $(A_i, B_i), i = 1,2,\dots,N$ are controllable.

*2)* All the state variables $x_i$, $i = 1,2,\dots,N$ are locally available for measurement for all time.

*3)* The subsystem interactions are globally bounded, that is, $\|h_i(x,t)\| < \beta_i < \infty$ for some unknown constant $\beta_i > 0$. The interactions satisfy the same quadratic constraint as, for example [3],[5]-[7] , that is:

$$h_i^T(x,t)h_i(x,t) \le \alpha_i^2 x H_i^T H_i x \qquad (2)$$

*4)* The external disturbance $f_i(t)$ is bounded by a known constant $\varepsilon_i$, i.e. $\|f_i(x,t)\| < \varepsilon_i$

The overall interconnected system can be rewritten in a compact form as:

$$\dot{x}(t) = Ax(t) + B\big(u(x,t) + f(t)\big) + h(x,t) \qquad (3)$$

where $A = diag(A_1,\dots,A_N)$, $B = diag(B_1,\dots,B_N)$, $f(t) = [f_1^T(t),\dots,f_N^T(t)]^{\mathrm{T}}$ and $h(x,t) = [h_1^T(x,t),\dots,h_N^T(x,t)]^{\mathrm{T}}$, and with the third assumption, the interconnections $h(x,t)$ are bounded as follows:

$$h^T(x,t)h(x,t) \le x\left(\sum_{i=1}^{N} \alpha_i^2 H_i^T H_i\right)x = xH^T Hx$$

where $\alpha_i$ is a bounding constant.

Because all of the subsystems are stabilizable, it is easy to verify that the overall system is controllable.

The objective is to design a totally decentralized sliding mode controller that robustly regulates the state of the overall system without any information exchange between the subsystems. And with this type of controllers, the overall system is robust to all the uncertainties and matched perturbations.

## III. Main result

It is well known that in SMC design, there are two steps: a). Sliding surface design and b) control law design. In this Section, the sliding surface is designed by an LMI and an adaptive control law is realised.

### *A. Sliding surface design*

Define $\Gamma$ as any basis of the null space of $B^T$, i.e. $\Gamma$ is an orthogonal complement of $B$.

Consider the following LMIs:

Minimize $\gamma_i$, subject to $X > 0$,

$$\begin{bmatrix} \Gamma^T(XA^T + AX + I)\Gamma & \Gamma^T XH_1^T & \cdots & \Gamma^T XH_N^T \\ H_1X\Gamma & -\gamma_1 I & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ H_N X\Gamma & 0 & \cdots & -\gamma_N I \end{bmatrix} < 0$$
$$(4)$$

Assume that the sliding surface $\sigma$ is given by

$$\sigma(x,t) = B^T X^{-1} x(t) \qquad (5)$$

where $X$ is a solution to the LMIs (4). This sliding surface is proposed in [12]

***Remark 1***: we should note that $X = diag(X_1,\dots,X_N)$ because that the LMIs (4) is based on the overall system (3). Also the sliding surface $\sigma(x,t) = [\sigma_1^T(x_1,t),\dots,\sigma_N^T(x_N,t)]^T$. The $\sigma_i(x_i,t)$ represents the sliding surface for the *i*-th subsystem.

***Theorem 1:*** Suppose the LMIs (4) have a solution $X$ and the sliding surface is given by Eq. (5). Then once the sliding surface (5) is reached and maintained thereafter, i.e. $\sigma(x,t) = 0$ and $\dot{\sigma}(\mathrm{x,t}) = 0$ and hence the overall system is stable.

***Proof:*** Define a transformation matrix and its inverse matrix as described in [12] as

$$T = \begin{bmatrix} \Gamma^T \\ S \end{bmatrix}, T^{-1} = [X\Gamma(\Gamma^T X\Gamma)^{-1} \quad B(SB)^{-1}] \qquad (6)$$

And the associated vector $z$ is given as

$$z(t) = \begin{bmatrix} z_1(\mathrm{t}) \\ z_2(\mathrm{t}) \end{bmatrix} = \begin{bmatrix} z_1(\mathrm{t}) \\ \sigma(t) \end{bmatrix} = Tx(t) \qquad (7)$$

where $z_1 \in \mathrm{R}^{\mathrm{n-m}}$ and $z_2 \in \mathrm{R}^{\mathrm{m}}$. And with the transformation (6), we can obtain a new state equation as:

$$\dot{z}(\mathrm{t}) = TAT^{-1}z(t) + TB(u(t) + f(x,t)) + Th(x,t) \quad (8)$$

where,

$$TAT^{-1} = \begin{bmatrix} \Gamma^T AX\Gamma(\Gamma^T X\Gamma)^{-1} & \Gamma^T AB(SB)^{-1} \\ \mathrm{S}A\Gamma(\Gamma^T X\Gamma)^{-1} & \mathrm{S}AB(SB)^{-1} \end{bmatrix}$$

$$TB = \begin{bmatrix} 0 \\ SB \end{bmatrix}$$

And note that once the system is on the sliding surface and maintained there, $\sigma(t) = 0$, $\dot{\sigma}(\mathrm{t}) = 0$. In this case, the system is insensitive to all the matched uncertainties or disturbances $f(x,t)$. The state equation (8) then becomes:

$$\dot{z}_1(\mathrm{t}) = \Gamma^T AX\Gamma(\Gamma^T X\Gamma)^{-1} z_1(t) + \Gamma^T h(x,t) \tag{9}$$

Now define a Lyapunov function for the system (9):

$$V(z_1) = z_1^T P z_1 \tag{10}$$

where P is a symmetric and positive definite (s.p.d) matrix.

The time derivative of $V(z_1)$ along the trajectories of Eq. (9) is given by:

$$\begin{aligned}\dot{V}(z_1) = z_1^T((\Gamma^T AX\Gamma(\Gamma^T X\Gamma)^{-1})^T P + P\Gamma^T AX\Gamma(\Gamma^T X\Gamma)^{-1})z_1 \\ +\mathrm{z}_1^{\mathrm{T}}\mathrm{P}\Gamma^T\mathrm{h} + \mathrm{h}^{\mathrm{T}}\Gamma P z_1\end{aligned} \tag{11}$$

To obtain a quadratic form of (11), we use the following result in [16]:

$$X^T Y + Y^T X \leq X^T X + Y^T Y$$

for any matrices (or vectors) $X$ and $Y$ with appropriate dimension.

It follows that:

$$z_1^T P\Gamma^T h + h^T \Gamma P z_1 \leq z_1^T P\Gamma^T \Gamma P z_1 + h^T h \tag{12}$$

Also, from the Assumption 3), the interactions satisfy the quadratic form:

$$h^T(x,t)h(x,t) \leq xH^T Hx = z^T (T^{-1})^T H^T H T^{-1} z \tag{13}$$

Because the system is in the sliding surface, $z_2 = \sigma = 0$. We can simplify the above inequality of interactions (13) as

$$h^T(x,t)h(x,t) \leq z_1^T(\Gamma^T X\Gamma)^{-1}\Gamma^T \mathrm{X} H^T HX\Gamma(\Gamma^T X\Gamma)^{-1} z_1 \tag{14}$$

Substitute the inequalities (12) and (14) into (11), we have a quadratic form of the derivative of the Lyapunov function as

$$\begin{aligned}\dot{V}(z_1) = z_1^T((\Gamma^T \mathrm{A}X\Gamma(\Gamma^T X\Gamma)^{-1})^{\mathrm{T}} P + P\Gamma^T \mathrm{A}X\Gamma(\Gamma^T X\Gamma)^{-1} + \\ P\Gamma^T \Gamma P + (\Gamma^T X\Gamma)^{-1}\Gamma^T \mathrm{X} H^T HX\Gamma(\Gamma^T X\Gamma)^{-1})z_1\end{aligned} \tag{15}$$

The stabilization of system (9) requires that:

$$\dot{V}(z_1) < 0 \tag{16}$$

for all $z_1 \neq 0$.

The development of (16) leads to

$$\begin{aligned}(\Gamma^T \mathrm{A}X\Gamma(\Gamma^T X\Gamma)^{-1})^{\mathrm{T}} P + P\Gamma^T \mathrm{A}X\Gamma(\Gamma^T X\Gamma)^{-1} + P\Gamma^T \Gamma P + \\ (\Gamma^T X\Gamma)^{-1}\Gamma^T \mathrm{X} H^T HX\Gamma(\Gamma^T X\Gamma)^{-1} < 0\end{aligned} \tag{17}$$

In this case, the problem is to find an s.p.d matrix $P$ such that the inequality (17) is satisfied. By pre-multiplying and post-multiplying the inverse matrix of $P$, we can simplify the inequality (17).

Define $Y = P^{-1}$, we have:

$$\begin{aligned}\mathrm{Y}(\Gamma^T \mathrm{A}X\Gamma(\Gamma^T X\Gamma)^{-1})^{\mathrm{T}} + \Gamma^T \mathrm{A}X\Gamma(\Gamma^T X\Gamma)^{-1}\mathrm{Y} + \Gamma^T \Gamma \\ + \mathrm{Y}(\Gamma^T X\Gamma)^{-1}\Gamma^T \mathrm{X} H^T HX\Gamma(\Gamma^T X\Gamma)^{-1} Y < 0\end{aligned}$$

We can easily find a matrix $\mathrm{Y} = \Gamma^T X\Gamma$ which satisfies the s.p.d condition so that the above inequality becomes:

$$\Gamma^T(\mathrm{A}X + XA^T + I + XH^T HX)\Gamma < 0 \tag{18}$$

With Assumption 3), we rewrite the inequality (18) as:

$$\Gamma^T\left(\mathrm{A}X + XA^T + I + \sum_{i=1}^{N}(\alpha_i^2 X H_i^T H_i X)\right)\Gamma < 0$$

By defining $\gamma_i = \frac{1}{\alpha_i^2}$ and using the Schur complement, the above inequality can be rewritten as the form of LMIs (4). □

After solving the LMIs (4), we have the s.p.d matrix $X$ which has the structure of $X = diag(X_1, \ldots, X_N)$. In this case, for each subsystem, the local sliding surface could be given by

$$\sigma_i(x_i,t) = S_i x_i(t) = B_i^T X_i^{-1} x_i(t) \tag{19}$$

*B. Adaptive Control law design*

It has proved above that once the system reaches the sliding surface and is maintained on it thereafter, the control objective is achieved. In the following a sliding control law is designed to drive the system to the sliding surface.

Introducing the control strategy for each subsystem:

$$u_i(x_i,t) = \begin{cases} -(S_i B_i)^{-1} A_i x_i(t) + \Psi_{\mathrm{i}} \frac{\sigma_i}{\|\sigma_i\|}, & \|\sigma_i\| \neq 0 \\ -(S_i B_i)^{-1} A_i x_i(t), & \|\sigma_i\| = 0 \end{cases} \tag{20}$$

where,

$$\Psi_{\mathrm{i}} = -\eta_{\mathrm{i}} - \|(\mathrm{SB})^{-1}\mathrm{S}\|\hat{\beta}_{\mathrm{i}}(\mathrm{t}),\ \dot{\hat{\beta}}_{\mathrm{i}}(\mathrm{t}) = \|\sigma_i\| \tag{21}$$

where, $\eta_{\mathrm{i}} > \varepsilon_{\mathrm{i}}$ is a positive constant. $\hat{\beta}_{\mathrm{i}}$ is the estimation of the upper bound of the interaction of *i*-th subsystem.

***Theorem 2***: For each subsystem of the form (1) and the sliding surface (19), by applying the control law (20) and (21) to each subsystem, the overall system trajectory converges to the sliding surface $\sigma_i(x_i,t) = 0$ in finite time and is maintained on the surface, i.e. $\dot{\sigma}_i(x_i,t) = 0$. Meanwhile, the system is insensitive to all the matched disturbances.

***Proof:*** To prove the reachability, we define the Lyapunov function with respect to the sliding surface as:

$$V = \sum_{i=1}^{N} \frac{1}{2}\left[\sigma_i^T (S_i B_i)^{-1}\sigma_i + \|(S_i B_i)^{-1} S\|\bar{\beta}_i^{\,2}\right] \tag{22}$$

The matrix $(S_i B_i)^{-1}$ satisfies the s.p.d condition since that $X_i^{-1}$is a s.p.d matrix and $(S_i B_i)^{-1} = (B_i^T X_i^{-1} B_i)^{-1} > 0$.

Moreover, we define the estimation error $\bar{\beta}_i(t) = \hat{\beta}(t) - \beta_i$.

Since $\beta_i$ is a constant, $\dot{\bar{\beta}}_i(t) = \dot{\hat{\beta}}_i(t) = \|\sigma_i\|$.

Differentiating (19) with respect to time yields:

$$\dot{\sigma}_i(x_i,t) = S_iA_ix_i(t) + S_iB_i(u_i(x_i,t) + f_i(t)) + S_ih_i(x,t)$$

Then taking the derivative of the Lyapunov function (22) we have:

$$\begin{aligned}\dot{V} &= \sum_{i=1}^{N}\left\{\sigma_i^T\left[\Psi_{\mathrm{i}}\frac{\sigma_i}{\|\sigma_i\|} + f_i(t) + (S_iB_i)^{-1}Sh_i(x,t)\right]\right.\\ &\qquad\left. + \|(S_iB_i)^{-1}S\|\bar{\beta}_i\dot{\bar{\beta}}_i\right\}\\ &\le \sum_{i=1}^{N}\left\{\sigma_i^T\left[\Psi_{\mathrm{i}}\frac{\sigma_i}{\|\sigma_i\|} + \varepsilon_i + (S_iB_i)^{-1}Sh_i(x,t)\right]\right.\\ &\qquad\left. + \|(S_iB_i)^{-1}S\|\bar{\beta}_i\|\sigma_{\mathrm{i}}\|\right\}\end{aligned} \tag{23}$$

We can easily verify that $\sigma_i^T\Psi_{\mathrm{i}}\frac{\sigma_i}{\|\sigma_i\|} = \Psi_{\mathrm{i}}\|\sigma_i\|$ , and on substituting (21) into (23), we have:

$$\begin{aligned}\dot{V} &\le \sum_{i=1}^{N}\{\Psi_{\mathrm{i}}\|\sigma_i\| + \|\sigma_i\|\varepsilon_{\mathrm{i}} + \|\sigma_i\|\|(S_iB_i)^{-1}S\|\beta_{\mathrm{i}}\\ &\qquad + \|(S_iB_i)^{-1}S\|\big(\hat{\beta}(t) - \beta_i\big)\|\sigma_{\mathrm{i}}\|\}\\ &\le \sum_{i=1}^{N}\{-\eta_{\mathrm{i}}\|\sigma_i\| - \|\sigma_i\|\|(\mathrm{SB})^{-1}\mathrm{S}\|\hat{\beta}_{\mathrm{i}}(\mathrm{t}) + \|\sigma_i\|\varepsilon_{\mathrm{i}}\\ &\qquad + \|(S_iB_i)^{-1}S\|\beta_{\mathrm{i}}\\ &\qquad + \|(S_iB_i)^{-1}S\|\big(\hat{\beta}(t) - \beta_i\big)\|\sigma_{\mathrm{i}}\|\}\\ &\le \sum_{i=1}^{N}(-(\eta_{\mathrm{i}} - \varepsilon_{\mathrm{i}})\|\sigma_i\|) < 0\end{aligned} \tag{24}$$

which implies that the overall system will reach the sliding surface in finite time [11] and be maintained on it.

***Remark 2:*** Following [15], with the adaptive mechanism, we do not need to know the upper bound of the interaction. It is worthwhile noting that for the purpose of reducing the "chattering" around the switching surface, a commonly used approach is to substitute $\frac{\sigma_i}{\|\sigma_i\|+\theta_{\mathrm{i}}}$ in the SMC for $\frac{\sigma_i}{\|\sigma_i\|}$, where $\theta_{\mathrm{i}}$ is a small positive constant (the so called boundary layer) [11]. However, if we use the boundary layer method here, $\hat{\beta}_{\mathrm{i}}(\mathrm{t})$ may keep growing in magnitude during sliding and the chattering will not be reduced since $\dot{\hat{\beta}}_{\mathrm{i}}(\mathrm{t}) \neq 0$. In this case, we should modify our controller (20) to:

$$u_i(x_i,t) = \begin{cases}-(S_iB_i)^{-1}A_ix_i(t) + \Psi_{\mathrm{i}}\frac{\sigma_i}{\|\sigma_i\|+\theta_{\mathrm{i}}}, & \|\sigma_i\| > \theta_{\mathrm{i}}\\ -(S_iB_i)^{-1}A_ix_i(t), & \|\sigma_i\| < \theta_{\mathrm{i}}\end{cases} \tag{25}$$

### *C. Feasibility of LMIs and improvement*

A lot of literature only describe the LMIs and do not prove the feasibility of the LMIs. It is always reasonable to prove the feasibility when raising an LMI problem. To prove the LMIs (4), we should introduce a preliminary Lemma:

***Lemma 1*** [13]: Given a symmetric matrix $G \in R^{n\times n}$ and two matrices $U \in R^{p\times n}$ and $V \in R^{n\times m}$ where $p,m < n$ . Consider the problem of finding some matrix *K* such that:

$$G + UKV^T + VK^TU^T < 0$$

Denote by $\tilde{U}$ and $\tilde{V}$ the orthogonal complements of $U$ and $V$. Then the above inequality is solvable for *K* if and only if

$$\tilde{U}^TG\tilde{U} < 0, \tilde{V}^TG\tilde{V} < 0$$

And the feasibility of the LMIs (4) is given by the following Lemma 2.

***Lemma 2:*** The optimization problem given by LMIs (4) is feasible if the pairs $(A_i, B_i)$ are controllable for all the subsystems.

***Proof:*** Because the LMIs (4) are based on the overall system equation (3), we first need to prove the controllability of the overall system (3) It then follows that the overall system is controllable as all of the overall system eigenvalues are changeable by the local inputs, i.e. there are no fixed modes in this decentralized system. [1] Assume there is no actuator disturbance or no interactions in the overall state equation (3). In this case, there exists a control law $u = Kx$ if and only if there exist an s.p.d matrix *X* such that:

$$XA^T + AX + \mathrm{X}K^TB + BKX < 0 \tag{26}$$

As a consequence of the system controllability, we can always find a gain matrix *K* and an s.p.d matrix *X* satisfying (25), Moreover, we can restrict (26) to:

$$XA^T + AX + \mathrm{I} + XK^TB + BKX < 0 \tag{27}$$

There also still exists a *K* and an s.p.d matrix *X* satisfying (27). Then by using *Lemma 1*, the following inequality is feasible:

$$\Gamma^T(XA^T + AX + \mathrm{I})\Gamma < 0 \tag{28}$$

Using the Schur complement in the LMIs (4):

$$\Gamma^T(XA^T + AX + \mathrm{I})\Gamma + \Gamma^T\textstyle\sum_{i=1}^{N}(\frac{1}{\gamma_i}XH_i^TH_iX)\,\Gamma < 0 \tag{29}$$

Therefore, from (28) and (29), the solution of the LMIs (4) are guaranteed by the existence of a set of large enough $\gamma_i$ □

The LMI optimization problem given by (4) does not pose any restriction on the size of the matrix *X*. Consequently, the results of these two optimization problems may yield inappropriate results. For example, very small $\gamma_i$ and *X* result in very large *S* values. In this case, we can restrict *X* by posing a further constraint on $\gamma_i$ as:

$$\gamma_i > \frac{1}{\bar{\alpha}_{\mathrm{i}}^2}, \bar{\alpha}_i > 0 \tag{30}$$

$\bar{\alpha}_i$ are given constants for $i = 1,2,\ldots,N$. Or equivalently:

$$X > \Lambda \tag{31}$$

where, $\Lambda = \mathrm{diag}(\kappa_1 I_{n_1}, \ldots, \kappa_N I_{n_N})$ , $\kappa_i > 0, i = 1,\ldots,N$ are given constants. The upper bound of the parameters could also be defined in the same way.

*Remark 3:* We should note that if the constants $\alpha_i, i = 1,\ldots,N$ are known, the optimization problem (4) becomes feasibility problem. By substituting $\gamma_i = 1/\alpha_i^2$ into (4), the solution of the LMIs is equivalent to finding a feasible solution *X*. However, the use of an optimization algorithm could still be appropriate to find a solution capable of dealing with potentially stronger interactions.

For the purpose of improving the performance, here we introduce a modification to the LMIs in (4). By adding a $\mu$-stability constraint, the system satisfies $lim_{t\to\infty} e^{\mu t}\|x(t)\| = 0$ for all solution trajectories *x*. Therefore, we can ensure a minimum positive decay rate $\mu$ >after arriving sliding surface. Moreover, the larger the rate is, the earlier the sliding surface could be reached. Hence, the Eq. (4) could be rewritten as:

Minimize $\gamma_i$, subject to $X > 0$,

$$\begin{bmatrix} \Gamma^T(X\bar{A}^T + \bar{A}X + I)\Gamma & \Gamma^T X H_1^T & \cdots & \Gamma^T X H_N^T \\ H_1 X\Gamma & -\gamma_1 I & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ H_N X\Gamma & 0 & \cdots & -\gamma_N I \end{bmatrix} < 0 \tag{32}$$

where, $\bar{A} = A + \mu I$. Other improvement of robust performance for the system (e.g. $H_\infty$, $H_2$, etc.) could also be constructed based on this LMI approach.

## IV. NUMERICAL EXAMPLE

In this Section, we illustrate the performance of the proposed decentralized controller with an example similar to the one used in [7]. This non-linear interconnected system model consists of three subsystems. The first subsystem is a second-order system and the others are third-order systems. Disturbance signals are also added to each subsystem.

The dynamic subsystems are given by:

$$\dot{x}_1 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x_1 + \begin{bmatrix} 0 \\ 1 \end{bmatrix} (u_1 + f_1(x_1,t)) + h_1(x,t)$$

$$\dot{x}_2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} x_2 + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} (u_2 + f_2(x_2,t)) + h_2(x,t)$$

$$\dot{x}_3 = \begin{bmatrix} -3 & 0 & 1 \\ 1 & 2 & 0 \\ 0 & 1 & -2 \end{bmatrix} x_2 + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} (u_3 + f_3(x_2,t)) + h_3(x,t)$$

where, $x_1 = [x_{11} \quad x_{12}]^T$ , $x_2 = [x_{21} \quad x_{22} \quad x_{23}]^T$ , $x_3 = [x_{31} \quad x_{32} \quad x_{33}]^T$

$$f_1(x_1,t) = 0.4\sin(\mathrm{x}_{11}) + 0.5\sin(10\mathrm{t}),$$

$$f_2(x_1,t) = 0.3\cos(\mathrm{x}_{22}) + 0.6\sin(5\mathrm{t}),$$

$$f_3(x_1,t) = 0.5\cos(\mathrm{x}_{33}) + 0.6\sin(7\mathrm{t})$$

$$h_1(x,t) = \alpha_1\, cos(x_{22})\, H_1 x,$$

$$h_2(x,t) = \alpha_2\, cos(x_{32})\, H_2 x,$$

$$h_3(x,t) = \alpha_3\, cos(x_{11})\, H_3 x$$

$$\alpha_1 = \alpha_2 = \alpha_3 = 0.1,$$

$$H_1 = \frac{1}{\sqrt{10}}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$H_2 = \frac{1}{\sqrt{15}}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$H_3 = \frac{1}{\sqrt{13}}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Choosing $\bar{\alpha}_1 = \bar{\alpha}_2 = 0.2$ and solving the optimization problem (4) and (29) results in

$$X_1 = \begin{bmatrix} 0.515 & -0.5 \\ -0.5 & 0.515 \end{bmatrix} X_2 = \begin{bmatrix} 1.454 & -0.515 & -0.722 \\ -0.515 & 0.738 & -0.514 \\ -0.722 & -0.514 & 1.451 \end{bmatrix}$$

$$X_3 = \begin{bmatrix} 0.221 & -0.112 & -0.013 \\ -0.112 & 0.184 & -0.138 \\ -0.013 & -0.138 & 0.228 \end{bmatrix}$$

with this solution, we have the sliding surface matrices for each of the subsystems:

$$S_1 = [32.94 \quad 33.94]$$
$$S_2 = [36.143 \quad 50.744 \quad 36.627]$$
$$S_3 = [15.772 \quad 28.942 \quad 18.459]$$

Using the control law (25) and choosing $\theta_i = 0.01$ and $\eta_i = 5$ for $i = 1,2$ . The initial conditions $x_1(0) = [0.5 \quad 0.5]^T$ , $x_2 = [0.5 \quad 0.5 \quad 0.5]^T$. $x_3 = [0.5 \quad 0.5 \quad 0.5]^T$

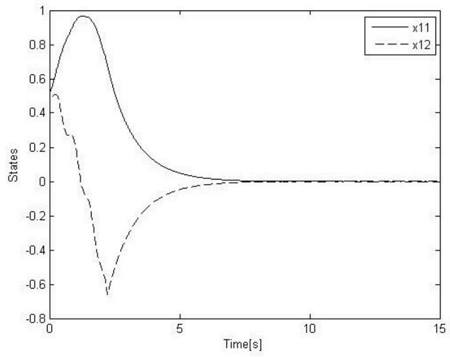


Figure 1. State responses for subsystem 1

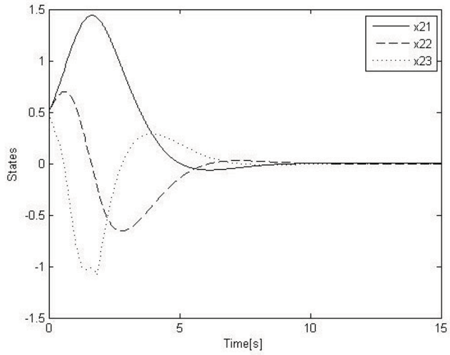


Figure 2. State responses for subsystem 2

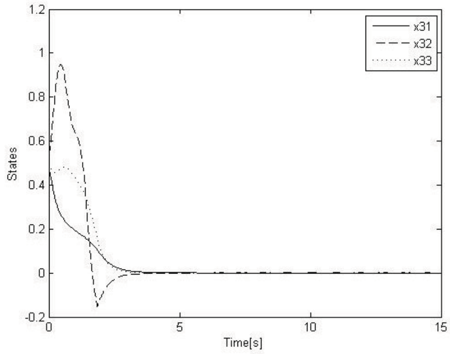


Figure 3. State responses for subsystem 3

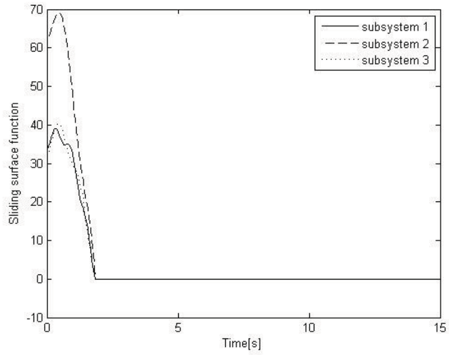


Figure 4. Sliding surface function for both subsystems

Figures 1-2 show the state responses for the system. From Figure 4, the sliding surface is reached in finite time. Moreover, combine with Figure 3, we note that after reaching the sliding surface, the system is stable and insensitive to the disturbances. However, in the reaching phase, the system is sensitive to the disturbances. This is also one of the main disadvantages of sliding mode theory.

## V. CONCLUSION

This paper proposes a simple and easy way to design a decentralized sliding mode system. The difference between the proposed approach and other decentralized sliding mode methods is that, this method requires the solution of only one LMI optimization problem in the sliding surface design, without a requirement for any unmatched and matched separation analysis. The mismatched uncertainties and interactions are handled well by the LMI approach. Meanwhile, all of the matched uncertainties and external disturbances are rejected by the SMC. The upper bound limitation of the uncertainties/interaction in the classical sliding mode design is removed by the adaptive mechanism. Other robust control methods could be easily extended based on this LMI approach.

## REFERENCES

[1] D. D. Šiljak, “Decentralized control of complex systems,” Academic Press, 1991.

[2] D. D. Šiljak, “Decentralized control and computations: status and prospects," *Annual Reviews in Control*, 1996, Vol. 20 pp.131-141.

[3] D. D. Šiljak, D. M. Stipanovic, and A. I. Zecevic, "Robust decentralized turbine/governor control using linear matrix inequalities," *IEEE Transactions on Power Systems*, 2002, Vol. 17(3), pp. 715-722.

[4] Y. Xie, W. Gui, and Z. Jiang, "Decentralized robust H-infinity descriptor output feedback control for value-bounded uncertain descriptor large-scale systems." *IET Journal of Control Theory and Applications*, 2006, Vol.4, pp. 193-200.

[5] K. Kalsi, J. Lian, and S. H. Zak, "On decentralized control of nonlinear interconnected systems," *International Journal of Conrtol*, March 2009, Vol. 82, pp. 541-554.

[6] K. Kalsi, J. Lian, and S. H. Zak, "Decentralized dynamic output feedback control of nonlinear interconnected systems," *IEEE Transactions on Automatic Control*, 2010, Vol. 55, pp. 1964-1970.

[7] Y. Zhu, and P. R. Pagilla. "Decentralized output feedback control of a class of large-scale interconnected systems," *IMA Journal of Mathematical Control and Information*, 2007, Vol. 24, pp. 57-69.

[8] W. J. Liu, "Decentralized control for large-scale systems with time-varying delay and unmatched uncertainties." *KYBERNETIKA*, 2011, Vol.47, pp. 285-299.

[9] F. Castanos and L. Fridman, “Integral sliding surface design using an h-infinity criterion for decentralized control,” *in 16th IFAC World Congress*, (Prague), 2005. paper Th-A09-TO/2

[10] N. Sandell, Jr., P. Varaiya, M. Athans and M. Safonov, "Survey of decentralized control methods for large scale systems," *IEEE Transactions on Automatic Control*, 1978, Vol. 23, pp.108-128.

[11] C. Edwards, S. K. Spurgeon, “Sliding mode control: theory and applications,” Taylor & Francis, 1998.

[12] H. H. Choi "A new method for variable structure control system design: A linear matrix inequality approach." *Automatica*, 1997, Vol.. 33, pp. 2089-2092.

[13] H. H. Choi, "An Explicit Formula of Linear Sliding Surfaces for a Class of Uncertain Dynamic Systems with Mismatched Uncertainties," *Automatica*, 1998, Vol.34, pp. 1015-1020.

[14] X. G. Yan, S. K. Spurgeon, et al. "Decentralized Output Feedback Sliding Mode Control of Nonlinear Large-Scale Systems with Uncertainties." *Journal of optimization theory and applications* Vol. 119 pp. 597-614.

[15] X. G. Yan, C. Edwards, S. K. Spurgeon, "Decentralised robust sliding mode control for a class of nonlinear interconnected systems by static output feedback." *Automatica*, 2004, Vol. 40, pp. 613-620.

[16] M. L. Hung and J. J. Yang, "Decentralized model-reference adaptive control for a class of uncertain large-scale time-varying delayed systems with series nonlinearities." *Chaos, Solitons and Fractals*, 2007, Vol. 33, pp. 1558-1568.

[17] S. Boyd, L. El Ghaoui, E. Ferron and V. Balakrishnan, "Linear matrix inequalities in systems and control theory", *Studies in Applied Mathematics, SIAM*, 1994, Philadelphia.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# A novel decoupling control method for multivariable systems with disturbances

Rui-Juan Liu[1,2,3] Guo-Ping Liu[2] Min Wu[1,3]

1. School of Information Science and Engineering, Central South University, Changsha 410083, China
2. Faculty of Advanced Technology, University of Glamorgan, Pontypridd CF37 1DL, U.K
3. Hunan Engineering Laboratory for Advanced Control and Intelligent Automation, Changsha 410083, China
Email: liuruijuan0313@163.com, gpliu@glam.ac.uk, min@csu.edu.cn

***Abstract*—This paper presents a novel decoupling method for multivariable systems with disturbances. In this method, the undesirable coupling parts in each loop are treated as the output disturbances. These disturbances, as well as the external disturbances, can be actively rejected by the equivalent-input-disturbance (EID) approach. The parameters of the controller in each loop can be designed independent of each other. A typical example demonstrates the simplicity in parameters design and good performance in decoupling control and disturbance rejection.**

***Index Terms*—Decoupling control, linear systems, disturbance rejection, state observer, equivalent-input-disturbance, pole assignment**

## I. Introduction

Multiple-inputs and multiple-outputs (MIMO) processes are often encountered in industry, such as chemical reactor, continuous stirred tank reactor, *et. al*. Because of the interaction between each control loop, the controller design for MIMO system becomes a complex problem and the common single-input single-output (SISO) control theories are not effective for MIMO systems. However, the practical control engineers are always accustomed to deal with the MIMO system loop by loop, which requires the common SISO control methods, such as PID control, optimal control. To solve such a problem, one of the control strategies is trying to eliminate the interactions between control loops, which is known as decoupling control.

The decoupling control theory has been well developed and established for several decades. To measure the interaction of the MIMO systems, Bristol [1] introduced relative gain array method for the input-output pairings and controller design. For linear systems, an effective approach to the internal model design [2] is proposed for decoupling stable square multivariable process with delays. A transfer function matrix decoupling approach is presented for the MIMO Smith Scheme [3]. Although many significant results have been proposed, the decoupling performance is still desired to be improved and some key problems, such as robustness, disturbance rejection, and other practical concerns [4] continue to pose serious challenges.

It is well known that the disturbance is of a primary concern for control system design, and is even the ultimate objective. This problem, in conjunction with decoupling control, has attracted many researchers. Feedback control [5] may be effective for reducing the disturbance and making disturbance rejection tuning independent of the controller design, if the disturbance is measurable. Due to the difficulties in obtaining exact model or disturbance, decoupling control schemes are often constructed to estimate the cross-couplings and disturbance at the same time; and a number of observer-based methods are making progress. Such as the disturbance observer based method (DOB) [6], the perturbation observer approach (POB) [7], disturbance decoupling control [8] based on the active disturbance rejection control (ADRC) [9] and so on. However, DOB requires the inverse dynamics model and POB needs the accurate model of the plant. The ADRC based decoupling method overcomes some drawbacks of the existing methods. However, the parameters in the resulting controller are usually very complicated.

In an MIMO system, the output of one loop is always the sum of all inputs actions. The actions of other loops are the resource of the interactions and bring bad effect on the stability and performance, which is just like output disturbances occurring in this loop. Intuitionally, we may think that if such ”output disturbances” can be fully overcame or compensated, the output of one loop is only determined by the relevant input, and then the MIMO system can be well decoupled. One of the novelties of this paper is that, we treat the undesirable coupling parts of each loop as its output disturbances, and compensate them together with the external disturbance. The equivalent-input-disturbance (EID) approach [10], [11] is a relatively new disturbance rejection method. It can reject more than one disturbance simultaneously and compensate any kind of disturbance effectively without knowing prior information. These important features motivated us to design the decoupling controller using the EID-based method.

This paper first constructs a configuration of the decoupling control based on the EID approach. Then, the controller parameters for each loop are designed independently. Finally, simulation results are shown for validating the proposed method.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

## II. Configuration of the dynamic decoupling control

Consider a linear time-invariant MIMO system. Let $r_i$ and $u_i$, $i = 1, 2, \cdots, n$ denote the reference inputs and the control inputs, respectively. Suppose that the open loop transfer function matrix of the process is described by

$$G(s) = [g_{ij}(s)]_{n\times n} \tag{1}$$

Then for the *i*-th loop, the output is given by

$$y_i = \sum_{j=1}^{n} g_{ij}(s)u_j, \quad i = 1, 2, \cdots, n \tag{2}$$

We divide it into two parts:

$$y_i = g_{ii}(s)u_i + \sum_{j=1, j\neq i}^{n} g_{ij}(s)u_j \tag{3}$$

where $g_{ii}(s)$ is regarded as the transfer function of the plant in the *i*-th loop, the second term in the right hand side is treated as a disturbance signal. So that it becomes an SISO system with disturbances for each loop.

For clarity, we take a two-input two-output (TITO) system as an example, as shown in Fig.1. For the plant $g_{11}$, we only need to design a controller $c_1$ to reject the signal through $g_{12}$, as well as the disturbance $d_1$. Similarly, one designs controller $c_2$. Due to this process, we can also notice that only the diagonal elements of the transfer function are needed in the compensation instead of the whole information.

For convenience, the processes are formulated in state-space forms. Then the plant in the *i*-th loop is given by

$$\begin{cases} \dot{x}_i(t) = A_i x_i(t) + B_i u_i(t) + D_i d_i(t) \\ y_i(t) = C_i x_i(t) + \Phi_i \varphi_i(t) \end{cases} \tag{4}$$

where $A_i$, $B_i$ and $C_i$ are matrices obtained by the transfer function $g_{ii}(s)$, $i = 1, 2, \cdots, n$, $\Phi_i\varphi_i(t)$ denotes the signal $\sum\limits_{j=1, j\neq i}^{n} g_{ij}(s)u_j$, which need not to be known exactly, and $D_i d_i(t)$ is the external disturbance.

Now, a transformation is required. For the output $y_i(t)$, there must exist a matrix $\Psi_i$ such that

$$y_i(t) = C_i[x_i(t) + \Psi_i\varphi_i(t)] \tag{5}$$

Let

$$\tilde{x}_i(t) = x_i(t) + \Psi_i\varphi_i(t) \tag{6}$$

Combing (5) and (6), the state equation (4) is derived as

$$\begin{cases} \dot{\tilde{x}}_i(t) = A_i\tilde{x}_i(t) + B_i u_i(t) + [D_i d_i(t) - A_i\Psi_i\varphi_i(t)] \\ y_i(t) = C_i\tilde{x}_i(t) \end{cases} \tag{7}$$

where $D_i d_i(t) - A_i\Psi_i\varphi_i(t)$ is regarded as whole disturbances imposed on this loop.

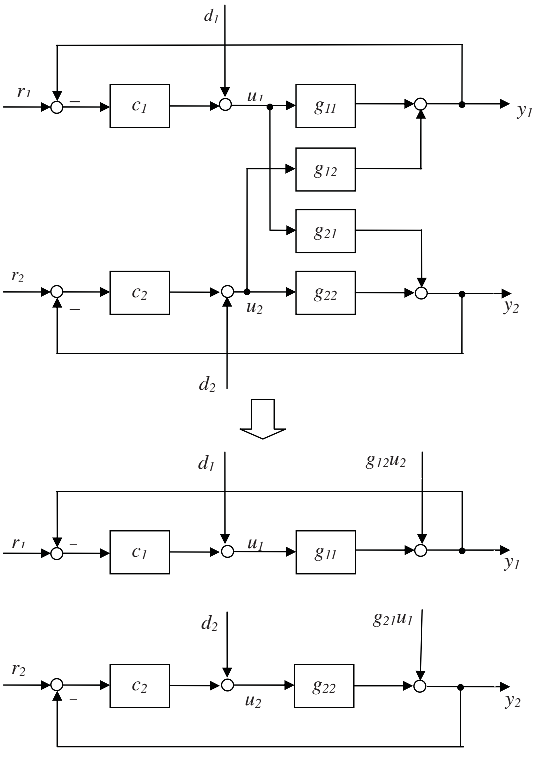


Fig. 1. Configuration of TITO decoupling control system

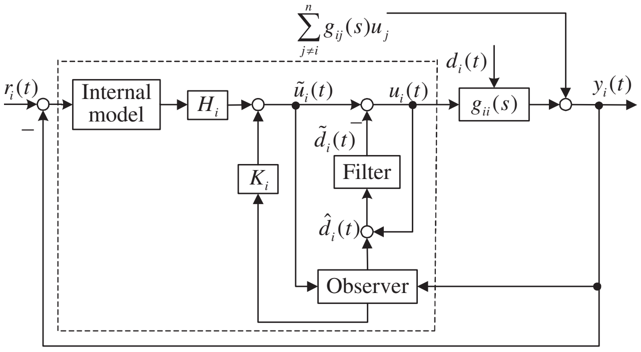


Fig. 2. Configuration of the EID method

An EID of the plant is defined to be a signal on the control input channel that produces the same effect on the output as disturbances do for all $t \geq 0$. The EID of the disturbance must exist under the condition that the plant does not have any zeros on the imaginary axis [10]. Suppose that a disturbance $de_i(t)$ is imposed on the control input channel of the *i*-th plant, then (7) can be described as the following equivalent model

$$\begin{cases} \dot{\tilde{x}}_i(t) = A_i\tilde{x}_i(t) + B_i[u_i(t) + de_i(t)] \\ y_i(t) = C_i\tilde{x}_i(t) \end{cases} \tag{8}$$

Note that the same variable $\tilde{x}_i$ is used for both the state of plant (7) and (8). This should not make confusion.

The configuration of the EID-based control system is shown in Fig. 2. The controller $c_i$ in Fig. 1 is just designed using the parts within the dashed line in Fig. 2. In this figure, the internal model

$$\dot{\bar{x}}_i(t) = \bar{A}_i\bar{x}_i(t) + \bar{B}_i[r_i(t) - y_i(t)] \tag{9}$$

is employed to improve the tracking precision, in which the parameters are determined by the reference input $r_i(t)$. The Luenberger observer

$$\left\{ \begin{array}{l} \dot{\hat{x}}_i(t) = A_i\hat{x}_i(t) + B_i\tilde{u}_i(t) + L_i(y_i(t) - \hat{y}_i(t)) \\ \hat{y}_i(t) = C_i\hat{x}_i(t) \end{array} \right. \tag{10}$$

plays a key role in the disturbance estimation.

Let

$$\Delta x_i(t) = \tilde{x}_i(t) - \hat{x}_i(t). \tag{11}$$

[10] gives us the expression of the estimation value

$$\hat{d}_i(t) = B_i^{+} L_i C_i \Delta x_i(t) + u_i(t) - \tilde{u}_i(t) \tag{12}$$

where $B_i^{+} = (B_i^T B_i)^{-1} B_i^T$.

Then, a low-pass filter is used to select the frequency band of the signal, *e.i.*,

$$\mathscr{L}[\tilde{d}_i(t)] = F(s)\mathscr{L}[\hat{d}_i(t)] \tag{13}$$

where $\mathscr{L}(\cdot)$ is the Laplace transform and $F_i(s)$ satisfies

$$|F_i(j\omega)| \approx 1 \tag{14}$$

for all $\omega \in \Omega$, where $\Omega$ is the chosen frequency band. Note that each loop becomes an SISO system in the design, a first-order filter

$$F_i(s) = 1/(T_i s + 1) \tag{15}$$

can work well.

Thus, $\tilde{d}_i(t)$ is just the estimated compensation of the whole disturbances in the *i*-th loop. So that the control law is given by

$$u_i(t) = \tilde{u}_i(t) - \tilde{d}_i(t) \tag{16}$$

where the state feedback control law is

$$\tilde{u}_i(t) = H_i\bar{x}_i(t) + K_i\hat{x}_i(t) \tag{17}$$

## III. Stability analysis and parameters design

Since the plant we considered in each loop may be influenced by any of the reference input, we set

$$r_i(t) = 0, i = 1, 2, \cdots, n, \ d_i(t) = 0 \tag{18}$$

for the *i*-th loop to guarantee that all the external inputs be zero.

Then combining (4), (10), (11) and (18) yields

$$\Delta\dot{x}_i(t) = -(A_i - L_iC_i)\Delta x_i(t) + B_i\tilde{d}_i(t) \tag{19}$$

and

$$\hat{d}_i(t) = B_i^{+} L_i C_i \Delta x_i(t) + \tilde{d}_i(t) \tag{20}$$

So the transfer function from $\tilde{d}_i(t)$ to $\hat{d}_i(t)$ can be derived as

$$G_i(s) = B_i^{+}(sI - A_i)[sI - (A_i - L_iC_i)]^{-1}B_i \tag{21}$$

In the EID-based control system, the state feedback design does not influence the stability of the whole system. Therefore, we have the following theorem.

*Theorem 1:* [10] For suitably designed $H_i$ and $K_i$, the control law (16) guarantees the stability of the *i*-th loop, if $A_i - L_iC_i$ is stable and

$$\|G_i(s)F_i(s)\|_\infty < 1 \tag{22}$$

where $||\cdot||_\infty$ denotes the upper bound of the maximum singular value of the function.

Since the plant in each loop is controllable and observable, the state feedback gains and the observer can be designed independently [14]. This brings us great convenience for parameters design.

As for the state feedback gains, the following augmented system including the original plant and the internal model is considered

$$\delta\dot{x}_i(t) = \tilde{A}_i\delta x_i(t) + \tilde{B}_iu_i(t) \tag{23}$$

where

$$\delta x_i(t) = \left[ \begin{array}{c} \tilde{x}_i(t) \\ \bar{x}_i(t) \end{array} \right], \ \tilde{A}_i = \left[ \begin{array}{cc} A_i & 0 \\ -\bar{B}_iC_i & \bar{A}_i \end{array} \right], \ \tilde{B}_i = \left[ \begin{array}{c} B_i \\ 0 \end{array} \right] \tag{24}$$

$H_i$ and $K_i$ can be optimized by using a well-known linear quadratic regulation (LQR) method [15]. The optimal state-feedback control law is given by

$$u_i^*(t) = -R_i^{-1}\tilde{B}_i^TP_i\delta x_i(t) \tag{25}$$

where $P_i = \left[ \begin{array}{cc} K_i & H_i \end{array} \right]$ is a solution of the Riccati equation

$$P_i\tilde{A}_i + \tilde{A}_i^TP_i + Q_i - P_i\tilde{B}_iR_i^{-1}\tilde{B}_i^TP_i = 0 \tag{26}$$

$Q_i > 0$ and $R_i > 0$ are diagonal matrices to be determined by the coefficient matrices of the augmented system.

The Luenberger observer has an important function that enables the coupling parts and disturbances to be compensated for. It will achieve good performance and be easily operated for a well designed gain $L_i$. We use the well-known pole-placement theory in this study instead of the perfect regulation method in [10], no matter if the plant is stable. First, we need to select a time parameter $T_i$ that satisfies (14) for the low-pass filter. Then $L_i$ can be obtained by the following procedure.

*Observer gain design algorithm*

Step 1: [14] For a prescribed $\zeta$, $0 \le \zeta \le 1$, the expected poles for the observer is chosen to be

$$\begin{aligned} \lambda_{1,2} &= -\zeta\sigma \pm j\sqrt{1-\zeta^2}\sigma, \\ \lambda_k &= -a_k\zeta\sigma, \ a_k \ge 5, \ k = 3, \cdots, m \end{aligned} \tag{27}$$

Step 2: Choose $\sigma$ in (27) and calculate the transfer function $G_i(s)$ in (21), such that (22) holds.

Step 3: Calculate the observer gain $L_i$ such that

$$\lambda_k(A_i - L_iC_i) = \lambda_k, \ k = 1, \cdots, m \tag{28}$$

*Remark 1:* Only if the plant is observable, the expected poles of the observer can be arbitrarily placed by the feedback matrix $L_i$. It is well known that the performance of the observer is determined by the dominant poles $\lambda_{1,2}$. In addition, the following part proves that there must exist a large enough $\sigma$ such that (22) holds.

It follows from (21) that

$$1/\left\|G_i(s)\right\|_\infty = \left\|\tilde{G}_i(s)\right\|_\infty \tag{29}$$

where

$$\tilde{G}_i(s) = B_i^{+}[sI - (A_i - L_iC_i)](sI - A_i)^{-1}B_i$$

Since $\lim\limits_{\sigma\to\infty} |\lambda_i| = \infty$ $(i = 1, \cdots, m)$, from (28), we have $\lim\limits_{\sigma\to\infty} \left\|A_i - L_iC_i\right\|_\infty = \infty$. For any fixed $\omega$, the matrices $j\omega I$ and $(j\omega I - A_i)^{-1}$ are constant matrices with finite norms, respectively. So,

$$\lim_{\sigma\to\infty} \left\|\tilde{G}_i\right\|_\infty = \infty. \tag{30}$$

Then, for a large enough $\sigma$, (22) holds. In view of the modeling uncertainties, $\sigma$ can be chosen a little large to handle the uncertainties and guarantee the stability of the system.

*Remark 2:* In the design of the proposed method, the coupling parts in each loop are treated as disturbances. In this case, even if strong multivariable interactions occur, the stability of the system will not be destroyed.

The controllers are designed loop by loop based on the analysis above. Although every loop needs a controller as discussed above, the parameters can be designed similarly among all loops and be independent of each other.

## IV. Numerical example

The Wood-Berry model of a pilot-scale distillation column [16] has been studied extensively. This section considers this multivariable system with delay set to zero, which is shown as

$$\begin{bmatrix} y_1(s) \\ y_2(s) \end{bmatrix} = \begin{bmatrix} \frac{12.8}{16.7s+1} & \frac{-18.9}{21s+1} \\ \frac{6.6}{10.9s+1} & \frac{-19.4}{14.4s+1} \end{bmatrix} \begin{bmatrix} u_1(s) \\ u_2(s) \end{bmatrix} \tag{31}$$

Let the reference inputs be $r_1 = 1$ and $r_2 = 1$, which were imposed at $t = 0$ s and $t = 20$ s, respectively. Then a disturbance $d_1(t) = 0.4$ was imposed on the first plant at $t = 40$ s.

The proposed method is applied to this model. Since the reference inputs are step signals, the parameters of the internal models are chosen to be

$$\bar{A}_1 = \bar{A}_2 = 0, \ \bar{B}_1 = \bar{B}_2 = 1 \tag{32}$$

As for the time constant $T_1$ and $T_2$, 0.01 will be suitable for both of the filters.

For the first loop, let

$$Q_1 = diag\left\{ \begin{array}{cc} 1 & 1 \end{array} \right\}, \ R_1 = 1 \tag{33}$$

Using LQR method yields

$$K_1 = -1.0, \ H_1 = 1.8233 \tag{34}$$

Set $\zeta = 1$ in (27) and choose $\sigma = 20$ in Step 2 of the observer gain design algorithm. So that they satisfy (22). Then

$$\lambda_1 = -20 \tag{35}$$

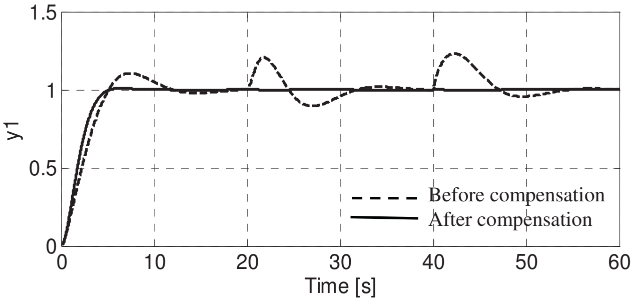


Fig. 3. The output responses for loop 1

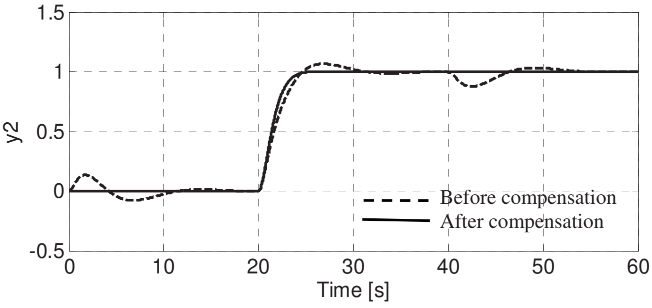


Fig. 4. The output responses for loop 2

Calculating the observer gain according to (28), yields

$$L_1 = 19.9401. \tag{36}$$

Similarly, for the second loop, choosing

$$Q_2 = diag\left\{ \begin{array}{cc} 1 & 1 \end{array} \right\}, \ R_2 = 1 \tag{37}$$

and

$$\lambda_2 = -20 \tag{38}$$

yields

$$K_2 = 1.0, \ H_2 = -1.5255 \tag{39}$$

and

$$L_2 = 19.9306. \tag{40}$$

The simulation results are shown in Fig.3 and Fig.4, respectively. In Fig.3, after the second step input and the disturbance being imposed, the peak to peak value (PPV) of the response in loop 1 is still less than 0.01. It occurs in loop 2 similarly; even the PPV at the beginning is only less than 0.01. It can be seen clearly that the proposed method achieves satisfactory performance in both decoupling control and disturbance rejection.

## V. Conclusion

A dynamic decoupling control method has been presented for MIMO systems with disturbances. The undesirable coupling parts and external disturbances in each loop are treated as ”disturbances”, respectively. So that these disturbances can be effectively compensated by using the EID-based approach. The decoupling control and disturbance rejection can be carried out simultaneity without knowing the prior information of all the treated disturbances. Although every loop needs a controller, the parameters design is still very simple and can be independent of each other. Simulation results demonstrated the good performance of the proposed method.

## ACKNOWLEDGMENT

The authors are grateful for the support of the National Natural Science Foundation of China (Grant No. 60974045). The authors would also like to thank the anonymous reviewers for their suggestions that help improve the manuscript.

## REFERENCES

[1] E. H. Bristol. "On a new method measure of interaction for multivariable process control". IEEE Transactions on Automatic Control, AC-133-4, 1966.

[2] Q. G. Wang,, T. H. Lee and J. B. He. "Internal stability of interconnected systems". IEEE Trans. Automatic Control. vol. 44, pp. 593-597, 1999.

[3] Q. G. Wang, B. Zou and Y. Zhang. "Decoupling smith predictor design for multivariable system with multiple time delay". Chemical Engineering Research and Design. vol. 78, pp. 565-572, 2000.

[4] Q. G. Wang. Decoupling control. Berlin: Springer. 2003.

[5] G. C. Goodwin, S. F. Graebe and M. E. Salgado, Control system design. Prentice Hall. 2011.

[6] K. J. Yang, Y. J. Choi and W. K. Chung, "On the tracking performance improvement of optical disk drive servo systems using error-based disturbance observer," IEEE Transactions on Industry Electronics, vol. 52, no. 1 pp. 270-278, 2005.

[7] S. Kwon and W. K. Chung, "A discrete-time design and analysis of perturbation observer for motion control applications". IEEE Transactions on Control Systems Technology, vol. 11, no.3, pp. 399-407, 2003.

[8] Q. Zheng, Z. Z. Chen, Z and Q. Gao. "A practical approach to disturbance decoupling control. Control engineering practice", vol. 17, pp. 1016-1025, 2009.

[9] J. Q. Han, "From PID to active disturbance rejection control," IEEE Transactions on Industrial Electronics, vol. 56, no. 3, pp. 900-906, 2009.

[10] J. She, M. X. Fang, Y. Ohyama, H. Kobayashi and M. Wu, "Improving disturbance-rejection performance based on an equivalent-input-disturbance approach," IEEE Transactions on Industry Electronics, vol. 55 no. 1, pp. 380-389, 2008.

[11] J. She, X. Xin and Y. Pan, "Equivalent-input-disturbance approach–Analysis and application to disturbance rejection in dual-stage feed drive control system," IEEE/ASME Transactions on Mechatronics, vol.16, no, 2, pp. 330-340, 2011.

[12] L. R. Hunt, G.Meyer, and R. Su, "Noncausal inverses for linear systems," IEEE Trans. Automatic Control, vol. 41, no. 4, pp. 608-611, Apr. 1996.

[13] M. Wu, W. Gui and Y. He, Advanced robust control (third edition). Science Press, Beijing, 2010.

[14] D. Z. Zheng. Linear system theory (second edition). TUP, Springer, 2002.

[15] B. D. O. Anderson and J. B. Moore, "Optimal Control-Linear Quadratic Methods. Englewood Cliffs", NJ: Prentice-Hall, 1989.

[16] R. K. Wood and M. W. Berry, "Terminal composition control of abinary distillation column". Chemical Engineering Science, vol. 28, pp. 1707C1717, 1973.

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# *Robust decentralized control design using integral sliding mode control*

Eshag Larbah[1] and Ron J Patton
Department of Engineering, University of Hull, Hull HU6 7RX, UK
(e-mail: r.j.patton@hull.ac.uk; e.y.larbah@2008.hull.ac.uk)

*Abstract*— **The problem of robust decentralization of uncertain inter-connected systems is concerned with the goal of de-coupling a Lipschitz non-linear systems into individual "decentralized" subsystems satisfying security and fault-tolerance objectives. This work proposes a new strategy for robust decentralized control in which each subsystem uses an observer-based state estimate structure invoking an approach to *separation principle recovery*, based on Integral Sliding Model Control (ISMC) with careful consideration of both matched and unmatched uncertainties arising from inter-connections and disturbances. The proposed design strategy for the linear observer and uncertainty de-coupling designs involves a single LMI. An example of 3 unstable inter-connected non-linear systems is used to illustrate the power of the approach.**

Keywords Decentralized control, Integral sliding mode control

## I. Introduction

The study of control of non-linear inter-connected systems has received considerable interest in recent years [1], [2], [3]. Some research focuses on interconnected systems with uncertainties, e.g. unknown nonlinear interconnections and disturbances, presenting robustness design challenges involving control specifications for each subsystem. These systems are particularly difficult to design when faced with limitations arising from uncertainty matching conditions and lack of available state information [4].

In most cases the design of robust decentralized systems focuses on state feedback problems. However, in reality only output information is available and this adds a further challenge to the robust design problem. It is often the case that the controller designs must depend to a degree on estimated states, and hence it is common practice in the literature to investigate the observer based feedback control approach with state estimates based on local information [5], [6]. In fact, the derivation of robust output feedback for decentralized control systems with uncertain inter-connection remains a difficult challenge in the literature [7].

Observer-based strategies represent a commonly used way of dealing with output feedback design and there are two observer-based control paradigms for decentralized systems. Firstly, a separate "decentralized" observer is designed for each subsystem, taking account of local information. The second approach involves the use of "inter-connected observers" [8] in which each observer measurement and input information is shared with observers from other local subsystems.

In many branches of control systems there is a need to compensate robustly for effects of system uncertainties or effects of input disturbances or even faults, to maintain required closed-loop performance and stability. One such approach is the use of sliding mode control (SMC) in which the system dynamic behavior can be forced to be independent of inputs, and certain disturbances and modeling uncertainties, once the so-called sliding regime has been reached. Several studies of inter-connected decentralized systems have focused on the use of SMC as a basis for solving robustness [9]. However, the classical approach to SMC requires (i) a reachability condition to guarantee that the SMC sliding or switching motion in state space can be reached from arbitrary initial conditions, and (ii) that two separate controllers be designed to achieve reachability and satisfy the sliding mode design objectives [10]. In the case of ISMC the requirements for both (i) & (ii) above are obviated, since the sliding motion is reached from initial time, making the use of ISMC very attractive for robust control of decentralized systems [11].

This paper focuses on the use of ISMC for decentralized control, based on state estimate feedback. It is assumed that the local system states are not measurable and hence the *decentralized observer* approach outlined above is used as a part of a state-estimate feedback design problem. Decentralized observers are used as a part of the strategy to de-couple the effects of inter-connections between subsystems. Although, each observer has linear feedback structure the observer-based control is formulated using a single LMI procedure to satisfy both *Lyapunov stability* and performance of the augmented state space form of the observer-controller state space system. This relates to the classical Separation Principle only in the sense that objective for each subsystem is to provide effective recovery of the Separation Principle and hence also effective decentralization. This is achieved through the use of the single LMI approach involving the feedback designs for each observer and controller [12]. The system description involves both matched and unmatched uncertainty components (arising from inter-connections and disturbance) and the paper deals with both forms of uncertainty.

The paper is structured as follows. Section II describes the problem formulation. Then section III considered the proposed control approach that includes output integral sliding mode

[1] *Eshag Larbah acknowledges funding support for his PhD scholarship from the ministry of education of Libya.*

control (OISLMC) in the first part and LMI observer–based control design in the second part. Section IV describes a numerical example with three interconnected systems to illustrate the design approach and simulation performance. Section V gives some conclusions.

## II. PROBLEM FORMULATION

Consider an interconnected system comprising subsystems described by:

$$\dot{x}_i(t) = A_i x_i(t) + B_i u_i(t) + Z_i(x_i,t) + W_i(x_i,t) + E_i d_i(t) + B_i f_i(t)$$
$$y_i(t) = c_i x_i(t) \qquad i = 1,2,□,n \qquad (1)$$

where $x_i(t) \in \mathbb{R}^n$is the state vector, $u_i(t) \in \mathbb{R}^m$ are the control inputs and $y_i(t) \in \mathbb{R}^p$ is the vector of system outputs. $A_i, B_i, C_i$ and $E_\mathrm{i}$ are known matrices of appropriate dimension. $Z_i(x_i,t) \in \mathbb{R}^n$ represent the unknown time-varying interactions between the subsystems, containing matched and unmatched components. Hence, $Z_i = Z_{mi} + Z_{ui}$ where $Z_{mi}$ is a matched component of $Z_i$and $Z_{ui}$ is the unmatched components [13].

Dropping the subscripts in $Z_i(x_i,t)$ and using the identity $I_n = BB^+ + B^\perp B^{\perp +}$ where $B^+ = (B^T B)^{-1} B^T$, $Z_i = B_i B_i{}^+ Z_i + B_i{}^\perp B_i{}^{\perp +} Z_i$ and $B^T B^\perp = 0$, then $Z_i = B_i B_i{}^+ Z_i + \zeta_i$ where $\zeta_i = B_i{}^\perp B_i{}^{\perp +} Z_i$ contains the unmatched uncertainty components.

$W_i(x_i,t)$ represents the subsystem unknown modeling uncertainties that satisfies the matching condition $W_i(x_i,t) = B_i Q_i(x_i,t)$ are unknown, $d_i(t)$ is an unknown bounded process disturbance, $f_i(t) \in \mathbb{R}^k$ denotes the actuator faults, where $f_i(t) = -K(t)u_i$ and $K(t) = diag(K_i)$ with $0 \leq K_i \leq 1$, $K_i = 0$ means that the actuator is working perfectly and if $K_i = 1$ the actuator has failed completely otherwise the fault is present.

***Assumptions*:**

- **A1**-The pair $(A_i, B_i)$ is controllable and $(C_i, A_i)$ is an observable pair.
- **A2**-$B_i$ has full rank $m_i$.
- **A3**-The initial state $x_i(t_o)$ is bounded.
- **A4**- The $Z_i(x_i,t)$ are Euclidean bounded norms as: $\|Z_i(x_i,t)\| \leq \beta_i(x_i,t)$ where $\beta_i(x_i,t)$ is a known nonlinear function [14].
- **A5**- The $Q_i(x_i,t)$ are bounded as: $\|Q_i(x_i)\| \leq \kappa_i \|x_i\|$ where $\kappa_\mathrm{i} > 0$ are known Lipschitz constants [15].
- **A6**- The $d_i(t)$ are Euclidean bounded norms as: $\|d_i(t)\| \leq \gamma_i \|x_i\|$ where $\gamma_i > 0$ are known constants.
- **A7**- $f_i(t)$ are Euclidean bounded norms as: $\|f_i(t)\| \leq \eta_i \|x_i\|$ where $\eta_i > 0$ are known Lipschitz constants.

All these assumptions are applicable for real control system problems, since all designs are made off-line.

Following the assumptions above Eq. (1) becomes:

$$\dot{x}_i(t) = A_i x_i(t) + B_i u_i(t) + B_i B_i^+ Z_i(t) + \zeta_i(t) + B_i Q_i(x_i,t) + E_i d_i(t) + B_i f_i(t)$$
$$y_i(t) = c_i x_i(t) \qquad i = 1,2,\ldots\ldots,n \qquad (2)$$

The control signal contains *two* components:

$$u_i(t) = u_i^{OBC}(t) + u_i^{ISM}(t) \qquad (3)$$

where $u_i^{OBC}$ is responsible for stabilizing the system and affects the desired performance and decreases the affect of unmatched components where the state is not available. $u_i^{ISM}$ is a discontinuous control responsible for rejecting the effects of matched components (uncertainties and actuator faults).

## III. CONTROL DESIGN

As described in (3) the subsystem control signal includes two parts with each part designed using a different method where (i) $u_i^{ISM}(t)$ is designed by output integral sliding mode control OISMC where the state is not available and only the estimated state is obtainable, and (ii) $u_i^{OBC}(t)$ depends on the estimated state as shown in Fig. 1.

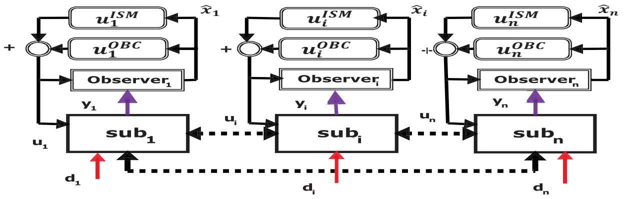


Figure 1 output control of interconnected systems via LMI+ISMC

### A. OUTPUT INTEGRAL SLIDING MODE CONTROL (OISLMC)

As outlined in Section 1 the integral sliding control can be used to remove the reachability problem. The output feedback case of integral sliding mode control can be developed by defining the following integral sliding switching surface:

$$\sigma_i(y_i, \hat{x}_i, t) =$$
$$G_i[y_i(t) - y_i(t_o) - \int_{t_o}^{t} (C_i A_i \hat{x}_i(t) + C_i B_i u_i^{LMI}(t)) dt] \qquad (4)$$

where $G_i \in \mathbb{R}^{mxp}$ is a design freedom matrix that must satisfy the invertibility of $G_i C_i B_i$ .

The two ISMC design steps are as:

1- Sliding surface design that satisfies the system performance and ensures that the system has required performance when the state trajectory is on the sliding surface.

2- Appropriate discontinuous control to maintain the system trajectory close to or on the sliding surface.

In the ISMC the design freedom of the integral action can be used to design a control law that satisfies the prescribed closed-loop performance.

The equivalent control $u_{eqi}(t)$ can be maintained on the sliding surface by forcing the time derivative of $\sigma_i(y_i, \hat{x}_i, t)$ in (4) to be zero-valued [16], i.e.:

$$\dot{\sigma}_i(y_i, \hat{x}_i, t) = G_i \dot{y}_i(t) - G_i C_i A_i \hat{x}_i(t) - G_i C_i B_i u_i^{OBC}(t) = 0 \qquad (5)$$

Then substituting (2) and (3) into (5) yields:

$$G_i C_i A_i x_i(t) + G_i C_i B_i u_i^{OBC} + G_i C_i B_i u_i^{ISM} + G_i C_i B_i B_i^+ Z_i(t) + G_i C_i \zeta_i(t) + G_i C_i B_i Q_i(x_i,t) + G_i C_i E_i d_i(t) + G_i C_i B_i f_i(t) - G_i C_i A_i \hat{x}_i(t) - G_i C_i B_i u_i^{OBC} = 0 \qquad (6)$$

Hence, the so-called *equivalent control* for the output feedback case is:

$$u_{eqi}(t) = u_i^{ISM} = -(G_iC_iB_i)^{-1}G_iC_iA_i(x_i(t) - \hat{x}_i(t)) - B_i^+Z_i(t) - (G_iC_iB_i)^{-1}G_iC_i\zeta_i(t) - Q_i(x_i,t) - (G_iC_iB_i)^{-1}G_iC_iE_id_i(t) - f_i(t) \quad (7)$$

Substituting (7) into (2) gives the $i^{\text{th}}$ subsystem state equation as:

$$\dot{x}_i(t) = A_ix_i(t) + B_iu_i^{OBC}(t) + [I_i - B_i(G_iC_iB_i)^{-1}G_iC_i]\zeta_i(t) + [I_i - B_i(G_iC_iB_i)^{-1}G_iC_i]E_id_i(t) - (G_iC_iB_i)^{-1}G_iC_iA_i(x_i(t) - \hat{x}_i(t) \quad (8)$$

From (8) the unknown matched uncertainties and actuator faults are completely nulled but the dynamics on the sliding surface contains the unknown unmatched uncertainties, disturbance and the state error. The unknown unmatched uncertainties and disturbances are multiplied by a matrix:

$$\Psi_i = [I_i - B_i(G_iC_iB_i)^{-1}G_iC_i]$$

To simplify the notation Eq. (10) can now be re-written as:

$$\dot{x}_i(t) = A_ix_i(t) + B_iu_i^{OBC}(t) + \Psi_i\zeta_i(t) + \Psi_iE_id_i(t) - M_ie_i(t) \quad (9)$$

where $M_i = (G_iC_iB_i)^{-1}G_iC_iA_i$ and $e_i(t) = x_i(t) - \hat{x}_i(t) \in \mathbb{R}^n$is the estimation error. The proposed discontinuous control is:

$$u_i^{ISM}(t) = -\mu_i \frac{\sigma_i(y_i,\hat{x}_i,t)}{\|\sigma_i(y_i,\hat{x}_i,t)\| + \mathfrak{z}_i} \quad (10)$$

The parameters $\mathfrak{z}_i > 0$ are chosen to reduce the amount of "chattering" of the motion around the sliding surface [15]. To satisfy subsystem stability the positive scalar $\mu_i$ is chosen according to the following derivation:

$$\mu_i > (G_iC_iB_i)^{-1}G_iC_i\beta_i(x_i,t) + \kappa_i\|x_i\| + (G_iC_iB_i)^{-1}G_iC_iE_i\gamma_i\,\|x_i\| + \eta_i\,\|x_i\| + (G_iC_iB_i)^{-1}G_iC_iA_i\|e_i(t)\| \quad (11)$$

To ensure sliding motion let $\sigma_i(y_i,\hat{x}_i,t) = 0$. Furthermore, the stability of the inter-connected system (1) is considered in terms of a positive definite summation of individual Lyapunov subsystems components as:

$$\textstyle\sum_{i=1}^n V_i\,(\sigma_i(y_i,\hat{x}_i,t)) = \sum_{i=1}^n\|\sigma_i(y_i,\hat{x}_i,t)\| > 0:$$

The derivative of the subsystem Lyapunov functions are:

$$\dot{V}_i(\sigma_i(y_i,\hat{x}_i,t)) = \frac{\sigma_i(y_i,\hat{x}_i,t)^T\dot{\sigma}_i(y_i,\hat{x}_i,t)}{\|\sigma_i(y_i,\hat{x}_i,t)\|} \quad (12)$$

Hence, from (4), (5) & (12) it can be shown that:

$$\textstyle\sum_{i=1}^n \dot{V}_i\left(\sigma_i(y_i,\hat{x}_i,t)\right) = \sum_{i=1}^n[-G_iC_iB_i\mu_i + \frac{\sigma_i(y_i,\hat{x}_i,t)}{\|\sigma_i(y_i,\hat{x}_i,t)\|}\,G_iC_iZ_i(t) + \frac{\sigma_i(y_i,\hat{x}_i,t)}{\|\sigma_i(y_i,\hat{x}_i,t)\|}\,G_iC_iB_iQ_i(x_i,t) + \frac{\sigma_i(y_i,\hat{x}_i,t)}{\|\sigma_i(y_i,\hat{x}_i,t)\|}G_iC_iE_id_i(t) + \frac{\sigma_i(y_i,\hat{x}_i,t)}{\|\sigma_i(y_i,\hat{x}_i,t)\|}\,G_iC_iB_if_i(t) + \frac{\sigma_i(y_i,\hat{x}_i,t)}{\|\sigma_i(y_i,\hat{x}_i,t)\|}G_iC_iA_ie_i(t)] \quad (13)$$

which can be re-written as:

$$\textstyle\sum_{i=1}^n \dot{V}_i\left(\sigma_i(y_i,\hat{x}_i,t)\right) \le \sum_{i=1}^n[-(G_iC_iB)[\mu_i - (G_iC_iB_i)^{-1}G_iC_i\|Z_i\| - \|Q_i\| - (G_iC_iB_i)^{-1}G_iC_iE_i\|d_i\| - \|f_i\| + (G_iC_iB_i)^{-1}G_iC_iA_i\|e_i(t)\|] \quad (14)$$

Then, according to A4, A5, A6 & A7:

$$\textstyle\sum_{i=1}^n \dot{V}_i\left(\sigma_i(y_i,\hat{x}_i,t)\right) \le \sum_{i=1}^n[-(G_iC_iB)[\mu_i - (G_iC_iB_i)^{-1}G_iC_i\beta_i(x_i,t) - \kappa_i\|x_i\| - (G_iC_iB_i)^{-1}G_iC_iE_i\gamma_i\,\|x_i\| - \eta_i\,\|x_i\| + (G_iC_iB_i)^{-1}G_iC_iA_i\|e_i(t)\|] \quad (15)$$

By suitable choice of $\mu_i$ in (11) then $\sum_{i=1}^n \dot{V}_i\left(\sigma_i(x_i,t)\right) \le 0$.

To minimize of norms $\|\Psi_i\zeta_i(t)\|$ and $\|\Psi_iE_id_i(t)\|$ corresponding to the unmatched uncertainty and disturbances, respectively, the matrix $G_i$ must be carefully chosen [13]. One choice is $G_i = B_i^TC_i^+$ which if substituted into (8) leads to the following:

**(i)** The term: $[I_i - B_i(B_i^+B_i)^{-1}B_i^+]B_i^\perp B_i^{\perp+}Z_i(t)$, with $B_i^TB_i^\perp = 0$, i.e.:

$$[I_i - B_i(B_i^+B_i)^{-1}B_i^+]B_i^\perp B_i^{\perp+}Z_i(t) = B_i^\perp B_i^{\perp+}Z_i(t) \quad (16)$$

**(ii)** The term:

$$[I_i - B_i(B_i^+B_i)^{-1}B_i^+]E_id_i(t) = [I_i - B_iB_i^+]E_id_i(t) \quad (17)$$

Substituting (16) & (17) into (8) yields the subsystem dynamics during sliding:

$$\dot{x}_i(t) = A_ix_i(t) + B_iu_i^{OBC}(t) + T_iZ_i(t) + H_id_i(t) - M_ie_i(t) \quad (18)$$

where $T_i = B_i^\perp B_i^{\perp+}$and $H_i = [I_i - B_iB_i^+]E_i$.

From (18) it can be observed that the unknown unmatched uncertainties and disturbances $T_iZ_i(t)\,, H_id_i(t)$ have not been minimized. Hence, another method must be found to minimize these terms and to limit their influence on the subsystem dynamics.

### B. *LMI OBSERVER-BASED CONTROL DESIGN*

After designing the ISMC, the subsystem sliding dynamics are:

$$\dot{x}_i(t) = A_ix_i(t) + B_iu_i^{OBC}(t) + \Gamma_iJ_i(t) - M_ie_i(t) \quad (19)$$

where $\Gamma_i = [T_i \quad H_i]$ and $J_i(t) = \begin{bmatrix} Z_i(t) \\ d_i(t) \end{bmatrix}$

The aggregated system dynamics are given by:

$$\dot{X}(t) = A_dX(t) + B_dU^{OBC}(t) + \Gamma_dJ(t) - M_de(t) \quad (20)$$

where: $X(t) = [x_1, x_2, \dots\dots, x_n]U^{OBC}(t) = [u_1, u_2, \dots\dots, u_n]$, $e(t) = [e_1, e_2, \dots\dots, e_n]$, $A_d = diag(A_i), B_d = diag(B_i)$, $\Gamma_d = diag(\Gamma_i)$ and $J(t) = [J_1, J_2, \dots\dots, J_n]$ , where "$diag$" represents the block diagonal matrix.

To develop a robust control law for the aggregate system consider a state estimate feedback of the form:

$$U^{OBC}(t) = K\hat{X}(t) = KX(t) - Ke(t) \quad (21)$$

where $K = diag(k_i)$ is the decentralized system gain that stabilizes the system under a specific performance objective. The design objective is to choose the gain $K$ to minimize the effect of $J(t)$ on the system. Suppose further that $J(t)$ is the unknown input disturbance which satisfies the quadratic inequality:

$$J^T(t)J(t) \le \alpha^2X^T(t)X(t) \quad (22)$$

where $\alpha > 0$ a positive constant. A suitable observer can estimate the aggregate system state $\hat{X}(t)$ any suitable observer can be used. However, the observer subsystems are given by:

$$\dot{\hat{x}}_i(t) = A_i\hat{x}_i(t) + B_i u_i^{OBC}(t) + L_i(y_i(t) - C_i\hat{x}_i(t)) \qquad (23)$$

where $L_i$ is the subsystem observer gain. The aggregate observer dynamics are thus:

$$\dot{\hat{X}}(t) = A_d\hat{X}(t) + B_d U^{OBC}(t) + L_d(Y(t) - C_d\hat{X}(t)) \qquad (24)$$

where $Y(t) = [y_1, y_2, \dots \dots, y_n]$ , $L_d = diag(L_i)$ and $C_d = diag(C_i)$

Subtracting (20) from (24) yields the state estimation error:

$$\dot{e}(t) = (A_d - L_d C_d)e(t) + \Gamma_d J(t) - M_d e(t) \qquad (25)$$

To check the stability of the observer-based closed-loop system the following candidate Lyapunov function is used:

$V(X,t) = X^T(t)PX(t) + e^T(t)Fe(t)$ where $P > 0$ & $F > 0$.

The time derivative of $V(X,t)$ is thus:

$$\dot{V}(X,t) = \dot{X}^T(t)PX(t) + X^T(t)P\dot{X}(t) + \dot{e}^T(t)Fe(t) + e^T(t)F\dot{e}(t) \qquad (26)$$

Substituting (21) and (20) into (26) and substituting (25) into (26):

$$\dot{V}(X,t) = X^T(t)[{A_d}^T P + K^T {B_d}^T P + PB_d K]X(t) - e^T[K^T {B_d}^T P - {M_d}^T P]X(t) + J^T(t){\Gamma_d}^T PX(t) + -X^T(t)[PB_d K - (t)PM_d]e(t) + X^T(t)P\Gamma_d J(t) + e^T(t)[{A_d}^T F - {C_d}^T {L_d}^T F - {M_d}^T F + FA_d - FL_d C_d - FM_d]e(t) + J^T(t){\Gamma_d}^T Fe(t) + e^T(t)F\Gamma_d J(t) \qquad (27)$$

The stability of subsystem (27) requires that $\dot{V}(X,t) < 0$ $\forall\, X(t) \neq 0$. Equation (27) can then be re-written as:

$$\mathcal{Z}^T\mathcal{D}\mathcal{Z} < 0 \qquad (28)$$

where: $\mathcal{Z} = \begin{bmatrix} X(t) \\ e(t) \\ J(t) \end{bmatrix}$ and

$$\mathcal{D} = \begin{bmatrix} {A_d}^T P + PA_d + K^T {B_d}^T P + PB_d K & -PB_d K - PM_d & P\Gamma_d \\ -K^T {B_d}^T P - {M_d}^T P & {A_d}^T F + FA_d + {C_d}^T {L_d}^T F + FL_d C_d - FM_d - {M_d}^T F & F\Gamma_d \\ {\Gamma_d}^T P & {\Gamma_d}^T F & 0 \end{bmatrix} \qquad (29)$$

To guarantee stability of the system (28) the matrix $\mathcal{D}$ must be negative-definite.

Furthermore, Eq (23) can be rewritten as:

$$\mathcal{Z}^T\mathcal{O}\mathcal{Z} \leq 0 \qquad (30)$$

where: $\mathcal{Z}_i = \begin{bmatrix} X(t) \\ e(t) \\ J(t) \end{bmatrix}$ and $\mathcal{O} = \begin{bmatrix} -\alpha^2 I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I \end{bmatrix}$.

To combine (28) & (30) into a single inequality the so-called S-procedure is now used [14].

If $\mathcal{D}$ and $\mathcal{O}$ can be considered as symmetric matrices then $\mathcal{Z}^T\mathcal{D}\mathcal{Z} < 0$ and $\mathcal{Z}^T\mathcal{O}\mathcal{Z} \leq 0$ . Hence, there is a number $\tau > 0$ where $-\tau\mathcal{O} < 0$ and it follows that:

$$\mathcal{D} - \tau\mathcal{O} = \begin{bmatrix} {A_d}^T P + PA_d + K^T {B_d}^T P + PB_d K + \tau\alpha^2 I & -PB_d K - PM_d & P\Gamma_d \\ -K^T {B_d}^T P - {M_d}^T P & {A_d}^T F + FA_d + {C_d}^T {L_d}^T F + FL_d C_d - FM_d - {M_d}^T F & F\Gamma_d \\ {\Gamma_d}^T P & {\Gamma_d}^T F & -\tau I \end{bmatrix} < 0 \qquad (31)$$

Substituting $\mathcal{Y} = \frac{P}{\tau}$ and $\Finv = \frac{F}{\tau}$ into (31) yields:

$$\Pi = \begin{bmatrix} {A_d}^T\mathcal{Y} + \mathcal{Y}A_d + K^T {B_d}^T\mathcal{Y} + \mathcal{Y}B_d K + \tau\alpha^2 I & -\mathcal{Y}B_d K - \mathcal{Y}M_d & \mathcal{Y}\Gamma_d \\ -K^T {B_d}^T\mathcal{Y} - {M_d}^T\mathcal{Y} & {A_d}^T\Finv + \Finv A_d + {C_d}^T {L_d}^T\Finv + \Finv L_d C_d - \Finv M_d - {M_d}^T\Finv & \Finv\Gamma_d \\ {\Gamma_d}^T\mathcal{Y} & {\Gamma_d}^T\Finv & -I \end{bmatrix} < 0 \qquad (32)$$

The inequality (32) cannot be solved via an LMI since it includes the term $B_d K$ to overcome this problem both sides of (32) must be multiplied by the matrix $\mathcal{W} = \begin{bmatrix} \mathcal{Y}^{-1} & 0 \\ 0 & \mathcal{T} \end{bmatrix}$ where $\mathcal{T} = \begin{bmatrix} \mathcal{Y}^{-1} & 0 \\ 0 & S \end{bmatrix}$ where $S$ is a design parameter.

Hence $\mathcal{W}\Pi\mathcal{W}^T = \begin{bmatrix} \mathcal{Y}^{-1} & 0 \\ 0 & \mathcal{T} \end{bmatrix}\begin{bmatrix} \Pi_{11} & \Pi_{12} \\ \Pi_{21} & \Pi_{22} \end{bmatrix}\begin{bmatrix} \mathcal{Y}^{-1} & 0 \\ 0 & \mathcal{T} \end{bmatrix} =$

$$\begin{bmatrix} \mathcal{Y}^{-1}\Pi_{11}\,\mathcal{Y}^{-1} & \mathcal{Y}^{-1}\Pi_{12}\mathcal{T} \\ \mathcal{T}\Pi_{21}\mathcal{Y}^{-1} & \mathcal{T}\Pi_{22}\mathcal{T} \end{bmatrix} \qquad (33)$$

Also, by making a partition of $\Pi_i$ into four parts:

$$\Pi_{11} = {A_d}^T\mathcal{Y} + \mathcal{Y}A_d + K^T {B_d}^T\mathcal{Y} + \mathcal{Y}B_d K + \tau\alpha^2 I \qquad (34)$$

$$\Pi_{12} = [-\mathcal{Y}B_d K - \mathcal{Y}M_d \qquad \mathcal{Y}\Gamma_d] \qquad (35)$$

$$\Pi_{21} = \begin{bmatrix} -K^T {B_d}^T\mathcal{Y} - {M_d}^T\mathcal{Y} \\ {\Gamma_d}^T\mathcal{Y} \end{bmatrix} \qquad (36)$$

$$\Pi_{22} = \begin{bmatrix} {A_d}^T\Finv + \Finv A_d + {C_d}^T {L_d}^T\Finv + \Finv L_d C_d - \Finv M_d - {M_d}^T\Finv & \Finv\Gamma_d \\ {\Gamma_d}^T\Finv & -I \end{bmatrix} \qquad (37)$$

The term $\mathcal{T}\Pi_{22}\mathcal{T}$ can then be described using [17] as:

$$\mathcal{T}\Pi_{22}\mathcal{T} \leq -\lambda(\mathcal{T} + \mathcal{T}^T) - \lambda^2\Pi_{22}^{-1} \qquad (38)$$

where $\lambda > 0$ is used for tuning to get an acceptable response.

$$\therefore\ \mathcal{W}\Pi\mathcal{W}^T = \begin{bmatrix} \mathcal{Y}^{-1}\Pi_{11}\,\mathcal{Y}^{-1} & \mathcal{Y}^{-1}\Pi_{12}\mathcal{T} \\ \mathcal{T}\Pi_{21}\mathcal{Y}^{-1} & -\lambda(\mathcal{T} + \mathcal{T}^T) - \lambda^2\Pi_{22}^{-1} \end{bmatrix} < 0 \qquad (39)$$

By using the Schur complement (39) could rewritten as:

$$\begin{bmatrix} \mathcal{Y}^{-1}\Pi_{11}\,\mathcal{Y}^{-1} & \mathcal{Y}^{-1}\Pi_{12}\mathcal{T} & 0 \\ \mathcal{T}\Pi_{21}\mathcal{Y}^{-1} & -2\lambda\mathcal{T} & \lambda I \\ 0 & \lambda I & \Pi_{22} \end{bmatrix} < 0 \qquad (40)$$

Substituting $\mathcal{P} = \mathcal{Y}^{-1}$ in (40) and also substituting (34), (35), (36) and (37) into (40) yields:

$$\begin{bmatrix} \mathbb{W} & -B_d K\mathcal{P} - M_d\mathcal{P} & \Gamma_d S & 0 & 0 \\ -\mathcal{P}K^T {B_d}^T - \mathcal{P}{M_d}^T & -2\lambda\mathcal{P} & 0 & \lambda I & 0 \\ S{\Gamma_d}^T & 0 & -2\lambda S & 0 & \lambda I \\ 0 & \lambda I & 0 & \mathbb{L} & F\Gamma_d \\ 0 & 0 & \lambda I & {\Gamma_d}^T\Finv & -I \end{bmatrix} < 0 \qquad (41)$$

where $\mathbb{W} = \mathcal{P}{A_d}^T + A_d\mathcal{P} + \mathcal{P}K^T {B_d}^T + B_d K\mathcal{P} + \alpha^2\mathcal{P}\mathcal{P}$ and $\mathbb{L} = {A_d}^T\Finv + \Finv A_d + {C_d}^T {L_d}^T\Finv + \Finv L_d C_d - \Finv M_d - {M_d}^T\Finv$

Choose $= I$ , and substitute $N = K\mathcal{P}$, $R = \Finv L_d$ and $\epsilon = \frac{1}{\alpha^2}$ , by using the Schur complement (41) is re-written as:

$$\begin{bmatrix} \overline{\overline{\mathbb{W}}} & -B_d N - M_d\mathcal{P} & \Gamma_d S & 0 & 0 & \mathcal{P} \\ -N^T {B_d}^T - \mathcal{P}{M_d}^T & -2\lambda\mathcal{P} & 0 & \lambda I & 0 & 0 \\ S{\Gamma_d}^T & 0 & -2\lambda S & 0 & \lambda I & 0 \\ 0 & \lambda I & 0 & \overline{\overline{\mathbb{L}}} & F\Gamma_d & 0 \\ 0 & 0 & \lambda I & {\Gamma_d}^T\Finv & -I & 0 \\ \mathcal{P} & 0 & 0 & 0 & 0 & -\epsilon I \end{bmatrix} < 0 \qquad (42)$$

where $\overline{\overline{\mathbb{W}}} = \mathcal{P}{A_d}^T + A_d\mathcal{P} + N^T {B_d}^T + B_d N$ ,
$\overline{\overline{\mathbb{L}}} = {A_d}^T\Finv + \Finv A_d + {C_d}^T R_i^T + RC_d - \Finv M_d - {M_d}^T\Finv$.
There are *two* approaches to solving the LMI (42)

*Algorithm 1:*
Minimize $\epsilon$ subject to $\mathcal{P} > 0$ , $\Finv > 0$ and (42).
To minimize the gain magnitude the conditioning of the matrices $N$ and $R$ in terms of norm bounds $\|N\|_2 < K_N I$ and $\|R\|_2 < K_R I$ are used as further inequality conditions [18]:
where $K_N$ and $K_R$ are scalar variables, and by using the Schur complement inequalities (43) and (44) can be added to (42) as follows:

$$\begin{bmatrix} -K_N I & N^T \\ N & -I \end{bmatrix} < 0 \text{ and } \begin{bmatrix} -K_R I & R^T \\ R & -I \end{bmatrix} < 0 \qquad (43)$$

Additional inequalities can be added to the matrices $\mathcal{P}$ and $\Finv$ [18].

$$\begin{bmatrix} \mathcal{P} & I \\ I & K_P I \end{bmatrix} > 0 \; and \; \begin{bmatrix} \Finv & I \\ I & K_{\Finv} I \end{bmatrix} > 0 \qquad (44)$$

where $K_P$ and $K_{\Finv}$ are scalar variables.

*Algorithm 2:*
Minimize $(\epsilon + K_N + K_P + K_R + K_{\Finv})$ subject to $\mathcal{P} > 0$ ,$\Finv > 0$ , (42) , (43) and (44).

## IV. Numerical Example

Consider a numerical example consisting of three inter-connected nonlinear systems.

*Subsystem 1:*

$$A_1 = \begin{bmatrix} 0 & -6 \\ 6 & 1 \end{bmatrix}, B_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, C_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, E_1 = \begin{bmatrix} 0.1 \\ 0.3 \end{bmatrix},$$

$$z_1 = (\begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} x_{31} \\ x_{32} \end{bmatrix})$$

$W_1(x_1, t) = \begin{bmatrix} 0 \\ 4\,cos(2t)x_{11} - 2\,sin(t)x_{12} \end{bmatrix}$, $x_1(0) = \begin{bmatrix} 0.4 \\ -0.1 \end{bmatrix}$ and $x_1(t) = \begin{bmatrix} x_{11}(t) \\ x_{12}(t) \end{bmatrix}$

*Subsystem 2:*

$$A_2 = \begin{bmatrix} 0 & -1 \\ -2 & -7 \end{bmatrix}, B_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, C_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, E_2 = \begin{bmatrix} 0.1 \\ 0.3 \end{bmatrix},$$

$$z_2 = (\begin{bmatrix} 0 & 0 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} x_{31} \\ x_{32} \end{bmatrix})$$

$W_2(x_2, t) = \begin{bmatrix} 0 \\ 2\,sin(t)x_{21} + 4\,cos(2t)x_{22} \end{bmatrix}$, $x_2(0) = \begin{bmatrix} 0.3 \\ -0.2 \end{bmatrix}$ and $x_2(t) = \begin{bmatrix} x_{21}(t) \\ x_{22}(t) \end{bmatrix}$

*Subsystem 3:*

$$A_3 = \begin{bmatrix} 0 & -1 \\ -4 & -5 \end{bmatrix}, B_3 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, C_3 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, E_3 = \begin{bmatrix} 0.1 \\ 0.3 \end{bmatrix},$$

$$z_3 = (\begin{bmatrix} 0 & 0 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ -2 & -1 \end{bmatrix}\begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix})$$

$W_3(x_3, t) = \begin{bmatrix} 0 \\ 6\,sin(t)x_{31} + 2\,cos(2t)x_{32} \end{bmatrix}$, $x_3(0) = \begin{bmatrix} -0.3 \\ -0.3 \end{bmatrix}$ and $x_3(t) = \begin{bmatrix} x_{31}(t) \\ x_{32}(t) \end{bmatrix}$

The systems without feedback are unstable.

*Simulation Results*

The continuous control $u_i^{OBC}(t)$ designed by LMI where the solution of Algorithm 2 yields the gains:

$K1 = [-4.9522 \quad -4.4306$ , $K2 = [5.6854 \quad 3.6946]$ and $K3 = [7.6854 \quad 1.6946]$

$L1 = \begin{bmatrix} 5.2947 & -0.1504 \\ 0 & 0 \end{bmatrix}$, $L2 = \begin{bmatrix} 7.7790 & 0 \\ 0 & 0 \end{bmatrix}$, $L3 = \begin{bmatrix} 7.7790 & 0 \\ 0 & 0 \end{bmatrix}$

where the value of the aggregate system tuning design parameter is $\lambda = 1.86$ , where $\mu_1 = \mu_2 = \mu_3 = 5$ and $\mathfrak{z}_1 = \mathfrak{z}_2 = \mathfrak{z}_3 = 0.2$

All three subsystems without controls are unstable. Fig.2 shows the response of all three subsystems using the output decentralized system (LMI+ISMC) with no faults. These results illustrate that the controllers give an acceptable response with stable subsystems.

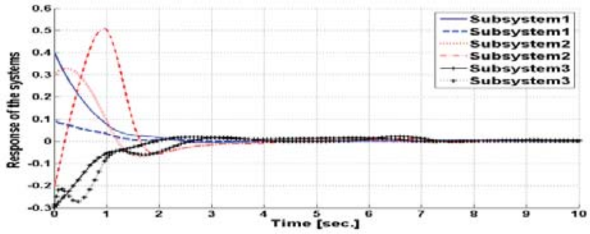


Figure.2 Three subsystems with controls and without faults

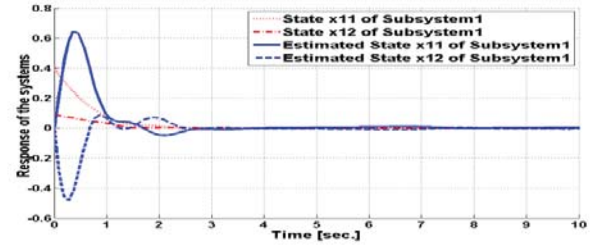


Figure.3 States of 1st subsystem and its estimated

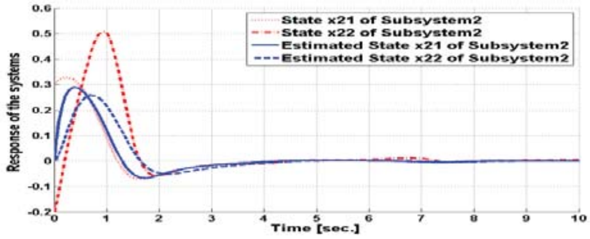


Figure.4 States of 2nd subsystem and it's estimated

Figs. 3, 4 & 5 illustrate the simulation results of the states of every subsystem and their estimates when there are no faults (actuators and sensors). In all subsystems from these results, the estimated and true state values are almost identical, with a little difference at the start of the simulation responses. These estimated states are used as feedback signals to control the subsystems.

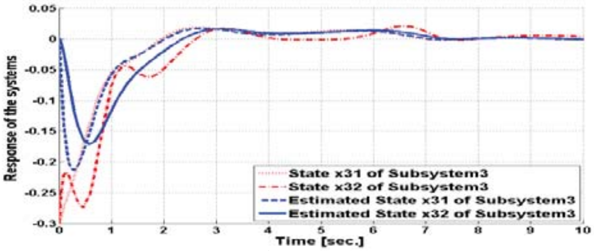


Figure.5 States of 3$^{th}$ subsystem and its estimated

For comparison with the results above, Fig.6 shows the response of all *three* independent linear subsystems using three separate observer-based controls to achieve output decentralized (LMI+ISMC) with no faults, no interconnections and no uncertainties. These results demonstrate very clearly the value of decentralized control approach described in the paper.

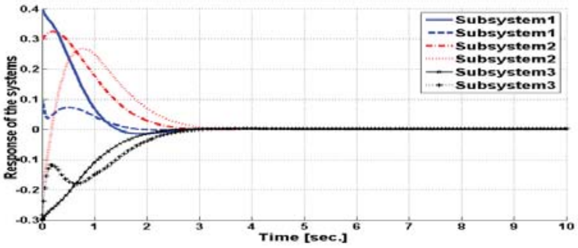


Figure. 6 Three linear subsystems with controls and without ( faults + interconnections + uncertainties)

## V. Conclusion

A major challenge of the control of uncertain inter-connected systems is to remove or compensate for the effects of uncertainly and disturbances acting in the subsystems so that an ideal decentralization can be achieved. In the ideal case, the resulting hitherto inter-connected system now becomes a truly decentralized structure in which the subsystems can be designed independently. This approach to control of complex systems has important consequences for security and fault-tolerance, e.g. if one subsystem fails then this failure does not influence the integrity of the remaining subsystems.

It is assumed in this work that the subsystem states are not available for control and hence the outputs are used together with the classical notion of state estimate feedback to develop a strategy for decentralization. Hence, the output decentralized control is achieved via ISMC together with linear observer design to give robust performance for both matched and unmatched uncertainty and disturbances. The design uses a single LMI to achieve stability of the aggregate system, minimization of matched/unmatched uncertainties and interactions and control performance specification. Whilst the design procedure is considerably complex the system implementation is nothing more than the ISMC and linear observers applied locally to each subsystem.

## References

[1] S. Dhbaibi, A. S. Tlili, S. Elloumi and N. B. Braiek, (2009), H infinity decentralized observation and control of nonlinear interconnected systems, *ISA Tranasactions*, <u>48</u>, pp. 458 - 467.

[2] C.-J. Mao and J-H. Yang,(1995), Decentralized Stabilization and Output Tracking of Large-scale Uncertain Systems, Automatica, <u>31</u>, 151 - 154.

[3] B. Labibi, H. J. Marquez and T. Chen,(2009), Decentralized robust output feedback control for control of affine nonlinear interconnected systems, *J. of Process Control*, <u>19</u> ,pp. 865- 878.

[4] B. Shafai, R. Ghadami, and M. Saif, (2011), Robust Decentralized PI Observer for Linear Interconnected Systems, *IEEE Symp. CACSD 2011.*

[5] P.R. Pagilla and Y. Zhu, (2005), A Decentralized Output Feedback Controller for a Class of Large-Scale Interconnected Nonlinear Systems, *ASME J. of Dynamic Systems, Measurement &Control*, <u>127</u>, pp. 167-172.

[6] S. Dhbaibi,A. S. Tlili and N. Benhadj , (2011), Decentrlized observer based tracking control of uncertain interconnected systems, 8th Int. Multi-Conference on Systems, Signals & Devices, SSD 201, March 22-25,Sousse, Tunisia

[7] S.S. Stankovic, D.M. Stipanovi¢ and D.D. Siljak, (2007), Decentralized Dynamic OutputFeedback for Robust Stabilization of a Class of Nonlinear Interconnected Systems, *Automatica*, <u>43</u> 861-867.

[8] S. Dhbaibi,A. S. Tlili and N. Benhadj , (2008), Decentralized observer based control for interconnected nonlinear systems. Application to multi-machine power systems, *J. Aut & Sys. Engineering*, <u>2</u>,(1), pp. 1- 17.

[9] R. Ghadami and B. Shafai,(2011), Decentralized PI Observer-Based Control of Nonlinear Interconnected Systems with Disturbance Attenuation, Proc. of *ACC '11*, San Francisco.

[10] A. Poznyak, L. Fridman, and F. J. Bejarano, (2004), Mini-Max Integral Sliding Mode Control for Multimodel Linear Uncertain Systems, *IEEE Tans. on Aut. Control*, <u>49</u>,(1), pp. 97- 102.

[11] F. Bejarano, L. Fridman, and A. Poznyak, (2007), Output integral sliding mode control based on algebraic hierarchical observer, *Int. Journal of Control,* <u>80</u>, pp. 443-453.

[12] P.R. Pagilla and Y. Zhu, (2007), Decentralized output feedback control of a class of large-scale interconnected systems , *IMA Journal of Mathematical Control and Information*, <u>24</u>, pp. 57 - 69

[13] F.Castaños, J-X.Xu, and L. Fridman, (2006), Integral sliding modes for systems with matched and unmatched uncertainties, Chapter 11 in: Advances in Variable Structure and Sliding Mode Control ( Edwards,Colet & Fridman, eds) , pp. 227 – 246, Springer.

[14] D. D. Siljak, and D.M. Stipanovic,(2001), Autonomus decentralized control,in Proceedings of the ASME International Mechanical Engineering Conference and Exposition, New York, NY, pp.761-765.

[15] C. Lin and R. J.Patton, (2010), Integral Hierarchical SMC of uncertain interconnected systems , *44$^{th}$ IEE CDC*, Atlanta, Dec 2010.

[16] Wen-Jun Cao and Jian –Xin Xu ,(2001), nonlinear integral- type sliding surface for both matched and unmatched uncertain systems, Proc. *ACC2001*, Arlington, USA, vol.6 , pp. 4369 - 4374

[17] D. Ichalal, B.Marx, J. Ragot and D. Maquin, (2010), Observer based actuator fault tolerant control for nonlinear Takagi-Sugeno systems : an LMI approach, *18th Mediterranean Conference on Control & Automation, MED'10,* Marrakech, June 23-25, 201

[18] A.Zecevic and D.Siljak, (2005), Global low-rank enhancement of decentralized control for large-scale systems, *IEEE Tans. Aut. Contr.*, <u>50</u>, (5), pp.740-744.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Synthesis of Variable Gain Controllers Based on LQ Optimal Control for a Class of Uncertain Linear Systems

Hidetoshi OYA
The Institute of Technology and Science,
The University of Tokushima,
2-1 Minamijosanjima, Tokushima, 770-8506 JAPAN
Email: hide-o@ee.tokushima-u.ac.jp

Yuhei UEHARA
The Graduate School of Advanced Technology and Science,
The University of Tokushima,
2-1 Minamijosanjima, Tokushima, 770-8506 JAPAN
Email: uehara@ee.tokushima-u.ac.jp

***Abstract*—This paper proposes a new variable gain controller for a class of uncertain linear systems. The proposed variable gain controller is based on optimal control for the nominal system and consists of the optimal feedback gain and a time-varying adjustable parameter which is designed so as to reduce the effect of uncertainties, i.e. the proposed variable gain controller can achieve good transient performance which is close to LQ optimal control for the nominal system. In this paper, we show sufficient conditions for the existence of the proposed variable gain controller for uncertain linear systems. Finally, numerical examples are presented.**

## I. Introduction

Robustness of control systems to uncertainties has always been the central issue in feedback control and therefore for linear systems with unknown parameters, a large number of design methods of robust controllers have been presented (e.g. [1] and references therein). For a system with structured uncertainties, several quadratic stabilizing control laws have also been suggested and a connection between quadratic stabilization and $\mathcal{H}^{\infty}$ control has been established[2]. It is well known that for robust control for linear dynamical systems with uncertainties, the concept of quadratic stabilization via fixed quadratic Lyapunov functions plays an important role in dealing with the controller design.

By the way in most practical situations, it is desirable to design robust control systems which achieve not only robust stability but also an adequate level of performance. Therefore robust controllers achieving some robust performances such as quadratic cost function, mixed $\mathcal{H}^{\infty}/\mathcal{H}^{2}$ control, robust $\mathcal{H}^{2}$ control and so on have been suggested (e.g.[3], [4], [5]). Additionally, synthesis problems of robust controllers with variable gain have also been tackled (e.g. [6], [7]). Yamamoto and Yamauchi[6] proposed a design method of a robust controller with the ability to adjust control performances adaptively. In [7], an adaptive robust controller with adaptation mechanism has been presented and the adaptive robust controller is tuned on-line based on the information about parameter uncertainties. Besides, we have proposed robust controllers with adaptive compensation inputs[8], [9]. Although the robust controllers in [8] and [9] can achieve not only asymptotical stability but also satisfactory transient behavior, these robust controllers include the additional dynamics of the nominal system. Namely, these robust controllers are dynamic one and their structure is more complex.

From these viewpoints, we propose a variable gain robust controller based on optimal control for a class of uncertain linear systems. The proposed variable gain controller consists of optimal feedback gain designed by using the nominal system and an adjustable time-varying parameter. The adjustable parameter is designed so as to reduce the effect of uncertainties. The proposed variable gain controller can achieve good transient performance which is close to the desirable trajectory generated by the nominal closed-loop system. This paper is organized as follows. In Sec. II, notation and useful lemmas which are used in this paper are shown, and in Sec. III, we introduce the class of uncertain linear systems under consideration. Sec. IV contains the main results. Finally, numerical examples are included to illustrate the results developed in this paper.

## II. Preliminaries

In this section, we show notations and useful and well-known lemmas which are used in this paper.

In the sequel, we use the following notation. For a matrix $\mathcal{A}$, The transpose of matrix $\mathcal{A}$ and the inverse of one are denoted by $\mathcal{A}^T$ and $\mathcal{A}^{-1}$ respectively and $\mathrm{rank}\left\{\mathcal{A}\right\}$ represents the rank of the matrix $\mathcal{A}$. Also, $H_e\{\mathcal{A}\}$ means $\mathcal{A}+\mathcal{A}^T$ and $I_n$ represents $n$-dimensional identity matrix and the notation $\mathrm{diag}\left(\mathcal{A}_1, \cdots, \mathcal{A}_N\right)$ denotes a block diagonal matrix composed of matrices $\mathcal{A}_i$ for $i = 1, \cdots, N$. For real symmetric matrices $\mathcal{A}$ and $\mathcal{B}$, $\mathcal{A} > \mathcal{B}$ (resp. $\mathcal{A} \geq \mathcal{B}$) means that $\mathcal{A} - \mathcal{B}$ is positive (resp. nonnegative) definite matrix. For a vector $\alpha \in \Re^n$, $||\alpha||$ denotes standard Euclidian norm and for a matrix $\mathcal{A}$, $||\mathcal{A}||$ represents a its induced norm. The symbols "$\stackrel{\triangle}{=}$" and "$\star$" means equality by definition and symmetric blocks in matrix inequalities, respectively.

Furthermore, the following well-known lemmas are used in this paper.

**Lemma 1:** For arbitrary vectors $\lambda$ and $\xi$ and the matrices $\mathcal{G}$ and $\mathcal{H}$ which have appropriate dimensions, the following relation holds.

$$H_e\left\{\lambda^T\mathcal{G}\Delta(t)\mathcal{H}\xi\right\} \leq 2\left\|\mathcal{G}^T\lambda\right\|\left\|\mathcal{H}\xi\right\|$$

where $\Delta(t) \in \mathbb{R}^{p\times q}$ is a time-varying unknown matrix satisfying $\left\|\Delta(t)\right\| \leq 1$.

*Proof:* The above relation is easily obtained by Schwartz's inequality[10]. ■

$$\begin{pmatrix} -\mathcal{Q}-(1+\tau_1)\mathcal{P}B\mathcal{R}^{-1}B^T\mathcal{P}+(\delta\tau_1+\tau_2)I_n & \Xi(\mathcal{P}) \\ \star & -\tau_2\Sigma_\sigma^{-1} \end{pmatrix} < 0 \tag{9}$$

**Lemma 2:** ($\mathcal{S}$-procedure) Let $\mathcal{F}(x)$ and $\mathcal{G}(x)$ be two arbitrary quadratic forms over $\mathbb{R}^n$. Then $\mathcal{F}(x) < 0$ for $\forall x \in \mathbb{R}^n$ satisfying $\mathcal{G}(x) \le 0$ if and only if there exist a nonnegative scalar $\tau$ such that

$$\mathcal{F}(x) - \tau\mathcal{G}(x) \le 0 \quad \text{for} \quad \forall x \in \mathbb{R}^n$$

*Proof:* See Boyd et al.[11] ■

**Lemma 3:** (Schur complement) For a given constant real symmetric matrix $\Xi$, the following arguments are equivalent.

(i) $\Xi = \begin{pmatrix} \Xi_{11} & \Xi_{12} \\ \Xi_{12}^T & \Xi_{22} \end{pmatrix} > 0$

(ii) $\Xi_{11} > 0$ and $\Xi_{22} - \Xi_{12}^T\Xi_{11}^{-1}\Xi_{12} > 0$

(iii) $\Xi_{22} > 0$ and $\Xi_{11} - \Xi_{12}\Xi_{22}^{-1}\Xi_{12}^T > 0$

*Proof:* See Boyd et al.[11] ■

## III. Problem Formulation

Consider the uncertain linear system described by the following state equation (see **Remark 1**).

$$\frac{d}{dt}x(t) = \left(A + \sum_{k=1}^{\mathcal{N}} \theta_k(t)\mathcal{D}_k\right)x(t) + Bu(t) \tag{1}$$

where $x(t) \in \Re^n$ and $u(t) \in \Re^m$ are the vectors of the state (assumed to be available for feedback) and the control input, respectively. In (1), the matrices $A$ and $B$ denote the nominal values of the uncertain system of (1). The matrices $\mathcal{D}_k$ $(k = 1, \cdots, N)$ which have appropriate dimensions represent the structure of uncertainties and the time-varying parameter vector $\theta(t) \in \Re^N(\theta(t) = (\theta_1(t), \cdots, \theta_{\mathcal{N}}(t))^T)$ shows unknown parameters which belong to the $\mathcal{N}$-dimensional ellipsoidal set expressed as

$$\begin{aligned} &\Delta \overset{\triangle}{=} \left\{\theta \in \Re^{\mathcal{N}} \mid \theta^T(t)\Sigma^{-1}\theta(t) \le 1\right\} \\ &\Sigma = diag\left(\sigma_1^2, \cdots, \sigma_{\mathcal{N}}^2\right) \end{aligned} \tag{2}$$

where $\Sigma \in \Re^{\mathcal{N}\times\mathcal{N}}$ represents the size of the ellipsoid. Beside the nominal system, ignoring the unknown parameters in (1), is given by

$$\frac{d}{dt}\overline{x}(t) = A\overline{x}(t) + B\overline{u}(t). \tag{3}$$

In this paper first of all, we consider the standard linear quadratic control problem for the nominal system of (3) in order to generate the desired response for the uncertain system of (1) systematically. Namely we define the following quadratic cost function for the nominal system of (3).

$$\mathcal{J} = \int_0^\infty \left(\overline{x}^T(t)\mathcal{Q}\overline{x}(t) + \overline{u}^T\mathcal{R}\overline{u}(t)\right)dt \tag{4}$$

where the matrices $\mathcal{Q} \in \Re^{n\times n}$ and $\mathcal{R} \in \Re^{m\times m}$ are positive definite. It is well-known that the optimal control input minimizing the quadratic cost function of (4) is given by $\overline{u}(t) = -K\overline{x}(t)$, where $K \in \Re^{m\times n}$ represent the optimal control gain matrix. Note that the closed-loop system matrix $A_K \overset{\triangle}{=} A - BK$ is stable and the optimal feedback gain matrix $K \in \Re^{m\times n}$ is derived as $K = \mathcal{R}^{-1}B^T\mathcal{P}$ where $\mathcal{P} \in \Re^{n\times n}$ is unique solution of the algebraic Riccati equation

$$H_e\left\{A^T\mathcal{P}\right\} - \mathcal{P}B\mathcal{R}^{-1}B^T\mathcal{P} + \mathcal{Q} = 0. \tag{5}$$

Now by using the optimal feedback gain matrix $K \in \Re^{m\times n}$ for the nominal system of (3), we consider the following control input.

$$u(t) \overset{\triangle}{=} \gamma(x,t)Kx(t) \tag{6}$$

where $\gamma(x,t) \in \Re^1$ is a time-varying adjustable parameter so as to compensate the effect of unknown parameters.

From eqs.(1) and (6), we have the closed-loop system

$$\frac{d}{dt}x(t) = Ax(t) + \Gamma(x,t)\theta(t) + \gamma(x,t)Kx(t). \tag{7}$$

In (7), $\Gamma(x,t)$ is a matrix expressed as

$$\Gamma(x,t) = \left(\ \mathcal{D}_1x(t), \mathcal{D}_2x(t), \cdots, \mathcal{D}_{\mathcal{N}}x(t)\ \right). \tag{8}$$

From the above discussion, our control objective in this paper is to design the robust stabilizing controller which achieves good transient performance for the uncertain closed-loop system of (7). That is to design the time-varying adjustable parameter $\gamma(x,t) \in \Re^1$ such that the closed-loop system of (7) is robustly stable and achieves satisfactory transient performance close to LQ optimal control for the nominal system of (3).

**Remark 1:** In this paper, we consider the uncertain dynamical system of (1) which has uncertainties in the state matrix only. The proposed design scheme of the variable controller derived in next section can also be applied to the case that the uncertainties are included in both the system matrix and the input matrix. By introducing additional actuator dynamics and constituting an augmented system, uncertainties in the input matrix are embedded in the system matrix of the augmented system[12]. Therefore the same design procedure can be applied.

## IV. Main Results

In this section, we show a design method of the proposed variable gain controller such that the uncertain system of (1) is asymptotically stable.

The following theorem gives sufficient conditions for the existence of the proposed controller.

**Theorem 1:** Consider the uncertain linear system of (1) and the control input of (6).

If there exist the positive scalars $\tau_1$ and $\tau_2$ satisfying the LMI of (9) then the adjustable time-varying parameter

$$\gamma(t) = \begin{cases} -\left(1 + \dfrac{\left\| \Sigma^{1/2} \Gamma^T(x,t) \mathcal{P} x(t) \right\|}{\left\| \mathcal{R}^{-1/2} B^T \mathcal{P} x(t) \right\|^2}\right) & \text{if} \quad x^T(t)\mathcal{P}B\mathcal{R}^{-1}B^T\mathcal{P}x(t) \geq \delta x^T(t)x(t) \\ -\left(1 + \dfrac{\left\| \Sigma^{1/2} \Gamma^T(x,t) \mathcal{P} x(t) \right\|}{\delta x^T(t)x(t)}\right) & \text{if} \quad x^T(t)\mathcal{P}B\mathcal{R}^{-1}B^T\mathcal{P}x(t) < \delta x^T(t)x(t) \end{cases} \tag{10}$$

$$\frac{d}{dt}\mathcal{V}(x,t) = x^T(t)\left[H_e\left\{A^T\mathcal{P}\right\}\right]x(t) + 2x^T(t)\Gamma(x,t)\theta(t) + 2\gamma(t)x^T(t)\mathcal{P}BKx(t) \tag{13}$$

$$\frac{d}{dt}\mathcal{V}(x,t) = -x^T(t)\left(\mathcal{Q} - \mathcal{P}B\mathcal{R}^{-1}B^T\mathcal{P}\right)x(t) + 2x^T(t)\Gamma(x,t)\theta(t) + 2\gamma(t)x^T(t)\mathcal{P}BKx(t) \tag{14}$$

$$\begin{aligned} \frac{d}{dt}\mathcal{V}(x,t) &= -x^T(t)\left(\mathcal{Q} - \mathcal{P}B\mathcal{R}^{-1}B^T\mathcal{P}\right)x(t) + 2x^T(t)\Gamma(x,t)\Sigma^{1/2}\Sigma^{-1/2}\theta(t) + 2\gamma(t)x^T(t)\mathcal{P}BKx(t) \\ &= -x^T(t)\left(\mathcal{Q} - \mathcal{P}B\mathcal{R}^{-1}B^T\mathcal{P}\right)x(t) + 2\left\|\Sigma^{1/2}\Gamma^T(x,t)x(t)\right\| + 2\gamma(t)x^T(t)\mathcal{P}B\mathcal{R}^{-1}B^T\mathcal{P}x(t) \end{aligned} \tag{15}$$

$$\frac{d}{dt}\mathcal{V}(x,t) = \begin{pmatrix} x(t) \\ \xi(t) \end{pmatrix}^T \begin{pmatrix} -\mathcal{Q} + \mathcal{P}B\mathcal{R}^{-1}B^T\mathcal{P} & \Xi(\mathcal{P}) \\ \star & 0 \end{pmatrix} \begin{pmatrix} x(t) \\ \xi(t) \end{pmatrix} + 2\gamma(t)x^T(t)\mathcal{P}B\mathcal{R}^{-1}B^T\mathcal{P}x(t) \tag{17}$$

$$\begin{aligned} &\begin{pmatrix} x(t) \\ \xi(t) \end{pmatrix}^T \begin{pmatrix} -\mathcal{Q} + \mathcal{P}B\mathcal{R}^{-1}B^T\mathcal{P} & \Xi(\mathcal{P}) \\ \star & 0 \end{pmatrix} \begin{pmatrix} x(t) \\ \xi(t) \end{pmatrix} + 2\gamma(t)x^T(t)\mathcal{P}B\mathcal{R}^{-1}B^T\mathcal{P}x(t) < 0 \\ &\text{s.t.} \quad x^T(t)\mathcal{P}B\mathcal{R}^{-1}B^T\mathcal{P}x(t) < \delta x^T(t)x(t) \text{ and } \xi^T(t)\Sigma_\sigma^{-1}\xi(t) \leq x^T(t)x(t) \end{aligned} \tag{20}$$

$$\begin{pmatrix} -\mathcal{Q} - \mathcal{P}B\mathcal{R}^{-1}B^T\mathcal{P} & \Xi(\mathcal{P}) \\ \star & 0 \end{pmatrix} < 0 \quad \text{s.t.} \quad x^T(t)\mathcal{P}B\mathcal{R}^{-1}B^T\mathcal{P}x(t) < \delta x^T(t)x(t) \text{ and } \xi^T(t)\Sigma_\sigma^{-1}\xi(t) \leq x^T(t)x(t) \tag{22}$$

$\gamma(t) \in \mathbb{R}^1$ is determined as (10). In (9), $\delta$ is a positive constant selected by designers and $\Xi(\mathcal{P})$ is a matrix given by

$$\Xi(\mathcal{P}) \overset{\triangle}{=} \begin{pmatrix} \mathcal{P}\mathcal{D}_1 & \mathcal{P}\mathcal{D}_2 & \cdots & \mathcal{P}\mathcal{D}_{\mathcal{N}} \end{pmatrix}. \tag{11}$$

Then the uncertain closed-loop system of (7) is robustly stable.

*Proof:* By using the unique solution $\mathcal{P} \in \mathbb{R}^{n \times n}$ of the algebraic Riccati equation of (5), we consider the following quadratic function.

$$\mathcal{V}(x,t) \overset{\triangle}{=} x^T(t)\mathcal{P}x(t) \tag{12}$$

The time derivative of the quadratic function $\mathcal{V}(x,t)$ can be written as (13). Additionally since the matrix $\mathcal{P}$ is the unique solution of the algebraic Riccati equation of (5), The time derivative of the quadratic function $\mathcal{V}(x,t)$ can be rewritten as (14).

Now, we consider the case of $x^T(t)\mathcal{P}B\mathcal{R}^{-1}B^T\mathcal{P}x(t) \geq \delta x^T(t)x(t)$. In this case using **Lemma 1**, we obtain (15). Here we have used the relation of (2) and $K = \mathcal{R}^{-1}B^T\mathcal{P}$. Besides, by using the adjustable time-varying parameter $\gamma(t)$ of (10), we find that the following relation holds.

$$\begin{aligned} \frac{d}{dt}\mathcal{V}(x,t) &= -x^T(t)\left(\mathcal{Q} + \mathcal{P}B\mathcal{R}^{-1}B^T\mathcal{P}\right)x(t) \\ &< 0 \quad \text{for} \quad \forall x(t) \neq 0 \end{aligned} \tag{16}$$

Next, we consider the case of $x^T(t)\mathcal{P}B\mathcal{R}^{-1}B^T\mathcal{P}x(t) < \delta x^T(t)x(t)$ and then the time derivative of the quadratic function $\mathcal{V}(x,t)$ of (14) can be described as (17). In (17), $\xi(t)$ is a $n \times \mathcal{N}$-dimensional vector given by

$$\xi^T(t) = \begin{pmatrix} \theta_1(t)x^T(t) & \theta_2(t)x^T(t) & \cdots & \theta_{\mathcal{N}}(t)x^T(t) \end{pmatrix}. \tag{18}$$

Note that from the relation of (2) the following inequality for the vector $\xi(t) \in \mathbb{R}^{n \times \mathcal{N}}$ is satisfied.

$$\xi^T(t)\Sigma_\sigma^{-1}\xi(t) \leq x^T(t)x(t) \tag{19}$$

In (19), $\Sigma_\sigma = \mathrm{diag}\left(\sigma_1^2 I_n, \sigma_2^2 I_n, \cdots, \sigma_{\mathcal{N}}^2 I_n\right)$. One can see that if the condition of (20) holds, then the following inequality is also satisfied.

$$\frac{d}{dt}\mathcal{V}(x,t) < 0 \quad \text{for} \quad \forall x(t) \neq 0 \tag{21}$$

Thus we consider the condition of (20). By using the adjustable time-varying parameter $\gamma(t)$ of (10), we have (22) which is a sufficient condition for the inequality of (20). Namely, if the condition of (22) holds, then the inequality of (20) is also satisfied. Applying **Lemma 2** ($\mathcal{S}$-procedure) to the condition of (22) and some trivial manipulations give the LMI of (9). Therefore for the case of $x^T(t)\mathcal{P}B\mathcal{R}^{-1}B^T\mathcal{P}x(t) < \delta x^T(t)x(t)$, if the LMI of (9) is feasible then the relation of (21) is satisfied.

From the above discussion, the quadratic function $\mathcal{V}(x,t)$ becomes a Lyapunov function and the uncertain linear system of (1) is ensured to be stable. It follows that the result of the theorem is true. The proof of **Theorem 1** is completed. ■

**Remark 2:** In this paper, the quadratic function $\mathcal{V}(x,t)$ of (12) is introduced and it becomes a Lyapunov function for the uncertain system of (1). On the other hand, the quadratic function $\mathcal{V}(x,t)$ of (12) is also a Lyapnov function for the nominal closed-loop system, i.e. the standard LQ regulator. Therefore, by selecting the design parameter $\delta \in \mathbb{R}^1$ the proposed controller can achieve the good transient performance and adjust the magnitude of the control input, because the Lyapunov function for the uncertain system of (1) and one of the nominal system of (3) have same level set.

$$\frac{d}{dt}x(t) = \begin{pmatrix} -3.0 & 1.0 \\ 0.0 & 1.0 \end{pmatrix} x(t) + \delta_1(t) \begin{pmatrix} 1.0 & 1.0 \\ 0.0 & 0.0 \end{pmatrix} x(t) + \delta_2(t) \begin{pmatrix} 0.0 & 0.0 \\ 0.0 & 1.0 \end{pmatrix} x(t) + \begin{pmatrix} 0.0 \\ 1.0 \end{pmatrix} u(t) \tag{23}$$

$$\begin{aligned} &\bullet \text{ Case 1)} \ : \ \delta_1(t) = \sqrt{3.0} &&, \ \delta_2(t) = -\sqrt{2.0} \\ &\bullet \text{ Case 2)} \ : \ \delta_1(t) = \sqrt{3.0} \times \sin(5\pi t) &&, \ \delta_2(t) = -\sqrt{2.0} \times \cos(5\pi t) \end{aligned} \tag{27}$$

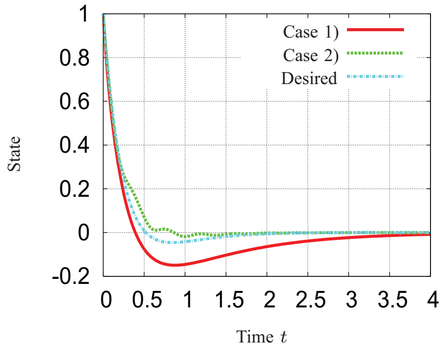


Fig. 1. Time histories of the state $x_1(t)$ : $\Sigma_1^*$

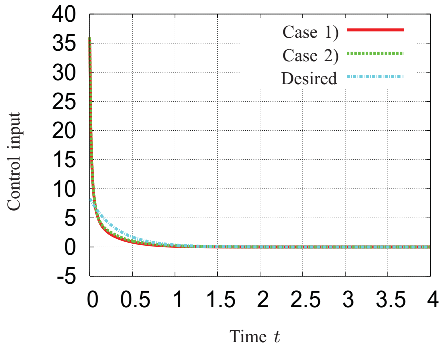


Fig. 3. Time histories of the control input $u(t)$ : $\Sigma_1^*$

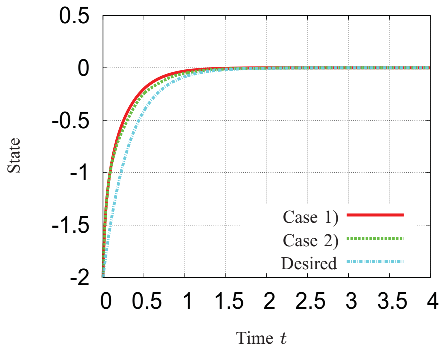


Fig. 2. Time histories of the state $x_2(t)$ : $\Sigma_1^*$

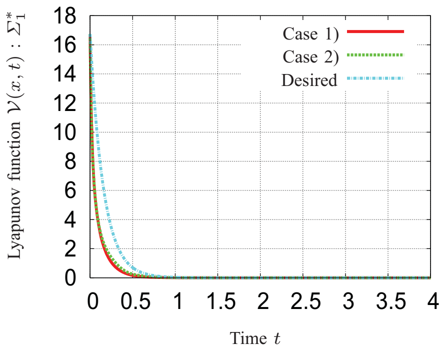


Fig. 4. Time histories of the Lyapunov function $\mathcal{V}(x,t)$ : $\Sigma_1^*$

## V. Illustrative Examples

In order to demonstrate the efficiency of the proposed control scheme, we have run a simple example. Consider the linear system with unknown parameters of (23) and we assume that the parameters $\sigma_1$ and $\sigma_2$ in the matrix $\Sigma \in \mathbb{R}^{2\times 2}$ in (2) are given by $\sigma_1 = 2.0$ and $\sigma_2 = 5.0 \times 10^{-1}$, respectively.

Firstly we select the weighting matrices $\mathcal{Q}$ and $\mathcal{R}$ such as $\mathcal{Q} = 1.0 \times I_2$ and $\mathcal{R} = 9.0$ for the quadratic cost function for the standard linear quadratic control problem, respectively. Then solving the algebraic Riccati equation of (5), we obtain

$$\begin{aligned} K &= \begin{pmatrix} 2.69892 \times 10^{-2} & -4.17080 \end{pmatrix} \\ \mathcal{P} &= \begin{pmatrix} 1.66545 \times 10^{-1} & 2.69892 \times 10^{-2} \\ \star & 4.17080 \end{pmatrix} \end{aligned} \tag{24}$$

In this example, we consider the following two kinds of the design parameters $\delta \in \mathbb{R}^1$ in (9) .

$$\bullet \ \Sigma_1^* \ : \ \delta = 1.0 \times 10^2, \quad \bullet \ \Sigma_2^* \ : \ \delta = 4.0 \times 10^5 \tag{25}$$

For these design parameters, solving the LMI condition of (9), we have

$$\begin{aligned} &\bullet \ \Sigma_1^* \ : \tau_1 = 1.00121 \times 10^{-7} &&, \tau_2 = 3.32679 \times 10^1 \\ &\bullet \ \Sigma_2^* \ : \tau_1 = 1.0 \times 10^{-7} &&, \tau_2 = 4.13381 \times 10^1 \end{aligned} \tag{26}$$

Now in this example, we consider the two cases for the unknown parameters in (27). Furthermore, the initial value for the uncertain system of (23) and its nominal system are selected as $x(0) = \overline{x}(0) = \begin{pmatrix} 1.0 & -2.0 \end{pmatrix}^T$.

The results of the simulation of this example are depicted in Figures 1 – 8. In these figures, "Case 1)" and "Case 2)" represent the time-histories of the state variables $x_1(t)$ and $x_2(t)$, the control input $u(t)$ and the Lyapunov function $\mathcal{V}(x,t)$. Besides, "Desired" represents the desirable transient behavior, the control input and the time-histories of the Lyapunov function $\mathcal{V}(x,t)$ generated by the nominal system.

From Figures 1 – 4, we find that the proposed variable gain controller ($\Sigma_1^*$) achieves good transient performance. However, the proposed control input is excessive comparing with the

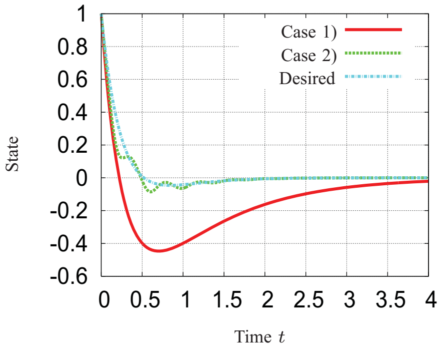


Fig. 5. Time histories of the state $x_1(t)$ : $\Sigma_2^*$

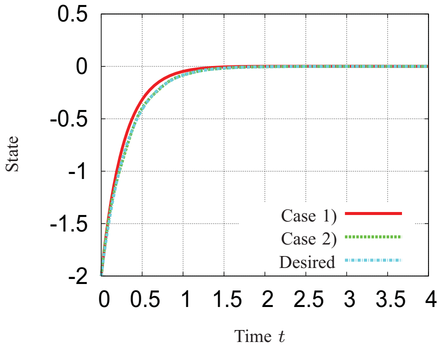


Fig. 6. Time histories of the state $x_2(t)$ : $\Sigma_2^*$

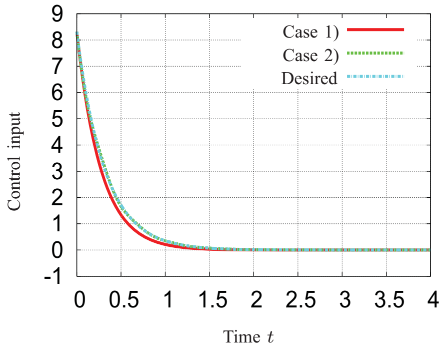


Fig. 7. Time histories of the control input $u(t)$ : $\Sigma_2^*$

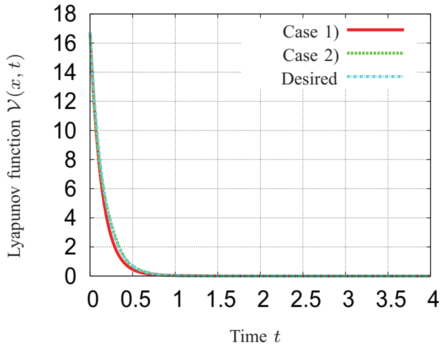


Fig. 8. Time histories of the Lyapunov function $\mathcal{V}(x,t)$ : $\Sigma_2^*$

nominal system. On the other hand, one can see from Figures 5 – 8 that although the error between the transient response for the proposed controller ($\Sigma_2^*$) and the one of the nominal system is large, the control input in $\Sigma_2^*$ is close to the desired one. Namely, the proposed controller can adjust the transient performance and the control input by means of selecting the design parameter $\delta \in \mathbb{R}^1$ in (10). Therefore the effectiveness of the proposed variable gain controller is shown.

## VI. Conclusions

In this paper we have proposed a new variable gain controller for a class of uncertain linear systems. Besides, by numerical simulations, the effectiveness of the proposed controller has been presented. One can see that the crucial difference between the existing results[8], [9] and our new one is that the structure of proposed controller is simple and the proposed variable gain controller can adjust the transient performance and the control input by means of selecting the design parameter.

The future research subjects are an extension of the proposed controller to such a broad class of systems as uncertain large-scale systems, uncertain discrete-time systems, uncertain time-delay systems and so on.

## Acknowledgment

The authors would like to thank CAE Solutions Corp. for providing its support in conducting this study.

## References

[1] K. Zhou, "Essentials of Robust Control", Prentice Hall Inc., 1998.

[2] J. C. Doyle and K. Glover and P. P. Khargonekar and B. A. Francis, "State-Space Solutions to Standarad $\mathcal{H}^2$ and $\mathcal{H}^\infty$ Control Problems", IEEE Trans. Automat. Contr., vol.34, no.8, pp.831–847, 1989.

[3] I. R. Petersen and D. C. McFarlane, "Optimal Guaranteed Cost Control and Filtering for Uncertain Linear Systems", IEEE Trans. Automat. Contr., Vol.39, No.9, pp.1971-1977, 1994.

[4] P. P. Khargonekar and M. A. Rotea, "Mixed $\mathcal{H}^2/\mathcal{H}^\infty$ Control , A Convex Optimization Approach", IEEE Trans. Automat. Contr.,Vol.36, No.7, pp.824-837, 1991.

[5] A. A. Stoorvogel, "The Robust $H_2$ Control Problem,A Worst-Case Design", IEEE Trans. Automat. Contr., Vol.38, No.9, pp.1358-1370, 1993.

[6] S. Yamamoto and K. Yamauchi, "A Design Method of Adaptive Control Systems by a Time-Varying Parameter of Robust Stabilizing State Feedback", Tran. ISCIE(in Japanese), vol.12, no.6, pp.319-325, 1999.

[7] M. Maki and K. Hagino, "Robust Control with Adaptation Mechanism for Improving Transient Behavior", Int. J. Contr., vol.72, no.13, pp.1218-1226, 1999.

[8] H. Oya and K. Hagino, "Robust Control with Adaptive Compensation Input for Linear Uncertain Systems", IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences, vol.E86-A, no.6, pp.1517-1524, 2003.

[9] H. Oya and K. Hagino, "Adaptive Robust Control Scheme for Linear Systems with Structured Uncertainties", IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences, Vol.E87-A, No.8, pp.2168-2173, 2004.

[10] F. R. Gantmacher, "The Theory of Matrices", Vol.1, Chelsea Publishing Company, New York, 1960.

[11] S. Boyd, L. El Ghaoui, E. Feron and V. Balakrishnan, Linear Matrix Inequalities in System and Control Theory, SIAM Studies in Applied Mathematics, 1994.

[12] K. Zhou and P. P. Khargonekar, "Robust Stabilization on Linear Systems with Norm Bounded Time-Varying Uncertainty", Systems & Control Letters, Vol.10, No.1, pp.17-20, 1988.

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# Output regulation for switched linear systems with different coordinate transformations

Xiao Xiao Dong*, Xi Ming Sun†, Jun Zhao ‡, and Georgi M. Dimirovski§
*State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang,
China 110819. Email: dongxiaoxiao0331@sina.com
†School of Control Science and Engineering, Dalian University of Technology, Dalian 116024, PRC;
Email: sunxm@dlut.edu.cn
‡State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang,
China 110819. Email:zhaojun@ise.edu.cn
§School of Engineering, Dogus University, Istanbul,
Turkey, TR-34722; School FEEIT, SS Cyril and Methodius University,
Skopje, Macedonia MK-1000. Email: gdimirovski@dogus.edu.tr

***Abstract*—This paper addresses the output regulation problem for switched linear systems. When each regulation equation has their own solution, we give a sufficient condition for the output regulation problem to be solvable. Firstly, we give the regulation equations of switched linear systems and the relation of the transformed states between two consecutive switching times. Secondly, the existence of a minimal average dwell-time for every switching sequence is assumed, and by virtue of an appropriate Lyapunov analysis, the output regulation is achieved. Our result is of much less conservativeness.**

## I. INTRODUCTION

A switched system is a special kind of hybrid system that consists of a family of continuous time or discrete-time subsystems and a rule orchestrating the switching among the subsystems [1]. This type of systems has wide applications in many fields, for instance, stochastic control [2], fault tolerant cooperative control [3], power systems [4], autonomous aircraft control [5]. A lot of methods have been presented to study stability and stabilization problems for switched systems, such as the convex combination technique, the common Lyapunov function technique, the multiple Lyapunov function method, the switched Lyapunov function method and the average dwell-time approach [6]-[9].

Output regulation is an important and interesting problem in control theory. This problem aims to achieve asymptotic tracking and disturbance rejection for a class of reference inputs and disturbances, which generated by an exosystem, besides closed-loop stability. Thus, the problem of output regulation is more challenging than stabilization and has attracted much attention. The problem for non-switched linear systems was widely studied in [10]-[12]. For non-switched nonlinear systems, there are also many results for the problem [13]-[15].

It is well known that for a non-switched linear system, in order to obtain the solvability condition for the output regulation problem, we need to solve a group of regulator equations. However, for a switched linear system, the solvability of the output regulation problem is more difficult and complicated. In order to get sufficient conditions for the problem, we need to solve a family of groups of regulator equations. If this family of systems has a common solution, then one can give the solvability conditions for the problem of switched linear systems. For instance, the problem for a class of switched systems with disturbances is solved under the assumption that the output regulation problem of a convex combination system is solvable [16]. A necessary and sufficient condition for the output regulation problem of switched systems under an arbitrary switching signal to be solvable is obtained in [17]. In addition, the multiple Lyapunov function method is used to solve the problem of output regulation for a class of discrete-time switched systems in [18]. The optimal output regulation is also guaranteed for the discrete switched linear system in [19]-[20]. However, to the authors' best knowledge, the output regulation problem has not been investigated for different coordinate transformations, which motivates the present study.

In this paper, for reducing the conservativeness, we consider the regulation equations of each subsystem has its own solution. This means that each subsystem of the switched linear system has the different transformation and complicates the solvability of the problem. Based on this case, we fist give the relation of the transformed states between two consecutive switching times. Then, the existence of a minimal average dwell-time for every switching sequence is assumed, and by virtue of an appropriate Lyapunov analysis, the output regulation is achieved.

This paper is organized as follows. Section 2 presents the problem statement and preliminaries. In section 3, the sufficient conditions for the output regulation problem to be solvable are given based on full information feedback controllers and error feedback controllers. Conclusion is stated in Section 4.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

## Problem statements and preliminaries

Consider a switched linear system modelled by equations of the form

$$\begin{aligned} \dot{x} &= A_{\sigma(t)}x + B_{\sigma(t)}u + P_{\sigma(t)}\omega, \\ e &= C_{\sigma(t)}x + Q_{\sigma(t)}\omega, \end{aligned} \tag{1}$$

with the state $x \in R^n$, the control input $u \in R^m$, the switching signal $\sigma : [0, \infty) \to I_N = \{1, \cdots, N\}$ is a piecewise constant function of time, the error variable $e \in R^p$, the exogenous input variable $\omega \in R^r$ satisfying the following exosystem

$$\dot{\omega} = S\omega. \tag{2}$$

We are in a position to state the output regulation problem of the switched system (1).

**Output regulation via full information**. Given $\{A_i, B_i, P_i, C_i, Q_i, S\}$ , $i \in I_N$, find full information controllers

$$u = K_i x + L_i\omega \tag{3}$$

and switching law $\sigma(t)$ such that:

1. the system (1) with the controllers (3) is asymptotically stable under the designed switching law $\sigma(t)$ without disturbance input.

2. for each $(x^0, \omega^0)$, the solution $(x(t), \omega(t))$ of

$$\begin{aligned} \dot{x} &= (A_{\sigma(t)} + B_{\sigma(t)}K_{\sigma(t)})x + (P_{\sigma(t)} + B_{\sigma(t)}L_{\sigma(t)})\omega, \\ \dot{\omega} &= S\omega \end{aligned} \tag{4}$$

satisfying $(x(0), \omega(0)) = (x^0, \omega^0)$ is such that

$$\lim_{t\to\infty}(C_{\sigma(t)}x + Q_{\sigma(t)}\omega) = 0.$$

**Output regulation via error feedback.** Given $\{A_i, B_i, P_i, C_i, Q_i, S\}$, $i \in I_N$, find error feedback controllers

$$\begin{aligned} \dot{\xi} &= F_i\xi + G_i e, \\ u &= H_i\xi \end{aligned} \tag{5}$$

and switching law $\sigma(t)$ such that:

1. the system (1) with the controllers (5) is asymptotically stable under the designed switching law $\sigma(t)$ without disturbance input.

2. for each $(x^0, \xi^0, \omega^0)$, the solution $(x(t), \xi(t), \omega(t))$ of

$$\begin{aligned} \dot{x} &= A_{\sigma(t)}x + B_{\sigma(t)}H_{\sigma(t)}\xi + P_{\sigma(t)}\omega, \\ \dot{\xi} &= G_{\sigma(t)}C_{\sigma(t)}x + F_{\sigma(t)}\xi + G_{\sigma(t)}Q_{\sigma(t)}\omega, \\ \dot{\omega} &= S\omega \end{aligned} \tag{6}$$

satisfying $(x(0), \xi(0), \omega(0)) = (x^0, \xi^0, \omega^0)$ is such that

$$\lim_{t\to\infty}(C_{\sigma(t)}x + Q_{\sigma(t)}\omega) = 0.$$

**Remark 1**. The system and the output regulation problem for non-switched system are given in [21]. In [16], the system model (1) and the description of output regulation problem are presented for switched linear systems.

In what follows, we assume the exosystem (2) satisfies the following assumption.

**Assumption 1(A1)** [22]. The system (2) is antistable, i.e. all the eigenvalues of $S$ have nonnegative real part.

Before obtaining the main results, we give a definition of the average dwell time.

**Definition 1** [23] [24]. For switching signal $\sigma(t)$ and any $T \geq t \geq 0$, let $N_\sigma(T, t)$ be the switching numbers of $\sigma(t)$ over the interval $(t, T)$. If for any given $N_0 \geq 0$ and $\tau_a > 0$, we have $N_\sigma(t, T) \leq N_0 + (T - t)/\tau_a$, then $\tau_a$ and $N_0$ are called average dwell time and the chatter bound, respectively.

## Main results

In this section, we will give the solvability conditions for the output regulation problem for the switched linear system (1) with the full information controllers and the error feedback controllers, where the regulation equations for the switched linear system (1) are given.

Now, we give a sufficient condition for the output regulation problem to be solvable based on the full information feedback.

**Theorem 1.** If there exist $\Pi_i$ , $\Gamma_i$ for $\forall i \in I_N$, satisfying the following equations

$$\begin{aligned} \Pi_i S &= A_i\Pi_i + B_i\Gamma_i + P_i, \\ 0 &= C_i\Pi_i + Q_i, \end{aligned} \tag{7}$$

and the system (1) with different coordinate transformation

$$\begin{aligned} \tilde{x}(t_k^-) &= x(t_k^-) - \Pi_{i_k}\omega(t_k^-), \\ \tilde{x}(t_k^+) &= x(t_k^-) - \Pi_{i_{k+1}}\omega(t_k^-), \end{aligned} \tag{8}$$

and

$$x(t_k^-) - \Pi_{i_{k+1}}\omega(t_k^-) = T_{i_k,i_{k+1}}(x(t_k^-) - \Pi_{i_k}\omega(t_k^-)), \tag{9}$$

where $T_{i_k,i_{k+1}}$, $k \in N$, are given matrices, then under the average dwell time

$$\tau_a > \tfrac{ln\mu}{\lambda_0}, \tag{10}$$

where

$$\mu = \sup_{1\leq i,j\leq N} \parallel P^{1/2}T_{i,j}P^{1/2} \parallel, \lambda_0 = \inf_{1\leq i\leq N} \tfrac{\lambda_{\min}(Q_i\prime)}{\lambda_{\max}(P)}, \tag{11}$$

and

$$(A_i + B_iK_i)^TP + P(A_i + B_iK_i) = -Q_i\prime, \tag{12}$$

the full information feedback controllers (3) solve the problem of output regulation for the switched system (1).

*Proof.* Set $L_i = \Gamma_i - K_i\Pi_i$ , and consider the coordinate transformation $\tilde{x} = x - \Pi_i\omega$. According to (7), the system (4) is rewritten as

$$\begin{aligned} \dot{\tilde{x}} &= (A_{\sigma(t)} + B_{\sigma(t)}K_{\sigma(t)})\tilde{x} = \tilde{A}_{\sigma(t)}\,\tilde{x}, \\ \dot{\omega} &= S\omega, \\ e &= C_{\sigma(t)}\tilde{x}. \end{aligned} \tag{13}$$

Therefore, the output regulation problem for the switched system (1) is equivalently converted to the stabilization problem of the system (13). According to (9), we have the relation of the transformed states between the two consecutive switching times $t_k^+$ and $t_k^-$ as follows

$$x(t_k^-) - \Pi_{i_{k+1}}\omega(t_k^-) = T_{i_k,i_{k+1}}(x(t_k^-) - \Pi_{i_k}\omega(t_k^-)).$$

We define the following Lyapunov function candidate

$$V(\tilde{x}) = \tilde{x}^T P \tilde{x}, \tag{14}$$

where $P$ is a positive definite matrix satisfying (12). When $t \in [t_k, t_{k+1})$, we have

$$\begin{aligned} V(t) &\leq \exp\left(-\tfrac{\lambda_{\min}(Q_{i_{k+1}})'}{\lambda_{\max}(P)}(t_{k+1}-t_k)\right)V(t_k^+) \\ &\leq \exp(-\lambda_0(t_{k+1}-t_k))V(t_k^+). \end{aligned}$$

Because the coordinate transformations are different, $V(t_k^+) \neq V(t_k^-)$. By virtue of the properties of the positive definite matrices, there exist $P^{\frac{1}{2}} > 0$ such that $P = P^{\frac{1}{2}}P^{\frac{1}{2}}$, applying (9), we obtain

$$\begin{aligned} V(t_k^+) &= \parallel P^{1/2}\tilde{x}(t_k^+) \parallel^2 = \parallel P^{1/2}T_{i_k,i_{k+1}}\tilde{x}(t_k^-) \parallel^2 \\ &\leq \parallel P^{1/2}T_{i_k,i_{k+1}}P^{-1/2} \parallel^2 \parallel P^{1/2}\tilde{x}(t_k^-) \parallel^2 \\ &\leq \parallel P^{1/2}T_{i_k,i_{k+1}}P^{-1/2} \parallel^2 V(t_k^-) \\ &\leq \mu V(t_k^-). \end{aligned}$$

Based on the above inequalities, we have

$$V(t_{k+1}^+) \leq \mu \exp(-\lambda_0(t_{k+1}-t_k))V(t_k^+). \tag{15}$$

Repeating the inequality (15) from $k=0$ to $k=N_\sigma - 1$ yields

$$V(t^-) \leq V(t_{N_\sigma}) \leq \mu^{N_{\sigma(t)}} \exp(-\lambda_0 t)V(0).$$

According to $N_\sigma(T,t) \leq N_0 + \frac{T-t}{\tau_a}$, we get $V(t^-) \leq \mu^{N_0} \exp\left(t\left(\frac{\ln\mu}{\tau_a} - \lambda_0 t\right)\right)V(0)$. Then, the system (4) without the disturbance input is asymptotically stable under the average dwell time (10). Meanwhile, we get

$$\lim_{t\to\infty}(C_{\sigma(t)}x + Q_{\sigma(t)}w) = \lim_{t\to\infty} C_{\sigma(t)}\tilde{x} = 0.$$

Therefore, we can conclude that the output regulation problem is solved.

**Remark 2.** Note that, each group of regulation equations (7) has their own solution for each subsystem, which leads to the switched system has different coordinate transformations. The above Theorem solves the problem of different transformations for the switched linear systems.

Next, we give another sufficient condition for the problem to be solvable via error feedback.

**Theorem 2.** If there exist $\Pi_i, \Sigma_i, H_i,\ F_i\ \bar{T}_{i,j}$ for $\forall i, j \in I_N$, satisfying the following equations

$$\begin{aligned} &\Pi_i S = A_i\Pi_i + B_iH_i\Sigma_i + P_i, \\ &\Sigma_i S = F_i\Sigma_i, \\ &0 = C_i\Pi_i + Q_i, \end{aligned} \tag{16}$$

and the system (1) with different coordinate transformations

$$\begin{aligned} &\tilde{\chi}(t_k^-) = \chi(t_k^-) - \bar{\Pi}_{i_k}\omega(t_k^-), \\ &\tilde{\chi}(t_k^+) = \chi(t_k^-) - \bar{\Pi}_{i_{k+1}}\omega(t_k^-), \end{aligned} \tag{17}$$

and

$$\chi(t_k^-) - \bar{\Pi}_{i_k}\omega(t_k^-) = \bar{T}_{i,j}(\chi(t_k^-) - \bar{\Pi}_{i_{k+1}}\omega(t_k^-)), \tag{18}$$

where

$$\tilde{\chi} = \left(\begin{array}{c} x - \Pi_i\omega \\ \xi - \Sigma_i\omega \end{array}\right), \chi = \left(\begin{array}{c} x \\ \xi \end{array}\right), \bar{\Pi}_i = \left(\begin{array}{c} \Pi_i \\ \Sigma_i \end{array}\right),$$

then under the average dwell time

$$\tau_a > \tfrac{ln\mu}{\lambda_0}, \tag{19}$$

where

$$\mu = \sup_{1\leq i,j\leq N} \parallel P^{1/2}\bar{T}_{i,j}P^{1/2} \parallel, \lambda_0 = \inf_{1\leq i\leq N} \tfrac{\lambda_{\min}(Q_i\prime)}{\lambda_{\max}(P)}, \tag{20}$$

and

$$\bar{A}_i^T P + P\bar{A}_i = -Q_i\prime, \tag{21}$$

the error feedback controllers (5) solve the problem of output regulation for the switched system (6), where $\bar{A}_i = \left(\begin{array}{cc} A_i & B_iH_i \\ G_iC_i & F_i \end{array}\right)$.

*Proof.* Based on (16), we can rewrite the regulation equations as follows

$$\begin{aligned} &\bar{\Pi}_i S = \bar{A}_i\bar{\Pi}_i + \bar{P}_i, \\ &0 = \bar{C}_i\bar{\Pi}_i + Q_i, \end{aligned} \tag{22}$$

where

$$\bar{P}_i = \left(\begin{array}{c} P_i \\ G_iQ_i \end{array}\right), \bar{C}_i = \left[\begin{array}{cc} C_i & 0 \end{array}\right].$$

Consider the transformations

$$\tilde{\chi} = \chi - \bar{\Pi}_i\omega.$$

According to (22), the system (6) is rewritten as

$$\begin{aligned} &\dot{\tilde{\chi}} = \bar{A}_{\sigma(t)}\tilde{\chi}, \\ &\dot{w} = Sw, \\ &e = \bar{C}_{\sigma(t)}\tilde{\chi}. \end{aligned} \tag{23}$$

For the switched linear system (23), we choose the Lyapunov function as follows

$$V(t) = \tilde{\chi}^T(t)P\tilde{\chi}(t).$$

Similar to the proof of Theorem 1, we can show that the problem of output regulation for the switched linear (6) is solved under the switching law (20).

## II. Conclusion

In this paper, the output regulation problem for switched linear systems is investigated. The problem is difficult for switched systems, therefore, the results are few. The different coordinate transformations for switched systems is always a complicated issue. Because once the coordinate transformation is different, the Lyapunov function is discontinuous, and the decreasing of the Lyapunov function is difficult to verify. For switched linear systems, the problem can be solved by common coordinate transformation, but the conservativeness is always relatively large.

In this work, we address the output regulation problem with the transformations of switching instant between the consecutive time satisfying a relational expressions, based on this relation, we solve the output regulation problem. We easily know that when the disturbance and the state of the switched systems are linearly dependent, the relational expressions given in this paper is automatically satisfied. By virtue of this, we can get that the conditions are easy to implement.

The main contribution in this paper is that the output regulation problem for the switched linear systems with different coordinate transformations is considered. Sufficient conditions for the problem to be solvable are given based on the average dwell time method. For the regulation equations of the switched linear system, we suppose that each subsystem has its own solution, which reduce the conservativeness.

Output regulation for switched systems is a complicated and important problem. Further, studying a robust output regulation problem for switched linear systems and switched nonlinear systems is another task for future work.

## ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grants 61174073 and 90816028, and Dogus University Fund for Science.

## REFERENCES

[1] D. Liberzon and A. S. Morse, 1999. ”Basic problems in stability and design of switched systems”. *IEEE Control Systems Magazine*, 19, pp. 59-70.

[2] L. G. Wu, Daniel W.C. Ho and C.W. Li, 2011. ”Sliding mode control of switched hybrid systems with stochastic perturbation”. *Systems & Control Letters*, 60(8), pp. 531-539.

[3] H. Yang, M. Staroswiecki, B. Jiang and J. Y. Liu, 2011. ”Fault tolerant cooperative control for a class of nonlinear multi-agent systems”. *Systems & Control Letters*, 60(4), pp. 271-277.

[4] S. M. Williams and R. G. Hoft, 1991. ”Adaptive frequency domain control of PWM switched power line conditioner”. *IEEE Trans. on Power Electronics*, 6, pp. 665-670.

[5] J. M. Eklund, J. Sprinkle, and S. S. Sastry, 2011. ”Switched and symmetric pursuit/evasion games using online model redictive control with application to autonomous aircraft”. *IEEE Trans. Control Systems Technology,* pp. 1-17.

[6] D. Cheng, 2004. ”Stabilization of planar switched systems”. *Systems & Control Letters*, 51, pp. 79-88.

[7] J. Zhao and G. M. Dimirovski, 2004. ”Quadratic stability of a class of switched nonlinear systems”. *IEEE Trans. Auto. Contr.*, 49, pp. 574-578.

[8] J. Daafouz, P. Riedinger and C. Jung, 2002. ”Stability analysis and control synthesis for switched systems: A switched lyapunov function approach”. *IEEE Trans. Auto. Contr.*, 47, pp. 1883-1887.

[9] J. L. Mancilla-Aguilar, 2000. ”A condition for the stability of switched nonlinear systems”. *IEEE Trans. Auto. Contr.*, 45, pp. 2077-2079.

[10] E. J. Davison, 1976. ”The robust control of a servomechanism problem for linear time-invariant multivariable systems”. *IEEE Trans. Auto. Contr.,* 21, pp. 25-34.

[11] B. A. Francis, 1977. ”The linear multivariable regulator problem”. *SIAM J. Control Optim.*, 15, pp. 486-505.

[12] B. A. Francis and W. M. Wonham, 1976. ”The internal model principle of control theory”. *Automatica,* 12, pp. 457-465.

[13] A. Serrani and A. Isidori, 2001. ”Semiglobal nonlinear output regulation with adaptive internal model”. *IEEE Trans. Auto. Contr.*, 46, pp. 1178-1194.

[14] H. K. Khalil, 2000. ”On the design of robust servomechanisms for minimum phase nonlinear systems”. *Int. J. Robust and Nonlinear Control*, 10, pp. 339-391.

[15] S. Seshagiri and H. K. Khalil, 2005. ”Robust output regulation of minimum phase nonlinear systems using conditional servocompensator”. *Int. J. Robust and Nonlinear Control*, 15, pp. 83-102.

[16] Y. Z. Liu and J. Zhao, 2001. ”Problem on output regulation via error feedback for a class of switched linear systems with disturbances”. *Control and Decision,* 16(Supp1), pp. 815-817.(In Chinese).

[17] Y. Z. Liu and J. Zhao, 2001. ”Output regulation of a class of switched linear systems with disturbances”. Proc. American Control Conf., Arlington, 2, pp. 882-883.

[18] Z. Y. Song, H. Nie and J. Zhao, 2006. ”Output regulation of linear discrete-time switched systems”. *Control and Decision*, 21, pp. 1249-1253.(In Chinese).

[19] J.W. Lee, and P. P. Khargonekar, 2008. ”Optimal output regulation for discrete-time switched and Markovian jump linear systems”, *SIAM J. Control Optim.*, 47(1), pp. 40-72.

[20] J.W. Lee, G. E. Dullerud, and P. P. Khargonekar, 2007. ”An Output Regulation Problem for Switched Linear Systems in Discrete Time”. *Proc. 46th IEEE Conf. on Decision and Control*, New Orleans, pp. 4993 - 4998.

[21] A. Isidori, Nonlinear control systems, 3-rd edition, Berlin: Springer-Verlag, 1995, pp. 391-402.

[22] H. W. Konbloch, A. Isidori and D. Flockerzi, Topics in Control Theory, Berlin: Birkh$\ddot{a}$user-Verlag, 1993.

[23] J. P. Hespanha and A. S. Morse, 1999. ”Stability of switched systems with average dwell-time”. *Proc. 38th IEEE Conf. on Decision and Control*, Phoenix, AZ, pp. 2655 - 2660.

[24] J. Q. Lu, Daniel W. C. Ho and J. D. Cao, 2010. ”A unified synchronization criterion for impulsive dynamical networks”, *Automatica,* 46, pp. 1215-1221.

# Design of Switching Adaptive Laws for the State Tracking Problem

Caiyun Wu
State Key Laboratory of Synthetical
Automation for Process Industries,
Northeastern University
Shenyang, 110819, China
wucaiyun123@126.com.

Georgi M. Dimirovski
School of Engineering, Dogus
University, TR-34722 Istanbul, R.
Turkey
School FEIT, SS Cyril & Methodius
University, MK-1000, Skopje, R,
Macedonia
gdimirovski@dogus.edu.tr

Jun Zhao
State Key Laboratory of Synthetical
Automation for Process Industries,
Northeastern University
Shenyang, 110819, China
zhaojun@ise.edu.cn

*Abstract*— **In this paper, we address the issue of the state tracking control for model reference adaptive control systems. For a given plant with a known structure and unknown parameters, this problem can be solved by designing an adaptive law for traditional model reference adaptive control designs. In this paper, for a plant with finite fixed adaptive laws where none of them guarantees the states of plant track those of the reference model, there are not allowed to design other adaptive law. In this case, we formulate a switching mechanism between these adaptive laws to track the reference model. A sufficient condition is given for the problem to be solvable via the convex combination technique, and a switching law is designed. The theoretical derivations are illustrated by means of an example.**

***Keywords- MRAC; adaptive control; switching law; state tracking; convex combination***

## I. INTRODUCTION

Model reference adaptive control (MRAC) is one of the main approaches in adaptive control. For MRAC, an adaptive controller, which has adjustable parameters and the same structure as the ideal controller, is usually given first. Then, an adaptive law is designed to adjust the parameters of the controller such that it can approach the ideal one to realize the state tracking. Differing from conventional non-adaptive controllers, the adaptive mechanism can improve steady accuracy and transient performance when the parameters of the system are unknown[1]. Several adaptive law design methods, including Lyapunov theory[2, 3], hyperstability theory[4, 5] and dissipative theory[6, 7], ensure that the boundedness of all the signals and state error asymptotically approaching zero

In general, the adaptive law need to be redesigned once the system changes. However, in practice, owing to the restrictions of hardwires and environment, it is difficult to redesign or modify the adaptive law when it is in use in the system. On the other hand, a single adaptive law, though works theoretically, is sometimes too complicated to implement in reality. In both cases, it is necessary to use multiple pre-given fixed adaptive laws to achieve state tracking.

The switched control strategy is very important. This is mainly due to the following reasons. Firstly, in some cases, a single controller (continuous or discrete) in a conventional control system is usually too complicated for sensors and actuators to realize. Secondly, a switched controller may stabilize the system when none of the controllers can stabilize the system alone [8-10]. At present, as a kind of hybrid control, the switched control strategy has been applied to automatic vehicles [11, 12], robot manipulators [13] and traffic system[14, 15], etc.

For adaptive control systems, the problem of stability with rapid variation of parameters can be solved by a switching strategy. At present, switched adaptive control is mainly focused on the multiple models adaptive control in which the transient performance is enhanced by a switching strategy of multiple adaptive controllers with different initial values. Though the system is non-switched, the closed-loop system of multiple model adaptive control is a switched system[16-19] . Another newly-arisen issue in switching adaptive control is the adaptive control problem of a switched system in which individual adaptive controllers are designed for the subsystems[20-23]. In these issues, a single adaptive controller is effective if the switching signal is fixed. How to design a switching law for given ineffective adaptive controllers to stabilize the error system, to the best of the authors' knowledge, has not been addressed in the existing literatures, which partly motivates our present work.

This paper studies the state asymptotically tracking problem of MRAC. For a system with finite fixed adaptive laws, none of which can guarantee state tracking, a sufficient condition is given to design a switching law between these adaptive laws via the convex combination

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

technique. State tracking is achieved for the closed-loop system.

The result of this paper has two features. First, unlike conventional MRAC adaptive law design, we design a switching law that orchestrates finite fixed adaptive laws to solve the state tracking problem. Additionally, the proposed method can deal with the case where the adaptive law must be designed with new system, which enlarges the applicability of adaptive control theory.

The rest of this paper is organized as follows. The state tracking MRAC problem is formulated in section II. In Section III, a switching strategy is proposed for the adaptive laws to solve the state tracking problem. Section IV gives a simulation to show the effectiveness of the proposed approach and Section V concludes the paper.

## II. Problem Statement

Consider a system

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \tag{1}$$

where $A(t) \in R^{n\times n}$ and $B(t) \in R^{n\times m}$ are the system matrix and input matrix, respectively, both of which have known structure and unknown parameters; $x(t) \in R^{n}$ is the system state; and $u(t) \in R^{m}$ is the control input.

The control objective is that the state $x(t)$ of the system (1) tracks the state $x_m(t) \in R^n$ of a reference model specified by the LTI system

$$\dot{x}_m(t) = A_m x_m(t) + B_m r(t), \tag{2}$$

where $A_m \in R^{n\times n}$ is a constant Hurwitz matrix, $B_m(t) \in R^{n\times m}$ is a constant input matrix, $x_m(t)$ is the system state of the reference model, and $r(t) \in R^m$ is the bounded reference input. The reference model and the input $r$ are chosen so that $x_m(t)$ represents a desired trajectory that $x(t)$ has to follow.

In order to achieve state tracking, we introduce a state error vector

$$e = x_m - x. \tag{3}$$

However, as well-understood, the parameters of the system (1) are difficult to be adjusted directly. In order to solve the state tracking problem, we introduce an adjustable feedback compensation matrix $F(t) \in R^{m\times n}$ and an adjustable feedforward gain matrix $K(t) \in R^{m\times m}$. Thus we apply a control law as follows

$$u = K(t)r + F(t)x. \tag{4}$$

Combining (4) with (1), we get the closed-loop system

$$\dot{x} = [A(t) + B(t)F(t)]x + B(t)K(t)r. \tag{5}$$

Furthermore, from (3) and (5), we can get the dynamical equation of the state error for the closed-loop system

$$\dot{e} = A_m e + [A_m - A(t) - B(t)F(t)]x + [B_m - B(t)K(t)]r. \tag{6}$$

When the dynamic response of system (6) is identical to that of the reference model (2) under the input $r(t)$, the system (1) is matched with reference model (2), that is,

$$\begin{aligned} A_m &= A(t) + B(t)F^*, \\ B_m &= B(t)K^*, \end{aligned} \tag{7}$$

where $F^*, K^*$ denote the ideal values of $F(t), K(t)$ when the system matches with the reference model, and $F(t)$, $K(t)$ are the estimate of $F^*(t)$, $K^*(t)$, respectively.

In conventional MRAC system, the adaptive law is expressed as [24]

$$\begin{aligned} \dot{F} &= R_F^{-1}(B_m K^{*-1})^T Pex^T, \\ \dot{K} &= R_K^{-1}(B_m K^{*-1})^T Per^T, \end{aligned} \tag{8}$$

where $R_F, R_K$ and $P$ are the positive definite symmetric matrices.

Then, for the closed-loop system, the dynamical equation of the state error is given by

$$\dot{e} = A_m e + B_m K^{*-1}\tilde{F}x + B_m K^{*-1}\tilde{K}r, \tag{9}$$

where $\tilde{F} = F^* - F$ and $\tilde{K} = K^* - K$ are the parameter errors.

The following problem is considered in this paper. Several fixed adaptive laws are offered and none of them satisfies the matching condition (7), meanwhile, it is not allowed to design any other adaptive law. How to design a switching law between these adaptive laws to achieve state tracking, that is, the adaptive laws are given by

$$\begin{aligned} \dot{F}(t) &= \Gamma_i ex^T, \\ \dot{K}(t) &= \Phi_i er^T, \end{aligned} \quad i = 1, 2, ..., N, \tag{10}$$

where $\Gamma_i \in R^{m\times n}$ and $\Phi_i \in R^{m\times n}$ are the adaptive adjustable matrices, but none of them satisfies the matching condition (7). Again, the parameter error are $\tilde{F} = F^* - F$ and $\tilde{K} = K^* - K$ with $F$ and $K$ given by the $i$-th adaptive (10). In order to make the state of the closed-loop system (5) track the state of the reference model (2), that is $e \to 0$, $t \to \infty$, we design the switching law for the error system (9) and (10).

## III. Main Result

In this section, we will design a switching law for the adaptive laws by means of the convex combination technology.

*Theorem 1.* If there exist scalars $\alpha_i \in (0,1), (i = 1, \cdots N)$ satisfying $\sum_{i=1}^{N} \alpha_i = 1$ and some positive definite symmetric matrices $R_F, R_K$ and $P$ satisfying

$$\begin{aligned} &\sum_{i=1}^{N} \alpha_i \Gamma_i = -R_F^{-1}(B_m K^{*-1})^T P, \\ &\sum_{i=1}^{N} \alpha_i \Phi_i = -R_K^{-1}(B_m K^{*-1})^T P. \end{aligned} \tag{11}$$

Then, there exists a switching law such that the adaptive controllers (4) and (10) make the state $x(t)$ of the closed-loop system (1) asymptotically track the state $x_m(t)$ of a reference model (2).

*Proof.* Based on the error system (9) and (10), we construct the combined error system

$$\begin{aligned} &\dot{e} = A_m e + B_m K^{*-1} \tilde{F} x + B_m K^{*-1} \tilde{K} r, \\ &\dot{F} = \sum_{i=1}^{N} (\alpha_i \Gamma_i e x^T), \\ &\dot{K} = \sum_{i=1}^{N} (\alpha_i \Phi_i e r^T), \end{aligned} \tag{12}$$

where (12) is the convex combination of the adaptive control law (10).

Consider the following Lyapunov function candidate

$$V = \frac{1}{2}[e^T P e + tr(\tilde{F}^T R_F \tilde{F} + \tilde{K}^T R_K \tilde{K})]. \tag{13}$$

The time derivative of the Lyapunov function (13) along the trajectory of the combined error system (12) is

$$\begin{aligned} \dot{V} &= \frac{1}{2}\Big\{\dot{e}^T P e + e^T P \dot{e} \\ &+ tr[\sum_{i=1}^{N} (\alpha_i \Gamma_i e x^T)^T R_F \tilde{F} + \tilde{F}^T R_F \sum_{i=1}^{N} (\alpha_i \Gamma_i e x^T) \\ &+ \sum_{i=1}^{N} (\alpha_i \Phi_i e r^T)^T R_K \tilde{K} + \tilde{K}^T R_K \sum_{i=1}^{N} (\alpha_i \Phi_i e r^T)]\Big\} \\ &= \frac{1}{2}[e^T (P A_m + A_m^T P) e] + e^T P B_m K^{*-1} \tilde{F} x + e^T P B_m K^{*-1} \tilde{K} r \\ &+ \frac{1}{2} tr[\sum_{i=1}^{N} (\alpha_i \Gamma_i e x^T)^T R_F \tilde{F} + \tilde{F}^T R_F \sum_{i=1}^{N} (\alpha_i \Gamma_i e x^T) \\ &+ \sum_{i=1}^{N} (\alpha_i \Phi_i e r^T)^T R_K \tilde{K} + \tilde{K}^T R_K \sum_{i=1}^{N} (\alpha_i \Phi_i e r^T)]. \end{aligned} \tag{14}$$

With the help of

$$\begin{aligned} &e^T P B_m K^{*-1} \tilde{F} x = tr(x e^T P B_m K^{*-1} \tilde{F}), \\ &e^T P B_m K^{*-1} \tilde{K} r = tr(r e^T P B_m K^{*-1} \tilde{K}), \end{aligned} \tag{15}$$

(14) can be rewritten as

$$\begin{aligned} \dot{V} &= \frac{1}{2}[e^T (P A_m + A_m^T P) e] \\ &+ tr[\sum_{i=1}^{N} (\alpha_i \Gamma_i e x^T)^T R_F \tilde{F} + x e^T P B_m K^{*-1} \tilde{F}] \\ &+ tr[\sum_{i=1}^{N} (\alpha_i \Phi_i e r^T)^T R_K \tilde{K} + r e^T P B_m K^{*-1} \tilde{K}] \\ &= \frac{1}{2}[\sum_{i=1}^{N} \alpha_i e^T (P A_m + A_m^T P) e] \\ &+ tr[\sum_{i=1}^{N} (\alpha_i \Gamma_i e x^T)^T R_F \tilde{F} + x e^T P B_m K^{*-1} \tilde{F}] \\ &+ tr[\sum_{i=1}^{N} (\alpha_i \Phi_i e r^T)^T R_K \tilde{K} + r e^T P B_m K^{*-1} \tilde{K}]. \end{aligned} \tag{16}$$

According to (11), for the combined error system (12), we get $\dot{V} < 0$ ,that is,

$$\begin{aligned} \dot{V} &= \sum_{i=1}^{N} \alpha_i \{\frac{1}{2}[e^T (P A_m + A_m^T P) e] \\ &+ tr\left[(\Gamma_i e x^T)^T R_F \tilde{F} + x e^T P B_m K^{*-1} \tilde{F}\right] \\ &+ tr[(\Phi_i e r^T)^T R_K \tilde{K} + r e^T P B_m K^{*-1} \tilde{K}]\} \\ &< 0. \end{aligned} \tag{17}$$

Because $\alpha_i > 0$ and $\sum_{i=1}^{N} \alpha_i = 1$, there exists $l \in N$ at least for (17) such that

$$\begin{aligned} &\frac{1}{2}[e^T (P A_m + A_m^T P) e] \\ &+ tr\left[(\Gamma_l e x^T)^T R_F \tilde{F} + x e^T P B_m K^{*-1} \tilde{F}\right] \\ &+ tr[(\Phi_l e r^T)^T R_K \tilde{K} + r e^T P B_m K^{*-1} \tilde{K}] \\ &< 0. \end{aligned} \tag{18}$$

Now, we turn to the error system (9) and (10).

Consider the following Lyapunov function candidate

$$V = \frac{1}{2}[e^T P e + tr(\tilde{F}^T R_F \tilde{F} + \tilde{K}^T R_K \tilde{K})]. \tag{19}$$

When the $j$ -th adaptive control law is active, the time derivative of the Lyapunov function (19) along the trajectory of the error system (9) and (10) is

$$\begin{aligned} \dot{V} &= \frac{1}{2}[e^T (P A_m + A_m^T P) e] \\ &+ tr\left[(\Gamma_j e x^T)^T R_F \tilde{F} + x e^T P B_m K^{*-1} \tilde{F}\right] \\ &+ tr[(\Phi_j e r^T)^T R_K \tilde{K} + r e^T P B_m K^{*-1} \tilde{K}]. \end{aligned} \tag{20}$$

Therefore, for the error system (9) and (10), the switching law can be designed as

$$\begin{aligned}\sigma(e,t) = \arg\min_i\{&\frac{1}{2}[e^T(PA_m + A_m^T P)e]\\ &+tr[(\Gamma_i ex^T)^T R_F \tilde{F} + xe^T PB_m K^{*-1}\tilde{F}] \\ &+tr[(\Phi_i er^T)^T R_K \tilde{K} + re^T PB_m K^{*-1}\tilde{K}]\}.\end{aligned} \qquad (21)$$

Owing to (18) and (21), adaptive laws (10) is orchestrated so that $V$ decreases along the solutions of the error system (9) and (10). Then, the state $x(t)$ of the system (1) asymptotically track the state $x_m(t)$ of the reference model (2) with the given controller (4) and (10).

*Remark 1.* For the given adaptive law (10), if there exists $h \in N$, such that the $h$-th adaptive control law satisfying (8), the state tracking problem can be solved by the $h$-th adaptive control law. In this case, we choose $\alpha_h = 1$, $\alpha_i = 0,\ i \neq h,$ therefore, Theorem 1 contains the result of [24] as a special case.

*Remark 2.* In the case of $N = 1$, the issue is degenerated into the design of a single controller.

*Remark 3.* The switching law is designed when $\tilde{F}$ and $\tilde{K}$ are available. If $\tilde{F}$ and $\tilde{K}$ they are unavailable, an estimator can be used instead[25] .

## IV. EXAMPLE

In order to show the effectiveness of the proposed switching adaptive controllers, we consider an example.

Consider the following system

$$\dot{x} = \begin{pmatrix} 0 & 1 \\ -6 & -7 \end{pmatrix} x + \begin{pmatrix} 0 \\ 8 \end{pmatrix} u.$$

The reference model is given by

$$\dot{x}_m = \begin{pmatrix} 0 & 1 \\ -10 & -5 \end{pmatrix} x_m + \begin{pmatrix} 0 \\ 2 \end{pmatrix} r.$$

We compare the simulation results of the system under two given adaptive laws none of which satisfies the state tracking respectively to those under the designed switching law.

For controller (4), there are two adaptive laws as follows

$$\left.\begin{aligned}\dot{F}(t) &= \Gamma_i ex^T, \\ \dot{K}(t) &= \Phi_i er^T,\end{aligned}\right. \quad i = 1,2, \qquad (22)$$

where

$$\Gamma_1 = [49.4 \quad 49.4], \Phi_1 = [4 \quad 4],$$

$$\Gamma_2 = [0.15 \quad 0.15], \Phi_2 = [1.5 \quad 1.5].$$

Simulations are performed for these adaptive laws, respectively.

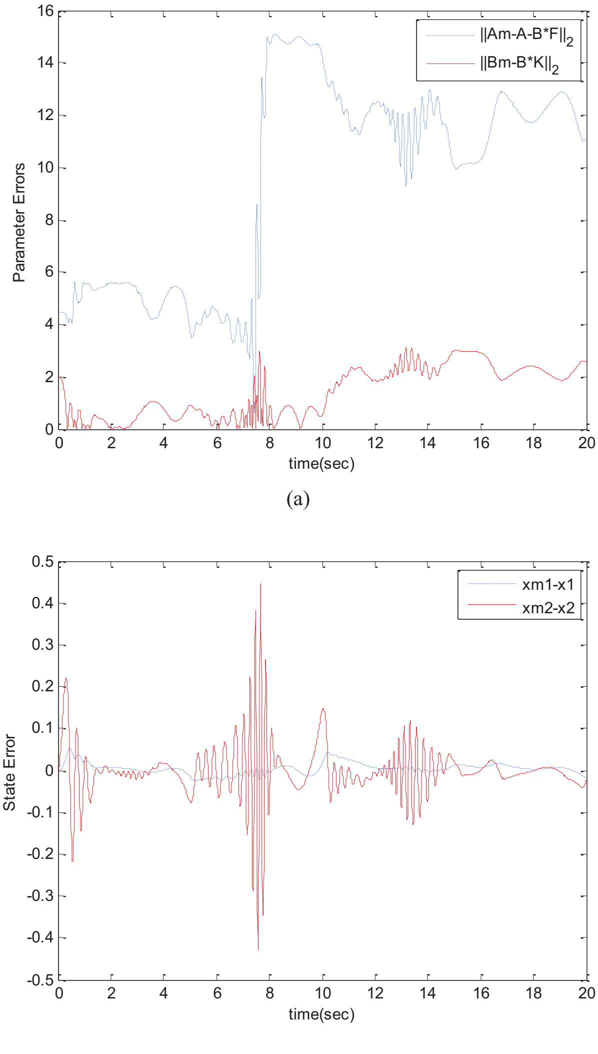
For $B_m = B(t)K^*$, we have

$$B(t) = B_m K^{*-1} = \begin{pmatrix} 0 \\ 8 \end{pmatrix}.$$

Choose $N = 2, P = \begin{pmatrix} 3 & 1 \\ 1 & 1 \end{pmatrix}.$

The parameter errors and the state error are shown in Figure 1 and Figure 2.

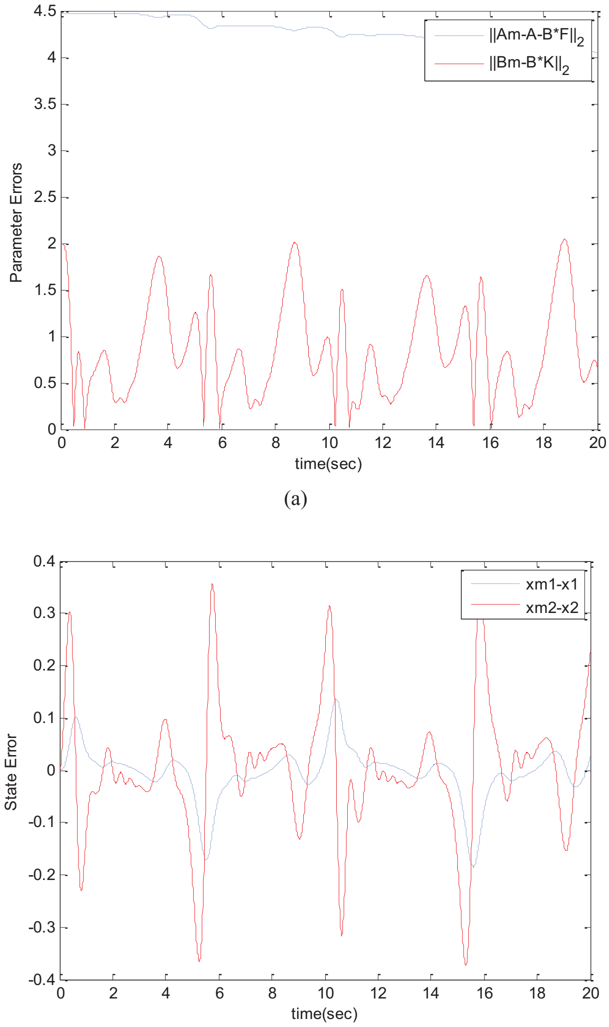
Obviously, Figure 1 and Figure 2 show that none of these adaptive laws makes the parameter errors and the state error convergent.



(a)

(b)

Figure 1. The simulation result ( $i = 1$ ).
(a): The parameter errors. (b): The state error



(a)

(b)

Figure 2. The simulation result ( $i = 2$ ).

(a): The parameter errors. (b): The state error

In order to achieve the objective, we will design a switching law by Theorem 1

With the help of (4) and (22), we obtain switching adaptive laws

$$\begin{aligned}\dot{F}(t) &= \Gamma_\sigma e x^T, \\ \dot{K}(t) &= \Phi_\sigma e r^T,\end{aligned} \quad \sigma = 1, 2,$$

where the switching law $\sigma$ is chosen by (21) as Figure 3.

Here, we choose $\alpha_1 = 0.2$, $\alpha_2 = 0.8$, $R_F = 1.25$, $R_K = 0.25$, such that (11) hold. From Figure 4, it is easy to see that the parameter errors and the state error are convergent under the switching law. Then, state tracking is achieved.

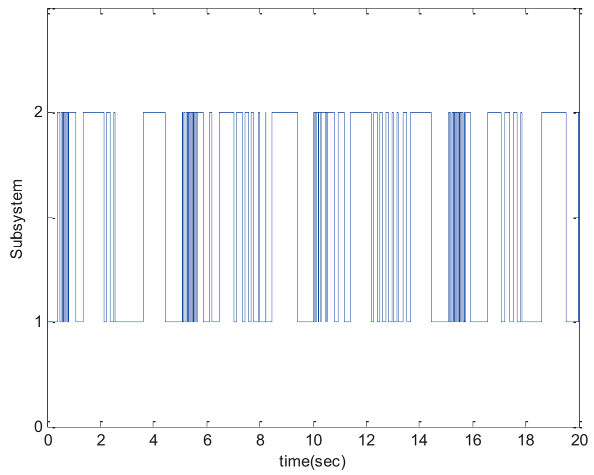
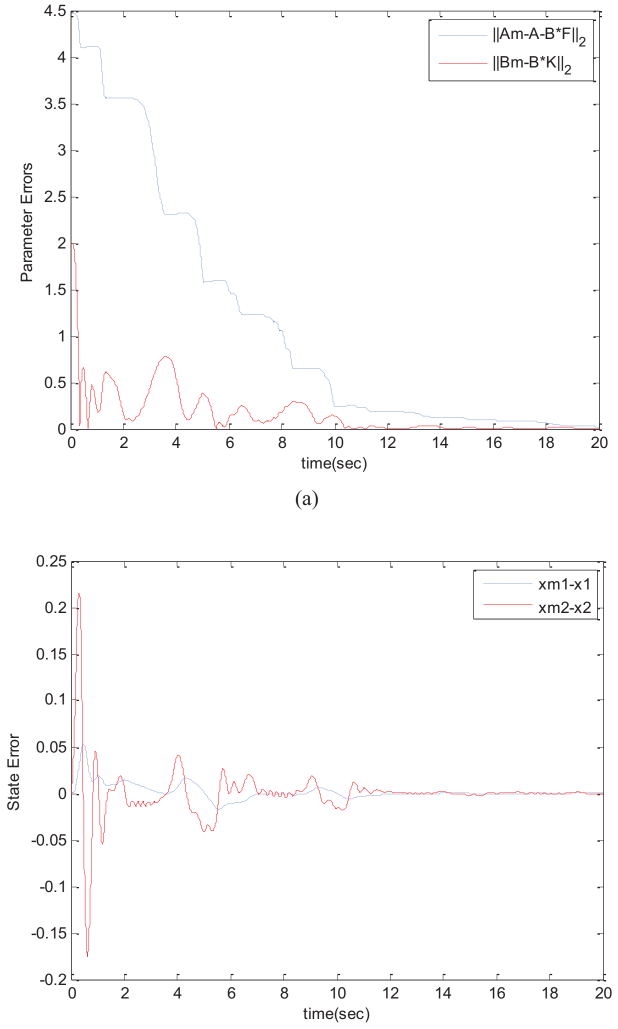


Figure 3. The switching signal



(a)

(b)

Figure 4. The simulation results under switching law.
(a): The parameter errors. (b): The state error

## V. Conclusions

We have studied the state asymptotically tracking control problem for model reference adaptive control systems. For a system with several fixed adaptive laws, a sufficient condition has been developed to solve the state tracking problem via the convex combination technique by designing the controller switching strategy. A simulation example has been given to show the effectiveness of the proposed method.

## Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grants 61174073 and 90816028.

## References

[1] H. Liu, *et al.*, "New results on output-feedback variable structure model-reference adaptive control: design and stability analysis," *IEEE Transactions on Automatic Control,* vol. 42, pp. 386-393, 1997.

[2] M. Jianqing and X. Zibin, "Adaptive control for a class of mismatched uncertain system," in *International Workshop on Intelligent Systems and Applications*, 2009, pp. 1-4.

[3] Z. Ding, "Adaptive control of triangular systems with nonlinear parameterization," *IEEE Transactions on Automatic Control,* vol. 46, pp. 1963-1968, 2001.

[4] G. Yang and T. H. Chin, "Adaptive-speed identification scheme for a vector-controlled speed sensorless inverter-induction motor drive," *IEEE Transactions on Industry Applications,* vol. 29, pp. 820-825, 1993.

[5] I. Landau, "A hyperstability criterion for model reference adaptive control systems," *IEEE Transactions on Automatic Control,* vol. 14, pp. 552-555, 1969.

[6] B. Brogliato, *et al.*, *Dissipative Systems Analysis and Control—Theory and Applications*, 2006.

[7] G. E. Valderrama, *et al.*, "Dissipativity-based adaptive and robust control of UPS in unbalanced operation," *IEEE Transactions on Power Electronics,* vol. 18, pp. 1056-1062, 2003.

[8] A. Y. Pogromsky, *et al.*, "On stability and passivity of a class of hybrid systems," in *IEEE Conference on Decision and Control*, 1998, pp. 3705-3710.

[9] E. Skafidas, *et al.*, "Quadratic stabilizability of state feedback hybrid control systems," *ICARCV'96,* pp. 1073-1077, 1996.

[10] A. V. Savkin, *et al.*, "Robust output feedback stabilizability via controller switching," *Automatica,* vol. 35, pp. 69-74, 1999.

[11] J. Lygeros, *et al.*, "Verified hybrid controllers for automated vehicles," *IEEE Transactions on Automatic Control,* vol. 43, pp. 522-539, 1998.

[12] V. Sankaranarayanan, *et al.*, "A switched controller for an underactuated underwater vehicle," *Communications in Nonlinear Science and Numerical Simulation,* vol. 13, pp. 2266-2278, 2008.

[13] M. Egerstedt, "Behavior based robotics using hybrid automata," *Lecture notes in computer science,* vol. 1790, pp. 103-116, 2000.

[14] A. V. Savkin and R. J. Evans, *Hybrid dynamical systems: controller and sensor switching problems*: Birkhauser, 2002.

[15] C. Tomlin, *et al.*, "Conflict resolution for air traffic management: a study in multiagent hybrid systems," *IEEE Transactions on Automatic Control* vol. 43, pp. 509-521, 1998.

[16] K. S. Narendra and J. Balakrishnan, "Adaptive control using multiple models," *IEEE Transactions on Automatic Control,* vol. 42, pp. 171-187, 1997.

[17] M. Kuipers and P. Ioannou, "Multiple model adaptive control with mixing," *IEEE Transactions on Automatic Control,* vol. 55, pp. 1822-1836, 2010.

[18] B. Anderson, *et al.*, "Multiple model adaptive control with safe switching," *International Journal of Adaptive Control and Signal Processing,* vol. 15, pp. 445-470, 2001.

[19] J. Boskovic, "A multiple model-based controller for nonlinearly-parametrized plants," in *Proceedings of American Control Conference*, 1997, pp. 2140-2144 vol. 3.

[20] K. Mutoh and A. Sano, "Adaptive control of piecewise linear systems," in *Asian Control Conference*, 2009, pp. 27-32.

[21] Q. Sang and G. Tao, "Adaptive control of piecewise linear systems with applications to NASA GTM," in *American Control Conference*, 2011, pp. 1157-1162.

[22] Q. Sang and G. Tao, "Adaptive control of piecewise linear systems: the state tracking case," *IEEE Transactions on Automatic Control,* vol. 57, pp. 522-528, 2010.

[23] M. L. Chiang and L. C. Fu, "Adaptive control of switched systems with application to HVAC system," in *IEEE International Conference on Control Applications*, 2007, pp. 367-372.

[24] P. A. Ioannou and J. Sun, *Robust adaptive control*: Prentice Hall, 1996.

[25] F. M. Pait and F. Kassab Jr, "On a class of switched, robustly stable, adaptive systems," *International Journal of Adaptive Control and Signal Processing,* vol. 15, pp. 213-238, 2001.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# An Adaptive Observer-based parameter estimation algorithm with application to Road Gradient and Vehicle's Mass Estimation

Muhammad Nasiruddin Mahyuddin*, Jing Na**, Guido Herrmann***, Xuemei Ren**** and Phil Barber*****

***Abstract*—A novel observer-based parameter estimation algorithm with sliding mode term has been developed to estimate the road gradient and vehicle weight using only the vehicle's velocity and the driving torque from the engine. The estimation algorithm exploits all known terms in the system dynamics and a low pass filtered representation to derive an explicit expression of the parameter estimation error without measuring the acceleration. The proposed algorithm which features a sliding-mode term to ensure the fast and robust convergence of the estimation in the presence of persistent excitation is augmented to an adaptive observer and analyzed using Lyapunov Theory. The analytical results show that the algorithm is stable and ensures finite-time error convergence to a bounded error even in the presence of disturbances. A simple practical method for validating persistent excitation is provided using the new theoretical approach to estimation. This is validated by the practical implementation of the algorithm on a small-scaled vehicle, emulating a car system. The slope gradient as well as the vehicle's mass/weight are estimated online. The algorithm shows a significant improvement over a previous result.**

## I. INTRODUCTION

In the automotive industry, reliable online vehicle parameter estimation is important to reduce emissions, improve fuel efficiency and enhance the safety of the vehicle. The vehicle's mass and the road grade are two parameters that largely influence a vehicles performance. This is particularly true for heavy duty vehicles where the loadings due to the mass and the grade can be significant [1]. The road gradient and mass estimation provides useful information to a vehicle in improving the transmission shift scheduling and vehicle longitudinal control, including cruise control, hill holding and traction control [2].

Having road inclination measured by a dedicated sensor such as an inclinometer may be inaccurate. Inclinometers are in fact accelerometers which can distort road gradient measurement due to its susceptibility to noise in dynamic conditions of a vehicle [3]. Therefore, the road grade should be accurately estimated [4-7]. It is evident that the transmission control unit and the anti-lock brake system can benefit from mass and road gradient estimates [7]. To address this issue, there has been significant interest in the estimation of the road gradient and vehicle mass [2-9]. Some of these results were developed based on on-board sensors and the use of sensor/data fusion methods. For instance, in [1], a GPS or barometer sensor is utilized in addition to torque and velocity sensors to obtain absolute road height information, while Barrho, et al. in [9] require accurate information of the vehicle mass which is not always possible. In recent work by [10], [11] and [12], low-cost sensors were placed in the vehicle to estimate the road grade ahead. In [13], the vehicle's mass in addition to vehicle speed and torque information is required. Bae et al. [4] suggest a recursive least squares approach which essentially requires acceleration information and then assumes the existence of sufficient data points to solve for the missing parameters, i.e. vehicle mass and gradient, by inverting a regressor matrix in a batch process. Similarly, the work in [7] and [2] estimate the road grade using the position, velocity and driving torque or force signal. In [8], a combination of an observer is suggested which provides for known mass the exact estimation of the road gradient. In all of the aforementioned approaches, however, some of the required information, e.g. acceleration, mass and vehicle's current location through GPS, may not be readily available. Although most of the approaches show good results, the convergence speed and complexity may cast some problems.

In this paper, we revisit the online road gradient and mass estimation of vehicular systems using only the vehicle's velocity and the driving torque. This is achieved based on a novel adaptive nonlinear observer design. Compared to previous results (e.g. [14]) concerning the parameter estimation, some appropriate information of the parameter error is derived, and then incorporated into the parameter adaptation for the observer design. In our work, the parameter estimation scheme uses a filtered regressor matrix. Measurable system states, a regressor vector and the known dynamics are collected and filtered to form auxiliary variables. Moreover, vehicle acceleration is not required in our estimation algorithm. Owed to the special feature of a sliding mode term, the adaptation algorithm guarantees robust finite-time convergence to a compact set, provided that there is a Persistent Excitation (PE) condition fulfilled so that the regressor matrix remains positive definite. In contrast to [15], our scheme calculates the inverse of the filtered and integrated regressor matrix without prior invertibility checking of the matrix and direct matrix inverse computation. The parameter error information can be explicitly formulated by virtue of the filtered auxiliary variables. The possible instability and infinite growth found in [15] due to the existence of an unstable integrator (as a result of auxiliary matrix and vector) are prevented in this paper. We also show robustness of our adaptive scheme and *we can verify* the PE condition in a straightforward and practical manner. The proposed method is verified experimentally in a reduced-scale vehicular system, which provide a significant improvement over a previous algorithm.

## II. SYSTEM FORMULATIONS

Consider a nonlinear system of the following structure:

$$\dot{x} = Ax + B_1 u_1 + B_2 f(x, u_2) + \zeta, \quad y = Cx \tag{1}$$

This work is supported by a joint grant between the Royal Society UK and National Natural Science Foundation of China under Grant No. 61011130163/JP090823, and the National Natural Science Foundation of China (No. 60974046). The authors are also particularly thankful to dSPACE Inc. for their provision of the MicroAutobox.

*Muhammad Nasiruddin Mahyuddin is currently a PhD student in the Department of Mechanical Engineering, University of Bristol, BS8 1TR, UK. and being sponsored by University Sains Malaysia. Email: memnm@bristol.ac.uk or nasiruddin@ieee.org

**Jing Na is with Plant Engineering Division, ITER Organization, St Paul Lez Durance, 13115, France. Email: najing25@163.com

***Guido Herrmann is a Senior Lecturer with the Department of Mechanical Engineering, University of Bristol, BS8 1TR, UK. Email: g.herrmann@bristol.ac.uk

****Xuemei Ren is a Professor at School of Automation, Beijing Institute of Technology, Beijing, 100081,China, E-mail: xmren@bit.edu.cn

*****Phil Barber is with Jaguar and Land Rover Research, W/2/021 Engineering Centre, Abbey Road, Whitley, Coventry CV 4LF, UK. E-mail: pbarber2@jaguarlandrover.com

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

where $A \in \mathbb{R}^{n\times n}$ is the known system matrix, $B_1 \in \mathbb{R}^{n\times m_1}$ and $B_2 \in \mathbb{R}^{n\times m_2}$ are known input matrices, $u_1 \in \mathbb{R}^{m_1}$ and $u_2 \in \mathbb{R}^{\bar{m}_2}$ are known inputs, whilst $C \in \mathbb{R}^{p\times n}$ is the corresponding output matrix and $\zeta \in L_\infty$ is bounded disturbance. The function $f(x, u_2) : \mathbb{R}^n \times \mathbb{R}^{\bar{m}_2} \to \mathbb{R}^{m_2}$ is partially unknown for which the detail will be outlined below and the pair $(A, B_1)$ is controllable. It is assumed that $p \geq m_2$. The following assumptions are made:

**Assumption 1** $(C, A, B_2)$ is minimum phase and $(CB_2)$ is full rank.

**Assumption 2** The function $f(x, u_2)$ can be represented in a linear parameterized form: $f(x, u_2) = \varphi(x, u_2)\Theta$, where $\varphi : \mathbb{R}^n \times \mathbb{R}^{m_2} \to \mathbb{R}^{m_2\times l}$ is a known Lipschitz continuous function, while $\Theta = const., \Theta \in \mathbb{R}^l$ is an unknown parameter vector which is to be estimated.

**Assumption 3** The signals $x$, $u_1$ and $u_2$ are measurable and bounded.

Assumption 3 is a common assumption for observer design and can be easily achieved by suitable choice of the control signal $u_1$ (e.g. [15] [17]).

Under these conditions, the system is assumed to take the following structure

$$\begin{aligned} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} &= \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \\ & \begin{bmatrix} B_{11} \\ B_{12} \end{bmatrix} u_1 + \begin{bmatrix} 0 \\ \bar{B}_2 \end{bmatrix} \varphi\Theta + \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix} \\ y &= [0 \quad I] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \end{aligned} \tag{2}$$

where $\bar{B}_2 \in \mathbb{R}^{p\times m_2}$, $I \in \mathbb{R}^{p\times p}$ and $x_2 = Cx$. Note that this reformulation is always possible from Assumptions 1 and 3 using Proposition 6.3 in [17]. Moreover, we make the following assumption.

**Assumption 4** $A_{21} = 0$, the second state equation in 2 is decoupled.

Assumption 4 is possibly a strong assumption but it will fit the generic practical system structures (e.g. vehicular) investigated in this paper.

## III. Adaptive Observer Design

We will design an adaptive observer to estimate the state vectors which will be suitably combined with a novel parameter estimation algorithm. The adaptive observer takes the following form:

$$\dot{\hat{x}} = A\hat{x} + B_1u_1 + B_2\varphi\hat{\Theta} + L(y - C\hat{x}) \tag{3}$$

where $\hat{x}$ is the estimated state vector, $\hat{\Theta}$ is the estimated parameter vector. $L$ is the observer gain matrix such that $A_c = A - LC$ is a stable matrix and there exist, according to Proposition 6.3 in [17], positive definite matrices, $P$ and $Q$ so that,

$$A_c^T P + PA_c = -Q, \quad P = \begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix} > 0 \tag{4}$$

$$Q = \begin{bmatrix} Q_1 & Q_{12} \\ Q_{12}^T & Q_2 \end{bmatrix} > 0, \quad PB_2 = C^T F^T \tag{5}$$

and $F \in \mathbb{R}^{m_2\times p}$ is a positive definite matrix. From (4), it follows,

$$\begin{bmatrix} A_{c11}^T P_1 + P_1A_{c11} & P_1A_{c12} \\ A_{c12}^T P_1 & P_2A_{c22} + A_{c22}^T P_2 \end{bmatrix} = -Q \tag{6}$$

Let $\tilde{x} = x - \hat{x}$ and $\tilde{\Theta} = \Theta - \hat{\Theta}$. We can then use (1) and (3) to define the error dynamics $\tilde{x} = x - \hat{x}$ as

$$\begin{aligned} \dot{\tilde{x}} &= (A - LC)\tilde{x} + B_2\varphi\tilde{\Theta} + \zeta \\ &= A_c\tilde{x} + B_2\varphi\tilde{\Theta} + \zeta \end{aligned} \tag{7}$$

where $\tilde{\Theta} = \Theta - \hat{\Theta}$ is the estimated parameter error vector.

In the next section, the adaptive laws that will update the estimated parameter vector, $\hat{\Theta}$ will be developed.

## IV. Adaptive Law Formulation

In this section, we shall define the adaptive law for our parameter estimator.

### A. Filter design

From (2), the second state equation can be expressed as,

$$\dot{x}_2 = (A_{22}x_2 + B_{12}u_1) + \bar{B}_2\varphi\Theta + \zeta_2 \tag{8}$$

Let,

$$\psi = A_{22}x_2 + B_{12}u_1, \quad \phi = \bar{B}_2\varphi \tag{9}$$

then, the following filtered variables can be defined as,

$$\begin{aligned} k\dot{x}_{2f} + x_{2f} &= x_2, \quad x_{2f}(0) = 0 \\ k\dot{\psi}_f + \psi_f &= \psi, \quad \psi_f(0) = 0 \\ k\dot{\phi}_f + \phi_f &= \phi, \quad \phi_f(0) = 0 \end{aligned} \tag{10}$$

In addition, we may introduce an auxiliary filter for the bounded disturbance (which is only used for analysis),

$$k\dot{\zeta}_{2f} + \zeta_{2f} = \zeta_2, \quad \zeta_{2f}(0) = 0 \tag{11}$$

i.e. $\zeta_{2f} \in L_\infty$. Consequently, we can obtain from (8) and (10) that,

$$\dot{x}_{2f} = \frac{x_2 - x_{2f}}{k}, \qquad \frac{x_2 - x_{2f}}{k} - \psi_f = \phi_f\Theta + \zeta_{2f} \tag{12}$$

### B. Auxillary integrated regressor matrix and vector

The filtered variables introduced above will be used in the definition of a filtered regressor matrix, $M(t)$, and a vector, $N(t)$ as,

$$\dot{M}(t) = -k_{FF}M(t) + k_{FF}\phi_f^T(t)\phi_f(t), M(0) = 0 \tag{13}$$

$$\dot{N}(t) = -k_{FF}N(t) + k_{FF}\phi_f^T(t)\left(\tfrac{x_2 - x_{2f}}{k} - \psi_f\right), \tag{14}$$

where, $k_{FF} \in \mathbb{R}^+$, can be implemented as a forgetting factor and the initial condition of $N(t)$ is $N(0) = 0$. Note that (14) is equivalent to:

$$\dot{N}(t) = -k_{FF}N(t) + k_{FF}\phi_f^T(t)(\phi_f(t)\Theta + \zeta_{2f}) \tag{15}$$

Consequently, we can find the solution to (13), (14) and (19),

$$\begin{aligned} M(t) &= \textstyle\int_0^t e^{-k_{FF}(t-r)} k_{FF}\phi_f^T(r)\phi_f(r)dr \\ N(t) &= \textstyle\int_0^t e^{-k_{FF}(t-r)} k_{FF}\phi_f^T(r)\left(\tfrac{x_2 - x_{2f}}{k} - \psi_f\right)dr \end{aligned} \tag{16}$$

and

$$N(t) = M(t)\Theta + \zeta_{2N} \tag{17}$$

where $\zeta_{2N} = \int_0^t e^{-k_{FF}(t-r)} k_{FF}\phi_f^T(r)\zeta_{2f}dr$. Note that $\phi$ is bounded since it is Lipschitz continuous and $x, u_2$ are bounded (Assumption 1). Thus, $\phi_f$ is bounded. Since $\zeta_{2f} \in L_\infty$, it follows that $N(t), M(t)$ and $\zeta_{2N}$ are bounded.

*Lemma 1: The auxiliary regressor matrix $M(t) \in \mathbb{R}^{l\times l}$ is positive definite, $M(t) > 0$, if and only if $\int_0^t \phi_f^T \phi_f > 0$.* •

*Proof :* It can be easily shown that

$$\int_T^t \phi_f^T(r)\phi_f(r)dr \geq \int_T^t e^{-k_{FF}(t-r)}\phi_f^T(r)\phi_f(r)dr \tag{18}$$
$$\geq e^{-k_{FF}t}\int_T^t \phi_f^T(r)\phi_f(r)dr$$

when $T < t$. For $T = 0$, the claim follows. ■

Thus, if $\phi_f$ is persistently excited, $M(t) > 0$ is positive definite. Clearly, if $\phi$ is persistently excited, then $\phi_f$ is also persistently excited and $M(t) > 0$ [22], [20] (as derived from the linear system (10) and definition (9)). Thus, if $\phi$ is persistently excited then $M(t) > 0$ and $\int_T^t \phi_f^T\phi_f > 0$. In this paper, it is important to achieve $M(t) > 0$ for our adaptation algorithm to work. This can be achieved through persistent excitation of $\phi$:

*Remark 1: The Persistent Excitation (PE) condition for the regressor $\phi$ can be achieved in the experiment through an appropriate control signal, $u$. For instance, the control signal can be augmented by a noise signal or the controller can introduce for the system states, $x$, a tracking demand which achieves 'sufficient richness' (SR) of $x$ and guarantees $M(t) > \lambda_m I, \lambda_m > 0$ as in [19]. Suitable analytical detail is avoided here due to space reasons.* ○

Another auxillary matrix $K(t)$ may be defined as,

$$\dot{K}(t) = k_{FF}K(t) - k_{FF}K(t)\phi_f^T(t)\phi_f(t)K(t), \tag{19}$$

where the initial condition $K(0) > 0$ is specified as a diagonal matrix, $K(0) = \frac{1}{\lambda}I$ with $\lambda > 0$ being constant. It will be seen that $K(t)$ is an approximation of the inverse of $M(t)$ where $\lim\limits_{t\to\infty} K(t)M(t) = I$. With the help of the following derivative matrix identity,

$$\frac{d}{dt}KK^{-1} = K\tfrac{d}{dt}K^{-1} + K^{-1}\tfrac{d}{dt}K = 0 \tag{20}$$

we can obtain,

$$K(t) = [e^{-k_{FF}t}K^{-1}(0) + M(t)]^{-1} \tag{21}$$

This also implies boundedness of $K(t)$, if $M(t) > 0$. To show the invertibility of $K(t)$ as well as $K(t)M(t)$ approaches unity, we are to employ the singular value decomposition for the matrix, $M(t)$,

$$M(t) = U(t)S(t)V^T(t) \tag{22}$$

where $S(t) = diag(s_1, \dots, s_n)$ is the matrix with $s_i$ being the singular values of matrix, $M(t)$ whilst, $U(t)$ and $V(t)$ are unitary matrices.

We know that $K(0) = \frac{1}{\lambda}I$ is a diagonal matrix, thereby,

$$K(t) = V(t)(S(t) + e^{-k_{FF}t}\lambda I)^{-1}U^T(t) \tag{23}$$

Then,

$$K(t)M(t) = V(t)diag(\tfrac{s_1}{s_1+e^{-k_{FF}t}\lambda}, \dots, \tfrac{s_n}{s_n+e^{-k_{FF}t}\lambda})V^T(t) \tag{24}$$

and $\lim\limits_{t\to\infty} \dfrac{s_i}{s_i + e^{-k_{FF}t}\lambda} = 1, \quad if \quad M(t) \geq \lambda_m I > 0,$ for $\lambda > 0$, adhering to the Persistent Excitation (PE) condition [14]. Since $V(t)$ is unitary, the matrix $K(t)M(t)$ can be represented as,

$$K(t)M(t) = I - \Delta(t) \tag{25}$$

where $\Delta$ converges to zero in infinite time. This shows that $K(t)$ is indeed a representation of the inverse of $M(t)$ where $\Delta(t)$ denotes the effects of the initial condition $K(0)$. Hence, the parameter estimation error vector can be written as,

$$\tilde{\Theta} = \Theta - \hat{\Theta} = [K(t)M(t) + \Delta(t)]\Theta - \hat{\Theta} \tag{26}$$

where $\lim\limits_{t\to\infty} \Delta(t) \to 0$

*Remark 2: Note that a practical test for $M(t) > 0$ is to verify in an experiment if $K(t)M(t) \approx I$ holds. This implies non-singularity of $K(t)$ and $M(t)$. Again, this condition can be achieved through PE of $\phi$ (see Remark 1) which can be verified experimentally as will be seen in Section VII on Practical Application Results.* ○

### C. Parameter Estimation

We shall write our adaptive law as,

$$\dot{\hat{\Theta}} = \Gamma[\varphi^T F(y - C\hat{x}) - \Omega R(t)] \tag{27}$$

In (27), $\Gamma$ and $\Omega$ are positive definite and diagonal design matrices, i.e. $\Gamma = diag(\gamma_1, \dots, \gamma_l)$ and $\Omega = diag(\omega_1, \dots, \omega_l)$ respectively. The term $R(t)$ contains a sliding mode type term to ensure fast parameter convergence,

$$R(t) = \Omega_1 \frac{\hat{\Theta} - K(t)N(t)}{\delta + \left\|\hat{\Theta} - K(t)N(t)\right\|} + \Omega_2(\hat{\Theta} - K(t)N(t)) \tag{28}$$

where $\Omega_1$ and $\Omega_2$ are diagonal positive definite matrices, whilst $\delta$ is a positive constant. It will be proven that the parameter error matrix, $\tilde{\Theta}$, converges to a small residual set around zero, $\left\|\tilde{\Theta}\right\| \leq c$, in finite time, where $c > 0$ is a positive constant.

*Remark 3: Compared to previous results (i.e. the parameter adaptation is only driven by the observer error in (27)), the extra term $R(t)$ taking parameter error information, $\hat{\Theta} - K(t)N(t)$ is employed, which could enhance the parameter convergence performance [21]. In particular, we incorporate the sliding mode technique in (28) such that the finite-time convergence to a set of ultimate boundedness is guaranteed as stated in the next section.* ○

## V. Stability and Performance

*Theorem 1: Given a system (1), which satisfies Assumption 1-4, an adaptive observer (3) with adaptation law (27) using (13) - (19), (28) can be designed for persistently excited $\phi$ (9) so that the unknown parameter vector $\Theta$ can be estimated via $\hat{\Theta}$ within finite time satisfying an ultimate bounded stability characteristic for $\tilde{\Theta}$ and the estimated state $\tilde{x}$. The set of ultimate boundedness can be arbitrarily small for $\zeta = 0$.* ◇

*Proof :* The following Lyapunov candidate shall be employed,

$$V(t) = \frac{1}{2}\tilde{x}^T P\tilde{x} + \frac{1}{2}\tilde{\Theta}^T\Gamma^{-1}\tilde{\Theta} \tag{29}$$

For ease of analysis, we shall decompose (29) as,

$$V(t) = \tfrac{1}{2}\tilde{x}_1^T P_1\tilde{x}_1 + \tfrac{1}{2}\tilde{x}_2^T P_2\tilde{x}_2 + \tfrac{1}{2}\tilde{\Theta}^T\Gamma^{-1}\tilde{\Theta} = V_1 + V_2 + V_3 \tag{30}$$

We now analyse the functions of $V_1 = \frac{1}{2}\tilde{x}_1^T P_1\tilde{x}_1$ and $\tilde{V} = V_2 + V_3 = \frac{1}{2}\tilde{x}_2^T P_2\tilde{x}_2 + \frac{1}{2}\tilde{\Theta}^T\Gamma^{-1}\tilde{\Theta}$ separately for convenience. The derivative of $\tilde{V} = \frac{1}{2}\tilde{x}_2^T P_2\tilde{x}_2 + \frac{1}{2}\tilde{\Theta}^T\Gamma^{-1}\tilde{\Theta}$ can be verified as,

$$\begin{aligned}\dot{\tilde{V}} &= \tfrac{1}{2}[\tilde{x}_2^T P_2(A_{c22}\tilde{x}_2 + \bar{B}_2\varphi\tilde{\Theta}) + (A_{c22}\tilde{x}_2 \\ &\quad +\bar{B}_2\varphi\tilde{\Theta})^T P_2\tilde{x}_2] + \tilde{\Theta}^T\Gamma^{-1}\dot{\tilde{\Theta}} + 2\tilde{x}_2^T P_2\zeta_2\end{aligned}$$

Using (27), it follows:

$$\begin{aligned}\dot{\tilde{V}} &= -\tfrac{1}{2}\tilde{x}_2^T Q_2\tilde{x}_2^T + \tilde{x}_2^T P_2\bar{B}_2\varphi\tilde{\Theta} \\ &\quad -\tilde{\Theta}^T\Gamma^{-1}[\Gamma(\varphi^T FC\tilde{x} - \Omega R(t))] + 2\tilde{x}_2^T P_2\zeta_2\end{aligned} \tag{31}$$

The observer error, $C\tilde{x}$ can be written as $\tilde{x}_2$ from (2), $Cx = x_2$. From (5), knowing that $P_2B_2 = (FC)^T$, equation (31) can be further simplified as,

$$\dot{\tilde{V}} = -\tfrac{1}{2}\tilde{x}_2^T Q_2\tilde{x}_2^T + \tilde{\Theta}^T\Omega R(t) + 2\tilde{x}_2 P_2\zeta_2 \tag{32}$$

Taking care of the diagonal positive definite matrices, i.e. $\tilde{\Omega}_1 = \Omega\Omega_1$ and $\tilde{\Omega}_2 = \Omega\Omega_2$ with $\zeta_{2KN} = K(t)\zeta_{2N}$ for ease of analysis, equation (32) can be written with the sliding-mode term, $R(t)$ using (26),

$$\begin{aligned}\dot{\tilde{V}} &= -\tfrac{1}{2}\tilde{x}_2^T Q_2\tilde{x}_2 + (\Theta - \hat{\Theta})^T\tilde{\Omega}_1\tfrac{\hat{\Theta}-K(t)N(t)}{\delta+\|\hat{\Theta}-K(t)N(t)\|} \\ &\quad +(\Theta - \hat{\Theta})^T\tilde{\Omega}_2[\hat{\Theta} - K(t)N(t)] + 2\tilde{x}_2P_2\zeta_2 \\ &\leq -\tfrac{1}{2}\tilde{x}_2^T Q_2\tilde{x}_2 - \lambda_{min}(\tilde{\Omega}_1)\tfrac{\|\hat{\Theta}-K(t)N(t)\|^2+\delta\|\hat{\Theta}-K(t)N(t)\|}{\delta+\|\hat{\Theta}-K(t)N(t)\|} \\ &\quad +\lambda_{min}(\tilde{\Omega}_1)\tfrac{\delta\|\hat{\Theta}-K(t)N(t)\|}{\delta+\|\hat{\Theta}-K(t)N(t)\|} + 2\tilde{x}_2P_2\zeta_2 \\ &\quad +(K(t)N(t) - \hat{\Theta})^T\tilde{\Omega}_2[\hat{\Theta} - K(t)N(t)] \\ &\quad +(\Delta(t)\Theta - \zeta_{2KN})^T[\tilde{\Omega}_2\left(\hat{\Theta} - K(t)N(t)\right) \\ &\quad +\tilde{\Omega}_1\left(\tfrac{\hat{\Theta}-K(t)N(t)}{\delta+\|\hat{\Theta}-K(t)N(t)\|}\right)] \\ &\leq -\tfrac{1}{2}\lambda_{min}(Q_2)\lambda_{min}(P_2^{-1})V_2 - \lambda_{min}(\tilde{\Omega}_2)\tfrac{\lambda_{min}(\Gamma)}{2}V_3 \\ &\quad -(\lambda_{min}(\tilde{\Omega}_1)\lambda_{min}(\Gamma^{1/2}) \\ &\quad -\lambda_{max}(\tilde{\Omega}_2)\|\Delta(t)\Theta\|\lambda_{max}(\Gamma^{1/2}))\sqrt{V}_3 \\ &\quad +2\lambda_{max}(\tilde{\Omega}_2)\|\Delta(t)\Theta - \zeta_{2KN}\|^2 + \lambda_{min}(\tilde{\Omega}_1)\delta \\ &\quad +2\lambda_{max}(\tilde{\Omega}_1)\|\Delta(t)\Theta - \zeta_{2KN}\| + 2\|\tilde{x}_2\|\|P_2\|\|\zeta_2\|\end{aligned} \tag{33}$$

There are suitable positive scalars $c_1, c_2, c_3$ for large enough time, $t > 0$ such that:

$$\tilde{V} \leq -c_1\tilde{V} - c_2\sqrt{V}_3 + c_4 \tag{34}$$

Therefore, $\tilde{x}_2$ and $\tilde{\Theta}$ converge to a compact set bounded by parameter $\delta$, $\|\Delta(t)\|(\lim_{t\to\infty}\Delta(t) \to 0)$ and $\|\zeta_{2KN}\|$. The term $\Delta(t)$ denotes the effect of the initial conditions of $K^{-1}(0)$. For $\zeta_{2KN} = 0$, the size of the compact set can be adjusted to be smaller by reducing $\delta$ and the elements $\lambda_i$ in matrix $K^{-1}(0)$. Note that ultimate bounded stability for $\tilde{x}_1$ and subsequently for $\tilde{x}$, now trivially follow. Again, for $\zeta = 0$, the set of ultimate boundedness can be arbitrarily small for suitable choice of $\delta$. ■

*Remark 4: The result in Theorem 5.1 in fact is quite generic. It also allows for analysis of measurement errors of $x_2$ and $\phi$. For this reason, we may have in the observer some measurement errors affecting both $x_2$ and also $\phi(x, u_2)$ measurement and in reality $\check{x}_2$ and $\check{\phi}(x, u_2)$ are provided in the practical system. Thus, the observer equation is*

$$\dot{\hat{x}} = A\hat{x} + B_1u_1 + B\check{\phi}\hat{\Theta} + L(\check{x}_2 - C\hat{x}) \tag{35}$$

*The plant dynamics in (8) can be rewritten as,*

$$\begin{aligned}\dot{\check{x}} &= (A\check{x} + B_1u_1) + B_2\check{\phi}\Theta + \zeta + (\dot{\check{x}} - \dot{x}) \\ &\quad +A(x - \check{x}) + B_2(\phi - \check{\phi})\Theta\end{aligned} \tag{36}$$

*where $\check{x} = [x_1^T \quad \check{x}_2^T]^T$. Assuming the measurement errors and its derivative are bounded, (i.e. $(x-\check{x}), (\dot{x}-\dot{\check{x}}), (\phi-\check{\phi}) \in L_\infty$), so that $\check{x}_2, \dot{\check{x}}_2 \in L_\infty$, then the plant dynamics (8) are,*

$$\dot{\check{x}} = (A\check{x} + B_2u_1) + B_2\check{\phi}\Theta + \check{\zeta}, \tag{37}$$

*where $\check{\zeta} = \zeta+(\dot{\check{x}}-\dot{x})+A(x-\check{x})+B_1(\phi-\check{\phi})\Theta$ can be regarded as a bounded disturbance. Defining the error dynamics as $\tilde{\check{x}} = (\check{x} - \hat{x})$ it follows that,*

$$\dot{\tilde{\check{x}}} = A_c\tilde{\check{x}} + B_2\check{\phi}\Theta + \check{\zeta} \tag{38}$$

*Under the assumption that $\check{\zeta}$ is bounded, we can continue the analysis as for Theorem 5.1. Boundedness of $\check{\zeta}$ might be achieved under suitable assumptions on the measurement errors affecting $x_2$ and an additional assumption on the nonlinear functions $\phi$.* ○

## VI. Parameter Estimation in the Vehicle Dynamics

In this section, we will discuss the previously formulated parameter estimation algorithm in the context of its application to road gradient and vehicle's weight estimation. Figure 1 shows the simplified model of the small-scaled model car used in the experiment to validate the parameter estimation algorithm.

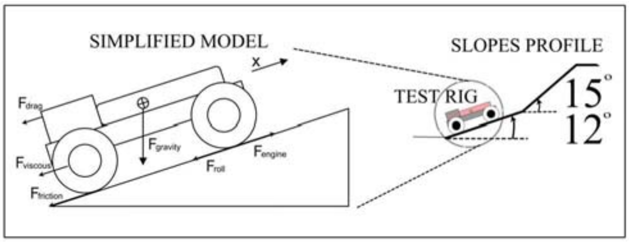


Fig. 1. Simplified model of the small-scaled model car and the slope profile

### A. Vehicle model

The parameters to be estimated are the road inclination, $\theta$, on which the vehicle traverses, the mass of the vehicle, $m$ and the viscous friction coefficient, $C_{vf}$. Referring to Figure 1, assuming the air drag, $F_{drag}$ and the rolling friction, $F_{roll}$ are negligible, and the braking force, $F_{brake}$ is subsumed in the driving force, $F_{engine}$, we may model the small-scaled model car using Newton's Second Law in the longitudinal direction to yield,

$$m\ddot{x} = F_{engine} - mgsin(\theta) - C_{VF}\dot{x} \tag{39}$$

where $m$ is the mass of the vehicle, $\theta$ is the road gradient on which the vehicle traverses, $\dot{x}$ is the vehicle's velocity and $C_{VF}$ is the viscous damping coefficient.

### B. Observer Design

Following the general structure presented in (3), the adaptive observer with finite-time parameter estimation can be written as,

$$\dot{\hat{\mathbf{x}}} = A\hat{\mathbf{x}} + \begin{bmatrix} 0 \\ 1 \end{bmatrix}[g \quad F_{engine} \quad \dot{x}]\begin{bmatrix} \hat{s} \\ \hat{b} \\ \hat{f} \end{bmatrix} + L(y - \hat{y}) \tag{40}$$

where $A$ is the system matrix (adheres to the Assumption 4), $\hat{s}$, $\hat{b}$, $\hat{f}$ are the estimated parameters of $-sin\theta$, $\frac{1}{m}$ and $-\frac{C_{VF}}{m}$ respectively whereas $L$ is the observer gain chosen to deliver the positive definite Lyapunov matrix, $P$, such that it satisfies (4). The engine driving force, $F_{engine}$ is assumed to be bounded to ensure that the system states, $x$ remains bounded. The vector $\hat{y} = C\hat{\mathbf{x}}$ will be the corresponding observer output and $\hat{\mathbf{x}} = [\hat{x} \quad \dot{\hat{x}}]$ is the observed state vector. Thus, using this structure it follows,

$$B_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \hat{\Theta} = \begin{bmatrix} \hat{s} \\ \hat{b} \\ \hat{f} \end{bmatrix}, \quad \varphi = \begin{bmatrix} g \\ F_{engine} \\ \dot{x} \end{bmatrix}^T \tag{41}$$

The observer adaptive weights are lumped such that,

$$\Gamma = diag(\gamma_1, \gamma_2, \gamma_3), \quad \Omega = \Gamma^{-1} diag(r_s, r_b, r_f) \tag{42}$$

## VII. Practical Application Results

The small-scale model car, previously built by Foreman *et al.* [18],was used in the experiment to evaluate the estimation algorithm. The vehicle's mass (nominally weighs 10kg) and the road gradients on which it traverses were the two parameters to be estimated. Figure 2 shows the implemented controller system network and architecture which emulates the system network of a road vehicle. Together with Matlab$^{TM}$, dSPACE MicroAutoBox, used in the experiment, is a dedicated Rapid Prototyping embedded system suited to test the proposed estimation algorithm. The drive train comprises of an EPOS 24/5 motor driver and the brushless DC motor (EC-i 40 Maxon) representing the vehicle's engine. The motor is current-controlled via the MicroAutoBox which subsequently provides the driving force, $F_{engine}$, proportional to the current signal being controlled. Gradient measurements provided by the installed SCA61T inclinometer is entirely for reference purpose and not to be used in the algorithm. In our experiment, we would avoid the occurrence of significant slippage as this would invalidate our estimation effort. The test slope was constructed using three stiff wooden planks of 2 m in length each. They are tilted and bolted together to give a slope profile with tilting angle of 12° for the first slope, 15° for the second slope and the last slope is horizontal i.e. parallel to the ground. The small-scaled model car was required to traverse up the designated slope at a constant speed of 0.2 m/s (Figure 1).

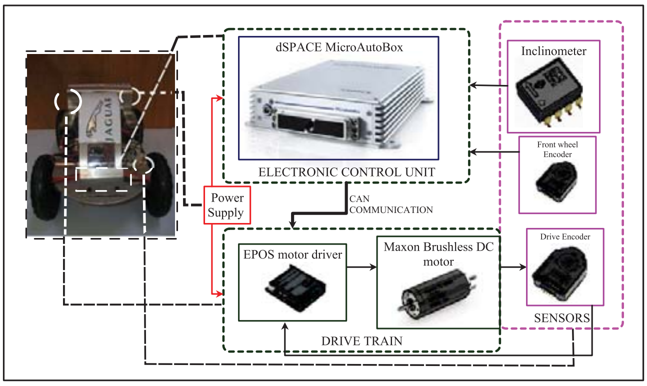


Fig. 2. Functional Structure of the System Onboard

### A. Parameters tuning

In our experiment, there were important parameters (see Table I) of the adaptive observer algorithm needed to be tuned to achieve satisfactory results. Realistic and acceptable physical bounds/limits were considered and shown in Table II. The corresponding values of the parameters displayed in the table are used in the experiment. Noise injected to the velocity and the control signal was kept constant in terms of power so that the actuating signal applied to the motor exerts sufficient and persistent excitation.

TABLE I
Parameters to be tuned

| Parameter Description | Symbols | Values |
|---|---|---|
| Observer Adaptive weights, $\Gamma$ | $\gamma_1$ | 0.01 |
| | $\gamma_2$ | 0.001 |
| | $\gamma_3$ | 0.001 |
| Sliding-Mode Adaptive weights, $\Omega$ | $r_s$ | 0.01 |
| | $r_b$ | 1 |
| | $r_f$ | 0.00001 |
| Forgetting Factor | $k_{FF}$ | 0.6 |
| Filter Poles | $k$ | 0.005 |
| Regressor Matrix, K initial condition | $K(0)$ | diag(0.4,0.4,1) |

TABLE II
Saturation Limits

| Plant Parameters | Estimation | Lower Limit | Upper Limit |
|---|---|---|---|
| Mass(m,kg) | $\hat{m}$ | 0.1 | 20 |
| | $\hat{b} = 1/\hat{m}$ | 0.05 | 10 |
| Gradient($\theta$,°) | $\hat{\theta}$ | -20 | 20 |
| | $\hat{s} = sin(\hat{\theta})$ | -0.4 | 0.4 |
| Friction Coefficient | $-C_{VF}$ | 240 | 0.1 |
| ($C_{VF}$,$kg/s$) | $\hat{f} = C_{VF}/\hat{m}$ | -12 | -1 |

### B. Results

The proposed adaptive observer algorithm with sliding-mode term performance is compared with that of the recent adaptive observer (without the term in (28)) previously carried out by [18]. Referring to Table III, the Integral Absolute Error (IAE), $\int_0^\infty |e(t)|dt$, is used as the performance index measuring the difference between the actual and the estimated road gradient. The low value of IAE translates to good parameter estimation as the estimated value converges to the true value. Figure 3(a) versus Figure 5(a) shows an excellent estimation performance of the new algorithm, having the estimated gradient converging close to the actual gradient during the climbing of the slope although there is a slight offset during the vehicle traversing the flat ground. A very consistent mass estimation of the new estimator is evident in Figure 3(b) as compared to the previous algorithm as shown in Figure 5(b). The estimated mass value settles at approximately 14 kg throughout the test slope profile and slightly descends to 12 kg towards reaching the flat ground at the end. The recorded IAE for the road gradient estimation for the proposed algorithm is 178.2402. In contrast, the previous algorithm lacks in performance when it exhibits a large increase in IAE of 229.9736. The estimated parameters, $\hat{s}, \hat{b}, \hat{f}$ in Figure

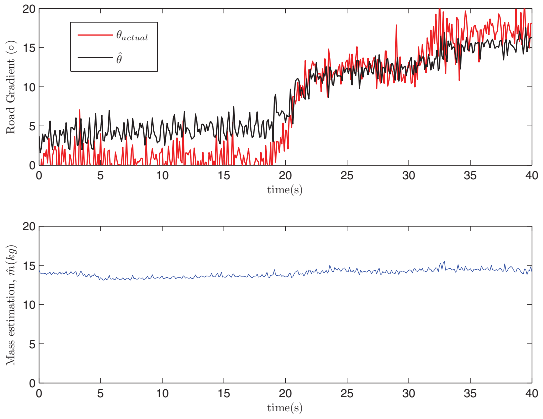


Fig. 3. (a) Gradient comparison(in Degree) between the actual, $\theta$ and estimated, $\hat{\theta}$ (b) Mass Estimation (kg) of the proposed novel algorithm

TABLE III
SUMMARY OF THE PERFORMANCE IN IAE

| Algorithm | IAE |
|---|---|
| Previous algorithm | 229.9736 |
| Proposed Finite-Time Estimation algorithm | 178.2402 |

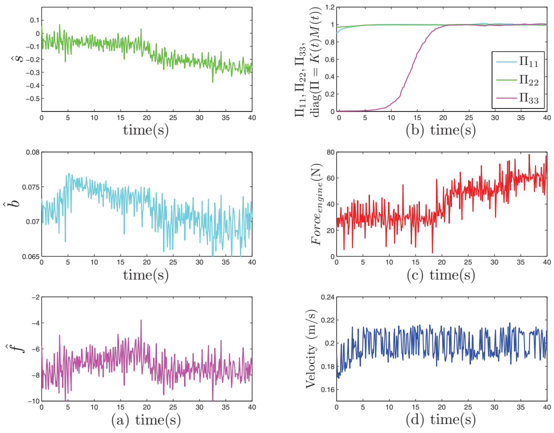


Fig. 4. (a) Estimated parameters($\hat{\Theta}$) consisting of $\hat{s}, \hat{b}, \hat{f}$, (b) proof of PE by which $K(t)M(t) \approx I$ holds experimentally, (c) Driving force,$F_{engine}$ and (d) vehicle's velocity, $V$

4(a) also show excellent behaviour as they remain within the bound of the given physical limit without saturation whereas in Figure 5(b), the estimated parameters saturate at the given bounds. Interestingly, the engine driving force, $F_{engine}$ signal in Figure 4(c) exhibits persistent excitation (PE) throughout the test slope profile which assures the finite-time convergence to the true value. Note that persistent excitation has been also verified via the approach of Remark 3. The product $K(t)M(t)$ has converged to unity as evident in the experimental data shown in Figure 4(b). Table III sums up the performance of the proposed algorithm compared with the previous one in terms of IAE.

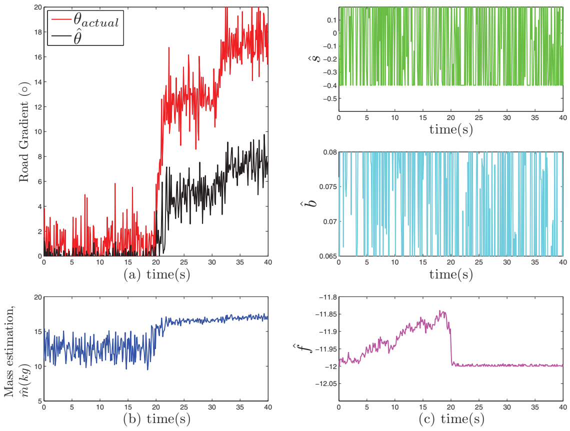


Fig. 5. (a) Gradient comparison between the actual, $\theta$ and the estimated, $\hat{\theta}$, (b) Mass Estimation (kg) and (c) Estimated parameters($\Theta$) consisting $\hat{s}, \hat{b}, \hat{f}$ of algorithm from [18]

## VIII. CONCLUSION

An adaptive observer with novel sliding-mode based parameter estimation algorithm to estimate the vehicle's mass and the road gradient is presented. The proposed parameter estimator with the sliding-mode term has been proven analytically to be finite time convergent to an error of well defined bound. The algorithm shows significant levels of robustness to disturbances and a particular class of measurement errors. The analytical results are further supported and validated by the practical implementation in a form of experiments conducted on a small-scale vehicle traversing a designated test slope profile with certain parameters tuned. The practical results show a significant improvement over the previous algorithm in terms of persistent excitation (PE), realistic values within the physical bound, estimation accuracy and convergence.

## REFERENCES

[1] H. Jansson, E. Kozica, P. Sahlholm, and K.-H. Johansson, "Improved road grade estimation using sensor fusion," in *Proceedings of the 12th Reglermote in Stockholm, Sweeden*, 2006.

[2] V. Winstead and I. Kolmanovsky, "Estimation of road grade and vehicle mass via model predictive control," in *Control Applications, 2005. CCA 2005. Proceedings of 2005 IEEE Conference on*, 2005, pp. 1588 –1593.

[3] S. Mangan, J. Wang, and Q. Wu, "Measurement of the road gradient using an inclinometer mounted on a moving vehicle," in *IEEE International Symposium on Computer Aided Control System and Design*, 2002.

[4] H. Bae, J. Ryu, and J. Gerdes, "Road grade and vehicle parameter estimation for longitudinal control using gps," in *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, 2001, pp. 166–171.

[5] P. Lingman and B. Schmidtbauer, "Road slope and vehicle mass estimation using Kalman filtering," *Vehicle System Dynamics*, vol. 37, pp. 12–23, 2002.

[6] P. Sahlholm, *Iterative road grade estimation for heavy duty vehicle control*. Elektrotekniska system, Kungliga Tekniska hogskolan, 2008.

[7] A. P. H. Vahidi, A.; Stefanopoulou, "Experiments for online estimation of heavy vehicle's mass and time-varying road grade," in *Proceedings IMECE, Jan 2003*, 2003.

[8] A. Vahidi, M. Druzhinina, A. Stefanopoulou, and H. Peng, "Simultaneous mass and time-varying grade estimation for heavy-duty vehicles," in *American Control Conference, 2003. Proceedings of the 2003*, vol. 6. IEEE, 2003, pp. 4951–4956.

[9] J. Barrho, M. Hiemer, U. Kiencke, and T. Matsunaga, "Estimation of Elevation Difference Based on Vehicle's Inertial Sensors."

[10] P. Sahlholm and K. Johansson, "Segmented road grade estimation for fuel efficient heavy duty vehicles," in *IEEE Conference on Decision and Control*, 2010.

[11] ——, "Road grade estimation for look-ahead vehicle control," in *Proceedings of the 17th IFAC World Congress*, 2008.

[12] ——, "Road grade estimation for look-ahead vehicle control using multiple measurement runs," *Control Engineering Practice*, vol. 18, no. 11, pp. 1328 – 1341, 2010.

[13] S. Mangan and J. Wang, "Development of a novel sensorless longitudinal road gradient estimation method based on vehicle can bus data," *Mechatronics, IEEE/ASME Transactions on*, vol. 12, no. 3, pp. 375 – 386, 2007.

[14] P. Iaonnou and J. Sun, *Robust Adaptive Control*. Prentice Hall, 1996.

[15] V. Adetola and M. Guay, "Finite-Time Parameter Estimation in Adaptive Control of Nonlinear Systems," *IEEE Transactions on Automatic Control*, vol. 53, no. 3, pp. 807–811, Apr. 2008. [Online]. Available: http://dx.doi.org/10.1109/TAC.2008.919568

[16] ——, "Performance Improvement in Adaptive Control of Linearly Parameterized Nonlinear Systems," *IEEE Transactions on Automatic Control*, vol. 55, no. 9, pp. 2182–2186, Sep. 2010. [Online]. Available: http://dx.doi.org/10.1109/TAC.2010.2052149

[17] C. Edwards and S. Spurgeon, *Sliding mode control: theory and applications*. CRC, 1998.

[18] J. Foreman, M. Yazdouni, S. Russo, G. Chan, A. Spiers, G. Herrmann, and P. Barber, "Hardware in the loop validation of a gradient and weight estimation algorithm and longitudinal speed control using a laboratory model car," in *International Conference on Systems Engineering (ICSE)*, 2009.

[19] J. Na, G. Herrmann, X. Ren, M. Nasiruddin, and P. Barber, "Robust adaptive finite-time parameter estimation and control of nonlinear systems," in *Multi-Conference on System and Control 2011, MSC2011*, 2011.

[20] N. Shimkin and A. Feuer, "Persistency of excitation in continuous system," *System and Control Letters*, pp. 225–233, 1987.

[21] J.-J. Slotine and W. Li, *Applied Nonlinear Control*. Prentice Hall, Oct. 1990. [Online]. Available: http://www.worldcat.org/isbn/0130408905

[22] T. Soderstrom and P. Stoica, *System Identification*, M. Grimble, Ed. Prentice Hall, 1989.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Experimental Study of a Capsubot for Two Dimensional Movements

M. Nazmul Huda, Hongnian Yu and Michael J. Goodwin

*Abstract*— **A capsubot (capsule robot) which works on the principle of internal reaction force has no external moving parts whereas a conventional robot has legs and/or wheels. It is an underactuated mechanical system and has a lot of potential applications such as medical diagnosis, underground pipe leakage detection, rescue work in the hazardous environment etc. However, most capsule robots studied/developed can only move in one dimension (1D). This paper presents the implementation of the recently proposed double parallel mass capsubot which uses two parallel inner masses (IMs) to move the capsubot in two dimensions (2D). A three stage control strategy is proposed to resolve the control issue of underactuated mechanical system. A closed-loop control approach is applied to the IMs of the capsubot. The comparison with the simulation studies are also obtained and analyzed.**

## I. INTRODUCTION

Minimally invasive diagnosis and interventions feature safe and reliable techniques and result in shorter hospital stays, less pain, more rapid return to daily work, and improved immunological response compare to the conventional ways. Thus relevant researches have gotten extensive interest to develop minimally invasive devices for surgical and diagnostic applications [1], [2], [3], [4]. Robot-assisted laparoscopic and thoracoscopic surgeries are becoming popular because of its less invasiveness and reliability [5]. Researchers are designing mobile robots to perform abdominal surgery which will further decrease the invasiveness [6], [7], [8].

Passive capsule endoscopy has been used to diagnose GI (gastro-intestinal) tract diseases which is safer for the patient and easier to perform compare to the conventional probe endoscopy. Research is ongoing in order to add mobility to capsule endoscope to increase the reliability, performance and add functionalities such as biopsy, drug delivery, surgery etc. Mobile robots designed for abdominal cavity as well as GI tract can be divided based on the propulsion mechanism used as 1) external propulsion robot (i. magnetic propulsion [9]) 2) internal propulsion robot (i. wheeled [10] ii. legged [11] iii. internal reaction force propulsion robot [12]) and 3) hybrid propulsion robot (combination of external and internal) [13].

The sharp edges of legs or wheels of the internal propulsion robots create the risk to injure the tender GI track and internal organs. Also to perform magnetic propulsion it requires large external magnet [9] which makes the system expensive.

On the contrary the capsubots based on internal reaction force are simple in construction and have no external legs or wheels [14], [15]. The structure of the principle is derived from [16] where impact force and dry friction is utilised to create motion. A mass attached to the main object through a piezoelectric element, is made to move away from the main object rapidly and then to return to the initial position slowly with a sudden stop. The main object moves during the rapid motion and at the stopping moment and remains stationary for the rest of the time. The object can move along a straight line by repeating the above process. In [17] the propulsion principle was analyzed from the viewpoint of physics and a control law and optimum parameters of the system were proposed. In [18], [19], the motion generation of a single mass capsubot was explained on the basis of a four step velocity profile which is, fast motion for first two steps and slow motion in the last two steps. In [15], the motion generation of the capsubot was explained on the basis of a seven step velocity profile which is, fast motion in the first three steps and slow motion in the rest of the steps. A pendulum-driven cart was analysed in [14] with a six step velocity profile. In [12], motion of a single mass capsubot was explained on the basis of a novel four step acceleration profile and a stand-alone prototype was developed based on the profile.

However, except [20] all the capsubots researched so far can only move in one dimension (1D) with their present designs. In practical applications, high dimensional movements of a capsubot is required. In [21] a double-pendulum driven cart was proposed for 2D movements and simulation results was presented. But this cart has wheels and thus lacks the advantages of a capsubot which is legless and wheelless. In [20] a capsubot with 2D movements was proposed and theoretical analysis was performed. In this paper practical implementation of the capsubot proposed in [20] is presented. The experimental results are analyzed and compared with the theoretical results.

The paper is structured as below. Section II presents the modelling and motion generation of the capsubot. Prototyping and programming of the capsubot prototype is explained in section III. Section IV proposes a control strategy for the underactuated mechanical system and explains the control of the 2D capsubot. Section V presents experimental results, compares them with simulation results and analyses them. Finally in section VI the paper is concluded and future direction of the research is presented.

This work has been supported by the European Erasmus-Mundus Sustainable eTourism project 2010-2359, the EPSRC UK-Japan Network on Human Adaptive Mechatronics Project (EP/E025250/1) and EU Erasmus Mundus Project-ELINK (EM ECW-ref.149674-EM-1- 2008-1-UK-ERAMUNDUS). The authors thank Mr. Sam Wane and Dr Ian Moorcroft for their comments and assistances while performing this research.

M. Nazmul Huda, Hongnian Yu and Michael J. Goodwin are with the Faculty of Computing, Engineering and Technology, Staffordshire University, UK. Email: {m.n.huda, h.yu and m.j.goodwin}@staffs.ac.uk

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

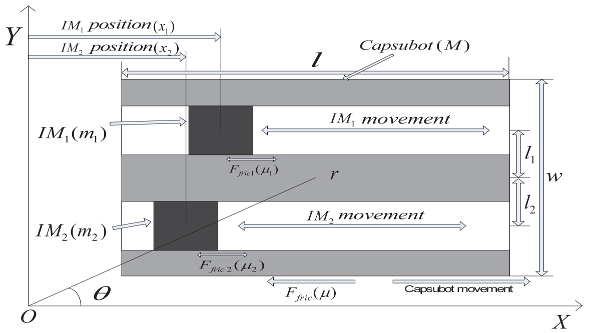


Fig. 1. Schematic diagram of double parallel mass capsubot [20]

## II. Modelling and Motion Generation

### A. Dynamic Modelling

The capsubot proposed in [20] is shown in Fig. 1. Two IMs are placed in the hollow spaces within the capsubot. IMs can move along the hollow spaces. By controlling the movements of IMs the capsubot can be moved in a 2D plane. It can perform linear, rotational, and 2D motions.

The capsubot dynamic models are as follows [20]:

$$F_i = m_i \ddot{x}_i + sgn(\dot{x}_i - \dot{r})\mu_i m_i g \quad \forall \;\; i = 1, 2 \qquad (1)$$

$$\sum F = -m_1 \ddot{x}_1 - m_2 \ddot{x}_2 = M\ddot{r} + sgn(\dot{r})\mu M g \qquad (2)$$

$$\sum M_G = m_2 \ddot{x}_2 l_2 - m_1 \ddot{x}_1 l_1 = I\ddot{\theta} + M_f \qquad (3)$$

where $sgn(\dot{x}_i - \dot{r})\mu_i m_i g$ is the friction force between the $IM_i$ and the capsubot; $sgn(\dot{r})\mu M g$ is the friction force between the capsubot with the environment; $M_f = sgn(\dot{\theta})\frac{2}{3}\mu_r P(r_2 + \frac{wl - \pi r_2^2}{\pi r_1})$ [20] is the frictional moment of the capsubot with surface of movement and $I = \frac{1}{12}M(l^2 + w^2)$ is the moment of inertia of the capsubot about z-axis through the mass centre of the capsubot. Rest of the parameters are explained in Table I.

### B. Motion Generation

In motion generation, we consider the following 3 cases: 1) Linear motion; 2) Rotational motion and 3) 2D motion. Here we design that $m_1 = m_2$ and $l_1 = l_2$.

*1) Linear Motion:* If both the IMs move with same acceleration i.e $\ddot{x}_1 = \ddot{x}_2$ then $\sum M_G = 0$, and $\sum F \neq 0$. Thus for linear motion both IMs follow the acceleration profile proposed in [12]. As $\sum M_G = 0$, so $\theta = 0$. Thus, $x = r \;\; and \;\; y = 0$ ( using $x = r\cos\theta \;\; and \;\; y = r\sin\theta$).

*2) Rotational Motion:* If both the IMs move with same acceleration in opposite direction i.e. $\ddot{x}_1 = -\ddot{x}_2$ then $\sum M_G \neq 0$, and $\sum F = 0$. Here one of the IMs follows the acceleration profile same as linear motion. The remaining IM follows an acceleration profile that is same in magnitude but opposite in direction. As $\sum F = 0$, so $r = 0$. Thus, $x = y = 0$. As $\sum M_G \neq 0$, so $\theta \neq 0$.

TABLE I

Parameters of the Capsubot

| | |
|---|---|
| $F_i$ | Force applied on $IM_i$ |
| $\sum F$ | Resultant reaction force acting on the mass centre of capsubot |
| $\sum M_G$ | Resultant moment on the capsubot about z-axis through the mass centre of the capsubot |
| $m_i$ | Mass of $IM_i$ (cylindrical rod with extra mass) |
| $x_i$ | Position of $IM_i$ with respect to an external fixed frame |
| $r$ | Position of the mass centre of the casubot with respect to an external fixed frame |
| $\theta$ | Angular position of the casubot with respect to an external fixed frame |
| $x$ | $r\cos\theta$ |
| $y$ | $r\sin\theta$ |
| $M$ | Total mass of the capsubot |
| $w$ | Width of the capsubot |
| $l$ | Length of the capsubot |
| $h$ | Height of the capsubot |
| $\mu$ | Dynamic friction coefficient between capsubot and environment (plywood table) |
| $\mu_r$ | Rotational friction coefficient between capsubot and environment (plywood table) |
| $\mu_i$ | Dynamic friction coefficient between $IM_i$ and linear DC motor ($LM_i$) |
| $g$ | Acceleration of gravity |
| k | Total stroke length in one direction |
| $F_{ipmax}$ | Maximum peak force on $IM_i$ |
| $F_{icmax}$ | Maximum continuous force on $IM_i$ |
| $\ddot{x}_{imax}$ | Maximum achievable acceleration of $IM_i$ |
| $l_i$ | Perpendicular distance of the mass centre of the capsubot from the direction of forces $F_i$ |
| $r_1$ | $\frac{1}{2}\sqrt{l^2 + w^2}$ |
| $r_2$ | $w/2$ |
| $P$ | $Mg$ |

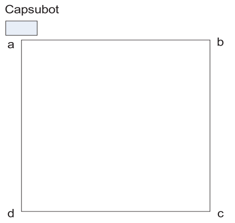


Fig. 2. Squared motion path

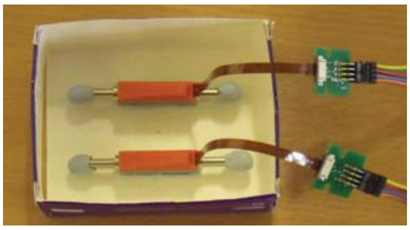
*3) 2D Motion:* By combining the linear and rotational motion, the capsubot can perform a 2D motion. Squared motion, rectangular motion, circular motion etc. are examples of this type of motion.

To travel a-b-c-d-a squared path of Fig. 2 the capsubot would move 'a' to 'b' using linear motion mode and then it would rotate 90 degree using rotational motion mode and then again would use linear motion. By using linear motion and rotational motion alternately the capsubot completes its travel path a-b-c-d-a.

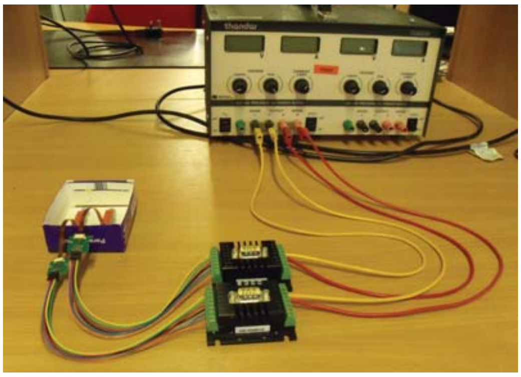
## III. Experimental Setup

A prototype shown in Fig. 3(a) has been developed to validate the theoretical analysis performed in [20]. Here the cylindrical rods of two linear DC motors (LM0830-015-01) [22] (see Fig. 4(A)) are used as two IMs. The linear

DC motors (LMs) are placed and attached using adhesive on a housing made of thin paperboard and thus forms the prototype. It is connected to a motion controller through wires. The parameters of the capsubot are listed in Tables I and II.



(a) Capsubot prototype



(b) Capsubot prototype with controllers and power supply

Fig. 3. Implemented Capsubot

The main components of the linear DC motor (LM) are a housing or motor shell which houses the coil, Hall sensors, a PCB (printed circuit board) etc. and a cylindrical rod which is a permanent magnet. The cylindrical rod can move back and forth through the housing. The cylindrical rod can move 7.5mm in each direction from its middle position. Here we used 6 mm and left the rest as a clearance. We added extra mass to the both ends of the cylindrical rod to increase the IM mass to capsubot mass ratio. We shall use the term IM (inner mass) for the cylindrical rod with extra mass in the rest of the paper.

The motion of the IM is controlled by the motion controller shown in Fig. 4(B). A linear force is applied to the IM when the coil in the motor shell is energised by the motion controller. The linear DC motor (LM) can be connected to the motion controller through wires and a connector. It takes power from the motion controller. The Hall sensors sense the position of the IM and feed the data to the motion controller to form a closed loop system.

The controller is programmed to move the IM from one location to another location by using a given acceleration and deceleration. The controller by itself calculates the time it has to use for acceleration and then deceleration to reach the desired location. The controller uses three Hall sensors on each linear DC motor (LM) to take feedback about the position of the IM and corrects the input to the IM accordingly to maintain the desired acceleration or deceleration and velocity.

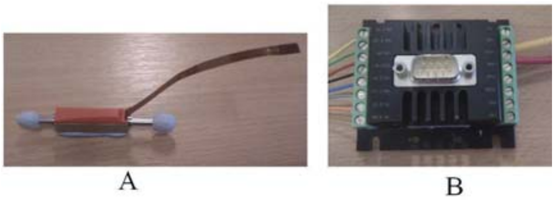


Fig. 4. A) Modified linear DC motor (LM) B) Motion controller

The motion controller is driven by $12V - 30V$ DC which is taken from a DC power supply. The motion controller of the capsubot system is programmed using the Motion Manager software [22] and the program is transferred from PC to the motion controller by a RS-232 cable and stored in the EEPROM of the motion controller. Then the motion controller can be disconnected from the PC. When the motion controller is powered the stored program is executed and the IM moves accordingly. If the motion controller is connected to the PC the Motion Manager software logs the data of the linear DC motor (LM).

## IV. CONTROL APPROACH

To use the capsubot in various applications the movement of the capsubot needs to be controlled precisely. The capsubot should be able to change its velocity while moving according to the requirement of the application. Thus we can state the problem as a position or velocity trajectory tracking problem. The capsubot is a underactuated system i.e degrees of freedom to be controlled are greater than number of control inputs. For underactuated systems trajectory tracking is still an open issue to be solved. To solve this problem we divide the problem into three stages which are described below. The schematic diagram of the complete control system is shown in Fig. 5.

- *Stage 1:* For a given trajectory of the capsubot, desired trajectories of IMs are calculated.
- *Stage 2:* For the desired trajectories of the IMs, control inputs i.e forces are calculated (open-loop). The closed-loop control is achieved by correcting the control inputs using the error which is the difference between the measured and the desired trajectories of the IMs.
- *Stage 3:* Feedback should also be taken from the capsubot position and the control input should be corrected according to the error value for tracking the position of the capsubot properly.

In this paper stage 2 of the control system is performed i.e. simulation analysis and experimentation of closed loop control of IMs are performed for the capsubot. Stages 1 and 3 shall be completed in our future research. The schematic diagram of the control system for stage 2 is shown in Fig. 6. By implementing this stage the capsubot can perform linear and rotational motion and by combining these two can perform 2D motion. If the IMs follow a fixed set of accelerations the capsubot would have a constant average linear or rotational velocity in every cycle. To change the velocity a different set of acceleration has to be chosen.

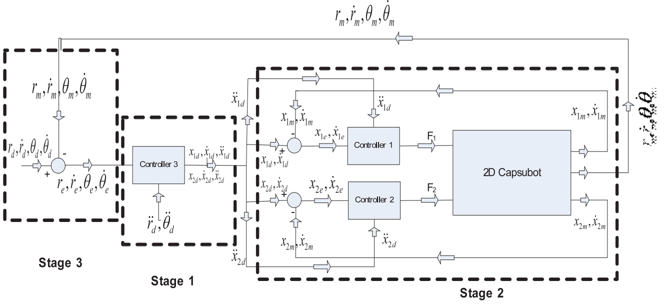

$r_m,\dot{r}_m,\theta_m,\dot{\theta}_m$
$r_m,\dot{r}_m,\theta_m,\dot{\theta}_m$
$r_d,\dot{r}_d,\theta_d,\dot{\theta}_d$
-
+
$r_e,\dot{r}_e,\theta_e,\dot{\theta}_e$
Controller 3
$x_{1d},\dot{x}_{1d},\ddot{x}_{1d}$
$x_{2d},\dot{x}_{2d},\ddot{x}_{2d}$
$\ddot{r}_d,\ddot{\theta}_d$
$\ddot{x}_{1d}$
$\ddot{x}_{2d}$
$x_{1m},\dot{x}_{1m}$
$x_{1e},\dot{x}_{1e}$
$x_{1d},\dot{x}_{1d}$
$x_{2d},\dot{x}_{2d}$
$x_{2e},\dot{x}_{2e}$
$x_{2m},\dot{x}_{2m}$
Controller 1
Controller 2
$F_1$
$F_2$
2D Capsubot
$x_{1m},\dot{x}_{1m}$
$x_{2m},\dot{x}_{2m}$
$r,\dot{r},\theta,\dot{\theta}$
Stage 3
Stage 1
Stage 2

Fig. 5. Schematic diagram of 3-stage control system for capsubot

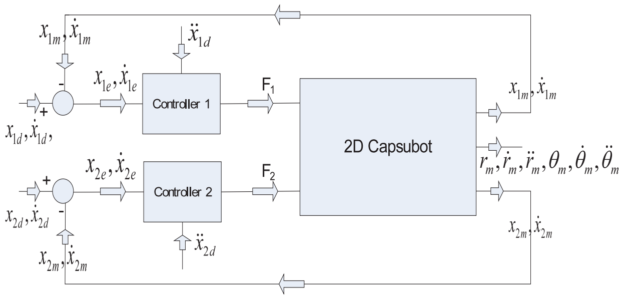


Fig. 6. Schematic diagram of Stage 2 of the control system

Two open loop control laws for the IMs for all the motion cases are:

$$F_{id} = m_i\ddot{x}_{id} + sgn(\dot{x}_{id} - \dot{r}_d)\mu_i m_i g \quad (4)$$

The closed loop control laws for the IMs can be selected, using partial feedback linearization, as [14]:

$$F_{id} = \alpha\tau_{id} + \beta \quad (5)$$

where $\alpha = m_i$ and $\beta = sgn(\dot{x}_{id} - \dot{r}_d)\mu_i mg$

Let $\tilde{x}_i = x_i - x_{id}$ be the tracking error; choosing the linear control law $\tau_{id} = \ddot{x}_i - k_1\dot{\tilde{x}}_i - k_2\tilde{x}_i$ and applying the control laws of (5) to (1) we get,

$$\ddot{\tilde{x}}_i + k_1\dot{\tilde{x}}_i + k_2\tilde{x}_i = 0 \quad (6)$$

The values of $k_1$ and $k_2$ can properly be selected using (6). Then by using the control laws of (5) the IMs can be made to follow the desired accelerations, velocities and positions which are given below. The impact of $\dot{r}_d$ on $\beta$ is neglected in simulation.

From the acceleration profiles in [12] and [20] desired IMs accelerations for linear and rotational motions are given below:

$$\ddot{x}_{id} = \begin{cases} b_{1\_i} & t \in [0, t_1) \\ b_{2\_i} & t \in [t_1, t_3) \\ b_{1\_i} & t \in [t_3, t_4] \end{cases}$$

where $b_{1\_i}$ is the acceleration of $IM_i$ in steps 1 and 4; $b_{2\_i}$ is the acceleration of $IM_i$ in steps 2 and 3; $t_1$, $t_2$, $t_3$ and $t_4$ are the time after steps 1, 2, 3 and 4 respectively.

Desired IMs velocities and positions can be calculated from the desired accelerations. All the simulations in this paper are performed using Matlab and Simulink with the help of the control law of (5) and motion equations (1), (2) and (3).

## V. EXPERIMENTAL RESULTS AND ANALYSIS

The acceleration of $IM_i$ is constrained by $\ddot{x}_i \leq min(\ddot{x}_{imax}, \frac{F_{icmax}}{m_i})$. Here $\ddot{x}_{imax}$ is $30ms^{-2}$ which is a physical constraint of $IM_i$. $F_{icmax}$ is the maximum force that can be achieved on the $IM_i$ continuously. On the other hand $F_{ipmax}$ is the maximum force that the $IM_i$ can sustain for a short time. In this experiment maximum acceleration used is $20ms^{-2}$.

The data of $IM_s$ are obtained from the Motion Manager software and then the curves are plotted using Matlab. To obtain the data for capsubot movements we recorded the motion of the capsubot using a video camera and then a video analysis software Quintic Biomechanics [23] was used to calculate the position, velocity and acceleration.

### A. Experimental Results

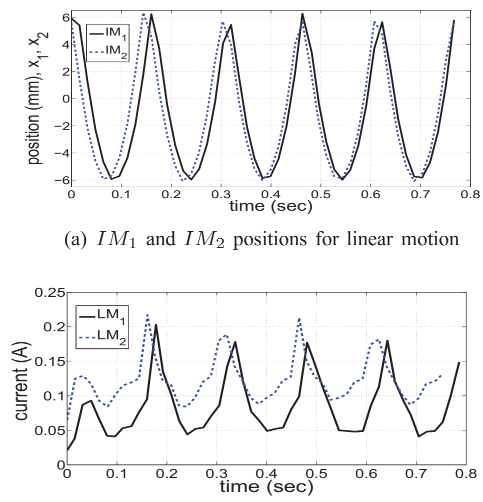
- Linear Motion: Fig. 7(a) shows the positions of $IM_1$ and $IM_2$, and Fig. 7(b) shows the currents of $LM_1$ and $LM_2$ for linear motion. We can see that the IMs move in the range of -6 mm to 6 mm with a cycle period of 0.15s. The shape of the curves of positions are similar with a very small difference. Motor currents are also similar in patterns though there is a small difference between them.
- Rotational motion: Fig. 8(a) shows the positions of $IM_1$ and $IM_2$, and Fig. 8(b) shows the currents of $LM_1$ and $LM_2$ for rotational motion. We can see that the two IMs move in the range of -6 mm to 6 mm in the opposite direction with a cycle period of 0.15s. Though the IMs are moving in the opposite direction the motor currents are similar in patterns as the magnitude of the accelerations are same.

### B. Comparison with Simulation

The parameters for the simulation of the capsubot is taken from the developed prototype and are listed in Table II.

Figs. 9(a)-9(d) and 10(a)-10(d) show the comparison between experimental and simulation results for the linear motion and rotational motion. For the linear motion both $IMs$ has the same acceleration profile. Thus comparison for $IM_1$ is shown only in Figs. 9(a)-9(d). For the rotational motion one of the $IMs$ follows the same acceleration profile as the linear motion and the other $IM$ follows an acceleration profile that is opposite to the previous one. Thus for the rotational motion comparison for $IM$ that has the opposite acceleration profile i.e. $IM_2$ is shown in Figs. 10(a)-10(d).

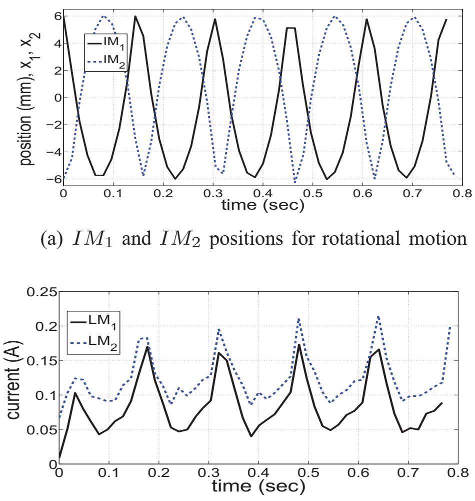
Although there are some differences between the experimental and simulation results, their trends are similar.



(a) $IM_1$ and $IM_2$ positions for linear motion

(b) Currents of $LM_1$ and $LM_2$ for linear motion

Fig. 7. Experimental results for linear motion



(a) $IM_1$ and $IM_2$ positions for rotational motion

(b) Currents of $LM_1$ and $LM_2$ for rotational motion

Fig. 8. Experimental results for rotational motion

The differences may be due to several reasons, such as motor dynamics, sensor dynamics, the other disturbances etc. which are not considered in the simulation model. We will investigate these issues further in future experiments.

From Figs. 9(d) and 10(d) we see that in the linear motion mode the capsubot moves with 8.4 mm/s average velocity. To move the capsubot in the opposite direction we just need to change the acceleration of the $IMs$ to the opposite direction. In the rotational motion the capsubot moves with 13 degrees/s average angular velocity. To rotate the capsubot in the opposite direction we need to swap the acceleration profiles between the IMs.

By using the linear motion and rotational motion alternately the capsubot can move in a 2D plane.

TABLE II

Parameter values of the Capsubot

| $m_1$, $m_2$ | $\mu_1$, $\mu_2$ | k | $w$ | $l$ | $h$ |
|---|---|---|---|---|---|
| $6.4gm$ | 0.2 | $6mm$ | $7cm$ | $8.7cm$ | $3.2cm$ |
| $g$ | $M$ | $F_{max}$ | $l_1$,$l_2$ | $\mu_r$ | $\mu$ |
| $9.8ms^{-2}$ | $42.9gm$ | $1.03N$ | $11.5mm$ | 0.08 | 0.28 |
| $F_{ipmax}$ | $F_{icmax}$ | | Linear | $b_{1\_1}$, $b_{1\_2}$ | $b_{2\_1}$, $b_{2\_2}$ |
| $2.74N$ | $1.03N$ | | Motion | $-20m/s^2$ | $5m/s^2$ |
| Rotational | $b_{1\_1}$ | $b_{1\_2}$ | $b_{2\_1}$ | $b_{2\_2}$ | |
| Motion | $-20m/s^2$ | $20m/s^2$ | $5m/s^2$ | $-5m/s^2$ | |

[]

## VI. Conclusions and Future Works

This paper has investigated the 2D capsubot from both simulation and experimentation. The paper has the following contributions: 1) proposed a control strategy for the motion control of underactuated mechanical systems which is an open problem till date; 2) validated the early proposed 2D capsubot design [20] through an initial experimentation; 3) implemented the closed-loop control strategy for the IMs of the 2D capsubot; 4) conducted both simulation and experimentation; 5) compared the experimental and simulation results for demonstrating the proposed capsule robot movability; 6) proposed measuring the position/velocity of a capsubot using video analysis software.

This has built our confidence to conduct further research along this line. We will conduct further experimental test using different control approaches. Also we will select/propose more realistic friction model and validate its performance in different environments.

## References

[1] P. Dario, B. Hannaford, and A. Menciassi, “Smart surgical tools and augmenting devices,” *IEEE Transactions on Robotics and Automation*, vol. 19, no. 5, pp. 782–792, 2003.

[2] N. Patronik, T. Ota, M. Zenati, and C. Riviere, “A miniature mobile robot for navigation and positioning on the beating heart,” *IEEE Transactions on Robotics*, vol. 25, no. 5, pp. 1109–1124, 2009.

[3] C. Briggs, C. Mann, G. Irving, C. Neal, M. Peterson, I. Cameron, and D. Berry, “Systematic review of minimally invasive pancreatic resection,” *Journal of Gastrointestinal Surgery*, vol. 13, no. 6, pp. 1129–1137, 2009.

[4] P. Gomes, “Surgical robotics: Reviewing the past, analysing the present, imagining the future,” *Robotics and Computer-Integrated Manufacturing*, vol. 27, no. 2, pp. 261–266, 2011.

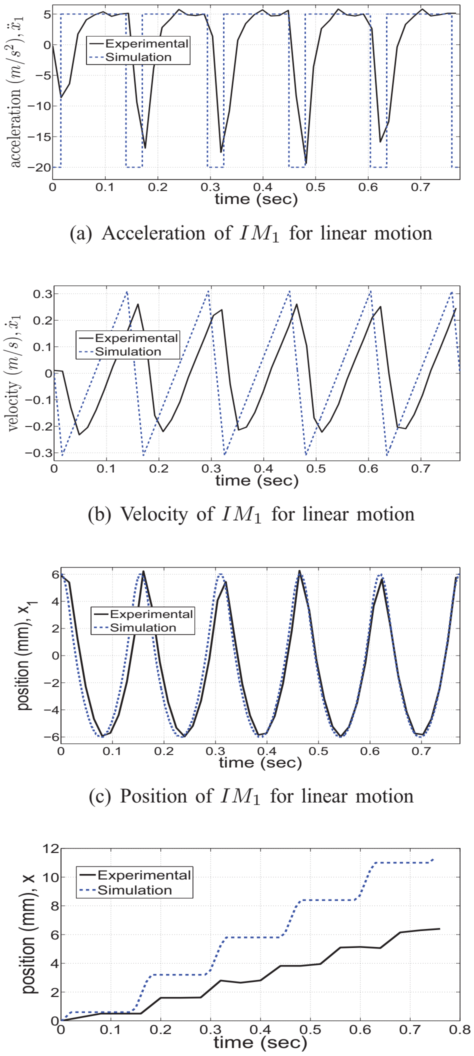
[5] C. Braumann, C. Jacobi, C. Menenakos, M. Ismail, J. Rueckert, and J. Mueller, “Robotic-assisted laparoscopic and thoracoscopic surgery with the da vinci system: a 4-year experience in a single institution,” *Surgical Laparoscopy Endoscopy & Percutaneous Techniques*, vol. 18, no. 3, p. 260, 2008.

[6] A. Lehman, K. Berg, J. Dumpert, N. Wood, A. Visty, M. Rentschler, S. Platt, S. Farritor, and D. Oleynikov, “Surgery with cooperative robots,” *Computer Aided Surgery*, vol. 13, no. 2, pp. 95–105, 2008.

[7] A. Lehman, M. Rentschler, S. Farritor, and D. Oleynikov, “The current state of miniature in vivo laparoscopic robotics,” *Journal of Robotic Surgery*, vol. 1, no. 1, pp. 45–49, 2007.

[8] S. Ohno, C. Hiroki, and W. Yu, “Design and manipulation of a suction-based micro robot for moving in the abdominal cavity,” *Advanced Robotics*, vol. 24, no. 12, pp. 1741–1761, 2010.

[9] M. Nokata, S. Kitamura, T. Nakagi, T. Inubushi, and S. Morikawa, “Capsule type medical robot with magnetic drive in abdominal cavity,” in *2nd IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics*, pp. 348–353, 2008.



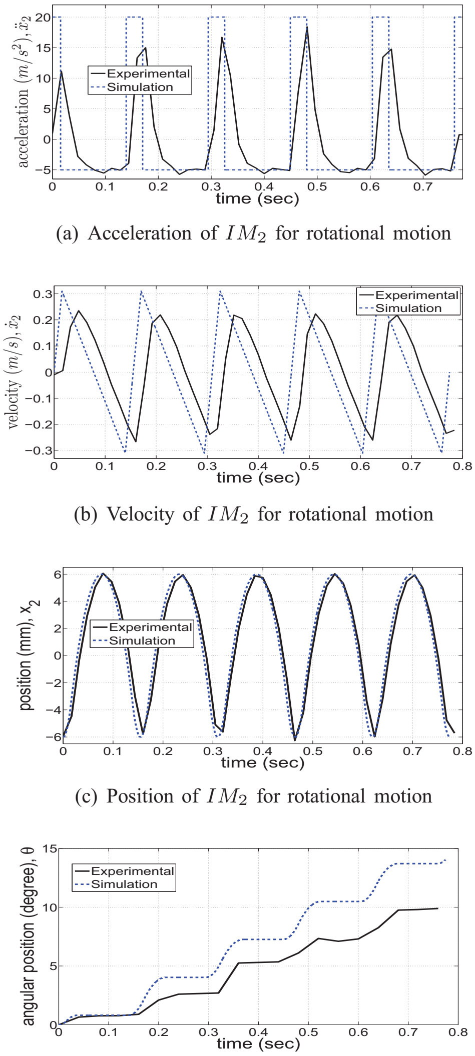
(a) Acceleration of $IM_1$ for linear motion

(b) Velocity of $IM_1$ for linear motion

(c) Position of $IM_1$ for linear motion

(d) Position of the capsubot for linear motion

Fig. 9. Comparison between experimental and simulation results for linear motion



(a) Acceleration of $IM_2$ for rotational motion

(b) Velocity of $IM_2$ for rotational motion

(c) Position of $IM_2$ for rotational motion

(d) Angular position of the capsubot for rotational motion

Fig. 10. Comparison between experimental and simulation results for rotational motion

[10] S. Platt, J. Hawks, and M. Rentschler, "Vision and task assistance using modular wireless in vivo surgical robots," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 6, pp. 1700–1710, 2009.

[11] P. Valdastri, R. Webster, C. Quaglia, M. Quirini, A. Menciassi, and P. Dario, "A new mechanism for mesoscale legged locomotion in compliant tubular environments," *IEEE Transactions on Robotics*, vol. 25, no. 5, pp. 1047–1057, 2009.

[12] H. Yu, M. N. Huda, and W. S. O., "A novel acceleration profile for the motion control of capsubots," in *IEEE International Conference on Robotics and Automation*, pp. 2437–2442, 2011.

[13] M. Simi, P. Valdastri, C. Quaglia, A. Menciassi, and P. Dario, "Design, fabrication, and testing of a capsule with hybrid locomotion for gastrointestinal tract exploration," *Mechatronics, IEEE/ASME Transactions on*, vol. 15, no. 2, pp. 170–180, 2010.

[14] H. Yu, Y. Liu, and T. Yang, "Closed-loop tracking control of a pendulum-driven cart-pole underactuated system," *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, vol. 222, no. 2, pp. 109–125, 2008.

[15] Y. Liu, H. Yu, and T. Yang, "Analysis and Control of a Capsubot," in *Proceedings of the 17th World Congress the International Federation of Automatic Control*, July 2008.

[16] Y. Yamagata and T. Higuchi, "A micropositioning device for precision automatic assembly using impact force of piezoelectric elements," in *1995 IEEE International Conference on Robotics and Automation*, vol. 1, pp. 666–671, May 1995.

[17] F. Chernous'ko, "The optimum rectilinear motion of a two-mass system," *Journal of applied Mathematics and Mechanics*, vol. 66, no. 1, pp. 1–7, 2002.

[18] N. Lee, N. Kamamichi, H. Li, and K. Furuta, "Control system design and experimental verification of capsubot," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1927–1932, 2008.

[19] G. Su, C. Zhang, R. Tan, and H. Li, "A linear driving mechanism applied to capsule robots," in *International Conference on Networking, Sensing and Control, 2009. ICNSC'09*, pp. 206–209, 2009.

[20] M. N. Huda and H. Yu, "Modelling and motion control of a novel double parallel mass capsubot," in *18th World Congress of the International Federation of Automatic Control (IFAC)*, pp. 8120–8125, 2011.

[21] Y. Liu, H. Yu, and S. Cang, "Modelling and motion control of a double-pendulum driven cart," *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, vol. 226, no. 2, pp. 175–187, 2012.

[22] S. MINIMOTOR SA, "http://www.faulhaber-group.com/." Online.

[23] Q. Biomechanics, "http://www.quintic.com/." Online.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# *Wind Turbine Sensor Fault Tolerant Control via a Multiple-Model Approach*

Montadher Sami and Ron J Patton

Department of Engineering, University of Hull, Hull HU6 7RX, UK

(e-mail: r.j.patton@hull.ac.uk; M.s.Shaker@2008.hull.ac.uk)

*Abstract*— **This paper presents a new strategy for wind turbine fault tolerant control (FTC) to optimise the wind energy captured by a wind turbine operating at low wind speeds. The FTC strategy uses Takagi-Sugeno (T-S) fuzzy observers with state feedback control to maintain nominal wind turbine control without changes in both the fault and fault-free cases. The proposed strategy obviates the need for sensor fault residual evaluation and observer switching by using a fuzzy proportional multiple integral observer (PMIO) to mask i.e. *'implicitly compensate'* the sensor fault(s) from the controller input and provide good estimation over a wide range of sensor fault scenarios. The proposed FTC method is applied to a 5 MW offshore wind turbine (OWT) benchmark model.**

## I. INTRODUCTION

Recently, FTC methods have stimulated research in a wide range of industrial control communities and academia, particularly for the systems that demand a high degree of reliability and availability (sustainability) and at the same time are characterised by expensive and/or safety critical maintenance work. The recently developed OWTs are foremost example for these systems having highly non-linear aerodynamics and with a stochastic and uncontrollable driving force as input in the form of wind speed. Moreover, the OWT site accessibility and system availability is not always ensured during or soon after malfunctions, primarily due to changing weather conditions. Indeed, maintenance work for OWTs is more expensive than the maintenance of onshore wind turbines by a factor of 5-10 times [1]. Hence, to be competitive with other energy sources, the main challenges for the deployment of wind turbine systems are to maximise the amount of good quality electrical power extracted from wind energy over a significantly wide range of weather conditions and minimise both manufacturing and maintenance costs.

In the literature, FTC and fault detection and diagnosis (FDD) systems have been recognised as the proper solution of ensuring the above mentioned requirements [2-6].

Specifically, in [5] the authors proposed linear parameter-varying (LPV) FTC systems for pitch actuator faults occurring in the full load operation with more emphasis on controller design rather than on fault estimation. In [6] fuzzy T-S observer based sensor FTC design is proposed to be capable of achieve maximization of wind power extraction. Their proposed FTC method is based on evaluation of *two* residual signals generated using the *generalised observer* idea of [7] to switch the estimation from faulty to healthy observers with the assumption that no simultaneous sensors faults are occur. It is clear that switching between two different observers produces unavoidable spikes that specifically affect the drive train torsion of low inertia wind turbines. Also the performance of the proposed FTC strategy is highly affected by the robustness and the computation time of the residual evaluation unit. Moreover, there is a significant probability of simultaneous occurrence of generator and rotor speed sensor faults [2].

This paper describes a new T-S fuzzy observer-based sensor FTC scheme designed to optimise the wind energy captured in the low wind speed range of operation. To cover a wider than usual range of sensor fault scenarios, the FTC strategy uses a fuzzy extension to the well-known PMIO [8] to provide simultaneous estimation of states and sensor faults. The nominal fuzzy controller remains unchanged during faulty and fault-free cases.

The main contributions of the paper are: (1) the use of the PMIO to mask or *implicitly compensate* the effect of drive train sensor faults, hence obviating the need for residual evaluation and observer switching. The PMIO simultaneously estimates the states and the sensor fault signals. Moreover, information about the fault severity can also be provided through the fault estimation signals; (2) the new fuzzy PMIO scheme is shown to cover a wide range of sensor fault scenarios [8].

The paper is organised as follows, Section 2 presents the flexible two mass OWT drive train model Section 3 describes the proposed FTC strategy. In Section 4 simulation results are presented showing the application of the FTC scheme to a 5MW wind turbine benchmark model. Section 5 gives a concluding statement.

## II. WIND TURBINE MODELLING

Normally, wind turbine models are obtained by combining the constituent subsystem models that make up the overall wind turbine dynamics. In this section a flexible low speed shaft, two mass wind turbine models are presented. The aerodynamic torque ($T_a$) representing the source of nonlinearity of the wind turbine. $T_a$ , depending on the rotor speed $\omega_r$, the blade pitch angle $\beta$ and the wind speed $v$ *is* given by:

$$T_a = 0.5\rho\pi R^2 C_p(\lambda,\beta)\frac{v^3}{\omega_r} \qquad (1)$$

where $\rho$ is the air density, $R$ is the radius of the rotor, and $C_p$ is the power coefficient that depends on the blade pitch angle ($\beta$) and the tip-speed-ratio ($\lambda$) (TSR) defined as:

$$\lambda = \omega_r R/v \qquad (2)$$

The drive train is responsible for gearing up the rotor speed to a higher generator rotational speed. The drive train model includes low and high speed shafts linked together by a

*Montadher Shaker acknowledges funding support for his PhD scholarship from the ministry of higher education and scientific research of Iraq.*

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

gearbox modelled as a gear ratio. The state space model of the wind turbine drive train has the form:

$$\begin{bmatrix}\dot{\omega}_r\\ \dot{\omega}_g\\ \dot{\theta}_\Delta\end{bmatrix} = \begin{bmatrix}a_{11} & a_{12} & a_{13}\\ a_{21} & a_{22} & a_{23}\\ a_{31} & a_{32} & a_{33}\end{bmatrix}\begin{bmatrix}\omega_r\\ \omega_g\\ \theta_\Delta\end{bmatrix} + \begin{bmatrix}b_{11} & 0\\ 0 & b_{22}\\ 0 & 0\end{bmatrix}\begin{bmatrix}T_a\\ T_g\end{bmatrix} \quad (3)$$

where:

$$a_{11} = -\frac{(B_{dt}+B_r)}{J_r} \qquad a_{12} = \frac{B_{dt}}{n_g J_r} \qquad a_{13} = -\frac{K_{dt}}{J_r}$$

$$a_{21} = \frac{B_{dt}}{n_g J_g} \qquad a_{22} = -\frac{(B_{dt}+n_g B_g)}{n_g^2 J_g} \qquad a_{21} = \frac{K_{dt}}{n_g J_g}$$

$$a_{31} = 1 \qquad a_{32} = -\frac{1}{n_g} \qquad a_{33} = 0$$

$$b_{11} = \frac{1}{J_r} \qquad b_{22} = -\frac{1}{J_g}$$

where $J_r$ is the rotor inertia, $B_r$ is the rotor external damping, $J_g$ is the generator inertia, $\omega_g$ and $T_g$ are the generator speed and torque, $B_g$ is the generator external damping, $n_g$ is the gearbox ratio, $K_{dt}$ is the torsion stiffness, $B_{dt}$ is the torsion damping coefficient, and $\theta_\Delta$ is the torsion angle.

It should be noted that the blade pitch subsystem model is not considered here as the pitch angle is held at the optimal pitch angle $\beta = 0$ value during low wind speed range to achieve maximum power extraction. The converter and generator subsystem is not considered here as the proposed controller presented here is designed as an outer controller to provide the inner generator controller with the required reference torque command. This approach to controller design separation is acceptable since the converter-generator subsystem is faster than the aerodynamic subsystem.

## III. THE PROPOSED STRATEGY

The nonlinear behaviour of the aerodynamic subsystem and its dependence on wind speed, it is decided to use the T-S fuzzy model based control strategy to design active sensor FTC.

This Section focuses on the description of the proposed active sensor FTC strategy for wind turbine power maximisation. Fig.1 schematically illustrates the proposed strategy.

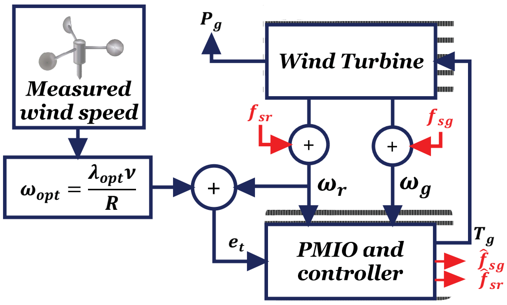


**Figure 1:** wind turbine sensor FTC scheme

The aim is to tolerate the effects of the drive train sensor fault. An estimator is used to estimate the fault signal and implicitly tolerate its effect on the state estimate signals delivered to the controller input. This strategy can be considered as a *fault-hiding* approach to FTC where the main objective of fault hiding is to maintain the same controller in both faulty and fault-free system cases.

The T-S fuzzy extension to the model given in (3) is:

$$\left.\begin{array}{l}\dot{x} = A(p)x(t) + \ Bu + E(p)v\\ y = \ Cx(t) + D_f f_s\end{array}\right\} \quad (4)$$

The system matrices $A(p) \in \mathcal{R}^{n*n}(= \sum_{i=1}^{r} h_i\,(p)A_i)$, $B \in \mathcal{R}^{n*m}$, $E(p) \in \mathcal{R}^{n*m_v}(= \sum_{i=1}^{r} h_i\,(p)E_i)$, $D_f \in \mathcal{R}^{l*g}$and $C \in \mathcal{R}^{l*n}$are known, $r$ is the number of fuzzy rules and the term $h_i\,(p)$ is the weighting function satisfying $\sum_{i=1}^{r} h_i\,(p) = 1$, and $1 \geq h_i\,(p) \geq 0$, for all $i$.

**Remark1:** The system given in Eq.(4) can be obtained from Eq.(3) by linearising the rotor aerodynamic torque equation. The central point to note in the T-S model given in Eq.(4) is that the system has common $B$ and $C$ matrices which will be utilized throughout the derivation of the proposed observer based controller. Details of linearization are given in [6, 9].

An augmented system consisting of the system (4) and the integral of the tracking error $e_t = \int (y_r - Sy)$ is defined as:

$$\left\{\begin{array}{l}\dot{\bar{x}} = \bar{A}(p)\bar{x} + \ \bar{B}u + \bar{E}(p)v + Ry_r\\ \bar{y} = \bar{C}\,\bar{x} + \bar{D}_f f_s\end{array}\right. \quad (5)$$

$$\bar{A}(p) = \begin{bmatrix}0 & -SC\\ 0 & A(p)\end{bmatrix}, \bar{x} = \begin{bmatrix}e_t\\ x\end{bmatrix}, \bar{B} = \begin{bmatrix}0\\ B\end{bmatrix}$$

$$\bar{E}(p) = \begin{bmatrix}0\\ E(p)\end{bmatrix}, R = \begin{bmatrix}I\\ 0\end{bmatrix}$$

$$\bar{C} = \begin{bmatrix}I_q & 0\\ 0 & C\end{bmatrix}, \bar{D}_f = \begin{bmatrix}0\\ D_f\end{bmatrix}$$

where $S \in \mathcal{R}^{w*l}$ is used to define which output variable is considered to track the reference signal.

Hence, the tracking problem is transferred to a fuzzy state feedback control, for which the proposed control signal is as:

$$u = K(p)\hat{\bar{x}} \quad (6)$$

where $K(p) \in \mathcal{R}^{m*(n+w)}(= \sum_{i=1}^{r} h_i\,(p)K_i)$ is the controller gain and $\hat{\bar{x}} \ \in \mathcal{R}^{(n+w)}$is the estimated augmented state vector.

If it is assumed that the $q^{th}$ derivative of the sensor fault signal is bounded, then an augmented state system consisting from the original local linear systems state and the $q^{th}$ derivative of the $f_s$, can be constructed.

Now let:

$$\varphi_i = f_s^{q-i} \qquad (i = 1,2, \dots, q)$$

$$\dot{\varphi}_1 = f_s^q; \dot{\varphi}_2 = \varphi_1;\ \dot{\varphi}_3 = \varphi_2; \dots; \dot{\varphi}_q = \varphi_{q-1}$$

Then the system (2) with augmented fault derivatives will become:

$$\left.\begin{array}{l}\dot{x}_a = A_a(p)x_a + B_a u + E_a(p)v + R_a y_r + Gf_s^q\\ y_a = C_a x_a\end{array}\right\} \quad (7)$$

where

$$x_a = \begin{bmatrix}\bar{x}^T & \varphi_1^T & \varphi_2^T & \varphi_3^T & \dots. & \varphi_q^T\end{bmatrix}^T \in \mathcal{R}^{\bar{n}}$$

$$A_a = \begin{bmatrix}\bar{A} & 0 & \dots & 0 & 0\\ 0 & 0 & \dots & 0 & 0\\ 0 & I & \dots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \dots & I & 0\end{bmatrix} \in \mathcal{R}^{\bar{n}\times\bar{n}}$$

$$B_a = [\bar{B}^T 0\ 0\ \dots 0]^T \ \in \mathcal{R}^{\bar{n}\times m}$$

$$G \ = [0\ I_k\ \ 0\ \dots 0]^T \ \in \mathcal{R}^{\bar{n}\times g}$$

$$C_a = \begin{bmatrix}\bar{C}\ 0\ \ 0\ \dots \bar{D}_f\end{bmatrix} \quad \in \mathcal{R}^{l*\bar{n}}$$

$$\bar{n} = (n + w) + gq$$

Hence, the following T-S fuzzy PMI observer is proposed to simultaneously estimate the system states and sensor faults:

$$\left.\begin{aligned}\dot{\hat{x}}_a &= A_a(p)\hat{x}_a + B_a u + E_a(p)v + R_a y_r + L_a(p)(y_a - \hat{y}_a)\\ \hat{y}_a &= C_a x_a\end{aligned}\right\} \qquad (8)$$

The state estimation error dynamics are obtained by subtracting Eq. (8) from Eq. (7) to yield:

$$\dot{e}_x = (A_a(p) - L_a(p)C_a)e_x + Gf_s^q \qquad (9)$$

The augmented system combining the augmented state space system (5), the controller (6), and the state estimation error (9) is given by:

$$\dot{\tilde{x}}_a(t) = \sum_{i=1}^{r} h_i\,(p)\{\tilde{A}_i\tilde{x}_a + \tilde{N}_i\tilde{d}\} \qquad (10)$$

where:

$$\tilde{A}_i = \begin{bmatrix}\bar{A}(p) + \bar{B}K(p) & -\bar{B}[K(p)\ 0_{m\times q}]\\ 0 & A_a(p) - L_a(p)C_a\end{bmatrix}$$

$$\tilde{x}_a = \begin{bmatrix}\bar{x}\\ e_x\end{bmatrix},\quad \tilde{N}_i = \begin{bmatrix}\bar{E}(p) & R & 0\\ 0 & 0 & G\end{bmatrix},\tilde{d} = \begin{bmatrix}d\\ y_r\\ f_s^q\end{bmatrix}$$

The objective here is to compute the gains $L_a(p)$ *and* $K(p)$ such that the effect of the input $\tilde{d}$ in Eq.(10) is attenuated below the desired level $\gamma$, to ensure robust stabilisation performance.

***Theorem1***: *For t>0 and* $h_i\,(p)h_j\,(p) \neq 0$, *The closed-loop fuzzy system in (6) is asymptotically stable and the $H_\infty$ performance is guaranteed with an attenuation level* $\gamma$, *provided that the signal* $(\tilde{d})$ *is bounded, if there exist symmetric positive definite matrices* $P_1, P_2$, *and matrices* $H_i, Y_i$, *and scalar* $\gamma$ *satisfying the following LMI constraints (11&12):*

$Minimise\ \gamma, such\ that$:

$$P_1 > 0, \qquad P_2 > 0 \qquad (11)$$

$$\begin{bmatrix}\Psi_{11} & \Psi_{12} & \bar{E}(p) & R & 0 & 0 & 0 & 0 & 0 & X_1C_p^T\\ * & -2\mu\bar{X}_1 & 0 & 0 & 0 & \mu I & 0 & 0 & 0 & 0\\ * & * & -2\mu I & 0 & 0 & 0 & \mu I & 0 & 0 & 0\\ * & * & * & -2\mu I & 0 & 0 & 0 & \mu I & 0 & 0\\ * & * & * & * & -2\mu I & 0 & 0 & 0 & \mu I & 0\\ * & * & * & * & * & \Psi_{55} & 0 & 0 & P_2G & 0\\ * & * & * & * & * & * & -\gamma I & 0 & 0 & 0\\ * & * & * & * & * & * & * & -\gamma I & 0 & 0\\ * & * & * & * & * & * & * & * & -\gamma I & 0\\ C_pX_1 & * & * & * & * & * & * & * & * & -\gamma I\end{bmatrix} < 0 \qquad (12)$$

where:

$K_i = Y_i X_1^{-1}$, $L_a = P_2^{-1}H_{ai}$ , $X_1 = P_1^{-1}, \bar{X}_1 = \begin{bmatrix}X_1 & 0\\ 0 & I_{q\times q}\end{bmatrix}$

$\Psi_{11} = \bar{A}_iX_1 + (\bar{A}_iX_1)^T + \bar{B}Y_i + (\bar{B}Y_i)^T$; $\Psi_{12} = [-\bar{B}Y_i \quad 0]$;

$\Psi_{55} = P_2A_{ai} + (P_2A_{ai})^T - H_{ai}C_a - (H_{ai}C_a)^T$.

***Proof:*** From Theorem 1, to achieve the performance and required closed-loop stability of (10) the following inequality must hold [10]:

$$\dot{v}(\tilde{x}_a) + \frac{1}{\gamma}\tilde{x}_a^T C_p^T C_p\tilde{x}_a - \gamma\tilde{d}^T\tilde{d} < 0 \qquad (13)$$

where $\dot{v}(\tilde{x}_a)$ is the time derivative of the candidate Lyapunov function $(v(\tilde{x}_a) = \tilde{x}_a^T\,\bar{P}\,\tilde{x}_a\ , where\ \bar{P} > 0)$ for the augmented system (10). Using Eq.(10), inequality (13) becomes:

$$\dot{v}(\tilde{x}_a) = \sum_{i=1}^{r} h_i\,\{\tilde{x}_a^T(\tilde{A}_i^T\bar{P}\ + \bar{P}\tilde{A}_i)\tilde{x}_a +\ \tilde{x}_a^T\bar{P}\tilde{N}_i\tilde{d} + \tilde{d}^T\tilde{N}_i^T\bar{P}\tilde{x}_a\} \qquad (14)$$

After simple manipulation, inequality (13) implies that the inequality (15) must hold:

$$\begin{bmatrix}\tilde{A}_{ij}^T\bar{P}\ + \bar{P}\tilde{A}_{ij} + \frac{1}{\gamma}I & \bar{P}\tilde{N}_{ij}\\ \tilde{N}_{ij}^T\bar{P} & -\gamma I\end{bmatrix} < 0 \qquad (15)$$

To be consistent with (10) $\bar{P}$ is structured as follows:

$$\bar{P} = \begin{bmatrix}P_1 & 0\\ 0 & P_2\end{bmatrix} > 0 \qquad (16)$$

Then after simple manipulation and using the variable change $(H_{ai} = P_2L_a(p))$ the inequality (15) can be reformulated as:

$$\Pi_{ij} = \left[\begin{array}{c:cccc}\Omega_{11} & -P_1[\bar{B}K_j\ 0] & P_1\bar{E}(p) & P_1R & 0\\ \hdashline * & \Omega_{22} & 0 & 0 & P_2G\\ * & * & -\gamma I & 0 & 0\\ * & * & * & -\gamma I & 0\\ 0 & (P_2G)^T & 0 & 0 & -\gamma I\end{array}\right] < 0 \qquad (17)$$

where:

$\Omega_{11} = \bar{A}_iX_1 + (\bar{A}_iX_1)^T + \bar{B}Y_i + (\bar{B}Y_i)^T + \frac{1}{\gamma}C_p^TC_p$

$\Omega_{22} = P_2A_{ai} + (P_2A_{ai})^T - \bar{H}_iC_a - (\bar{H}_iC_a)^T$

A single step design formulation of the matrix inequality in (17) is proposed to avoid the complexity of separate design steps characterised by repeated iteration to determine the gains required. Hence, $\Pi_{ij}$ as shown in (17) becomes:

$$\Pi_{ij} = \begin{bmatrix}\Pi_{11} & \Pi_{12}\\ * & \Pi_{22}\end{bmatrix} \qquad (18)$$

where

$\Pi_{11} = \Omega_{11}$ ; $\Pi_{12} = \begin{bmatrix}-P_1[\bar{B}K_j\ 0] & P_1\bar{E}(p) & P_1R & 0\end{bmatrix}$

$\Pi_{22}$ = lower right block

to do variable change, the following Lemma is required:

***Lemma 1*.** (Congruence) Consider two matrices $P$ and $Q$, if $P$ is positive definite and if $Q$ is a full column rank matrix, then the matrix $Q * P * Q^T$ is positive definite.

Let $Q = \begin{bmatrix} P_1^{-1} & 0 \\ 0 & X \end{bmatrix}$, $and\ X = \begin{bmatrix} \bar{X}_1 & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{bmatrix}$

Then $Q * \Pi_{ij} * Q^T < 0$ is also true and can be written as:

$$\begin{bmatrix} P_1^{-1}\Pi_{11}P_1^{-1} & P_1^{-1}\Pi_{12}X \\ * & X\Pi_{22}X \end{bmatrix} < 0 \tag{19}$$

Inequality (19) implies that $\Pi_{22} < 0$ so that the following inequality holds true [11, 12]:

$$(X + \mu\Pi_{22}^{-1})^T\Pi_{22}(X + \mu\Pi_{22}^{-1}) \leq 0 \Leftrightarrow X\Pi_{22}\mathrm{X} \leq -2\mu X - \mu^2\Pi_{22}^{-1} \tag{20}$$

where $\mu$ is a scalar.

By substituting (20) into (19) and using the Schur complement Theorem, then (19) holds if the following inequality holds:

$$\begin{bmatrix} P_1^{-1}\Pi_{11}P_1^{-1} & P_1^{-1}\Pi_{12}X & 0 \\ X\Pi_{12}P_1^{-1} & -2\mu X & \mu I \\ 0 & \mu I & \Pi_{22} \end{bmatrix} < 0 \tag{21}$$

After substitution for $\Pi_{11}, \Pi_{12}, \Pi_{12}, \Pi_{22}$ from (18) and by simple manipulation, the LMI in (12) is obtained. This completes the proof.

## IV. SIMULATION RESULTS

The simulation of the proposed T-S fuzzy observer based sensor FTC design is based on the wind turbine benchmark system described in [2]. The drive train subsystem is modelled by a two-mass system assuming a flexible low speed shaft. The model is implemented with band-limited measurements noise. The generator sensor faults are represented by two scale factor errors. The scale factors of 1.1 & 0.9 are multiplied by the simulated real generator rotational speeds. The expected fault effects would be a deviation of the wind turbine from the optimal operation.

**Remark2:** Without loss of generality, the output matrix parametric fault presented in wind turbine benchmark model [2] can be represented as an additive fault in which the fault signal depends on the measured state, as illustrated below:

$$y_f = C_f x = \begin{bmatrix} 1 & 0 \\ 0 & 0.9 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = Cx + \begin{bmatrix} 0 \\ 1 \end{bmatrix}(-0.1 * x_2) \tag{22}$$

Hence, parameter changes in the output matrix $C$ can be considered as a special case of additive faults in which the fault signal $(f_s)$ is a scaled version of the measured state.

**Remark3:** The sensor fault considered has a direct effect on the wind energy conversion efficiency since the controller starts to drive the wind turbine away from its optimal operation. Noticeably, in spite that this fault involves no damage risk, a study presented in [13] shows that a specific wind turbine operation in a 100 MW rated wind farm and operating with a realistic 35% capacity factor can generate about 307 GWh of energy in a given year. If the cost of the energy is \$0.04 per kWh, each GWh is worth about \$40,000, meaning that a 1% loss of energy on this wind farm corresponds to a loss of \$123,000 per year.

Fig. 2 shows the expected effects of the sensor bias fault on the power conversion efficiency. In fact, this sensor fault scenario simulates the effect of $C_p$ uncertainty encountered in the standard control law since in both cases the control signal drives the turbine away from the optimal trajectory.

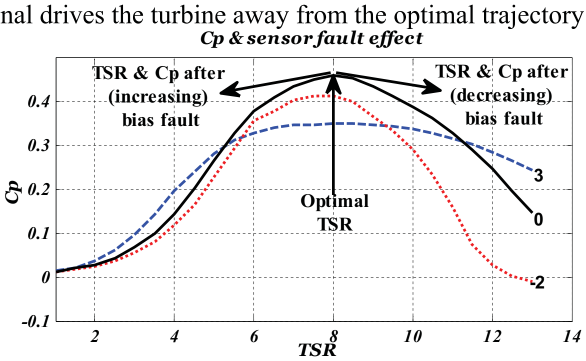


**Figure 2:** Effects of bias sensor fault on $C_p$ [$\boldsymbol{\beta}$=-2, 0, 3]

Fig. 3 shows how the wind turbine operation is affected by the two fault scenarios and helps to illustrate the success of the proposed strategy to tolerate the effects of the sensor faults, maintaining optimal wind turbine operation.

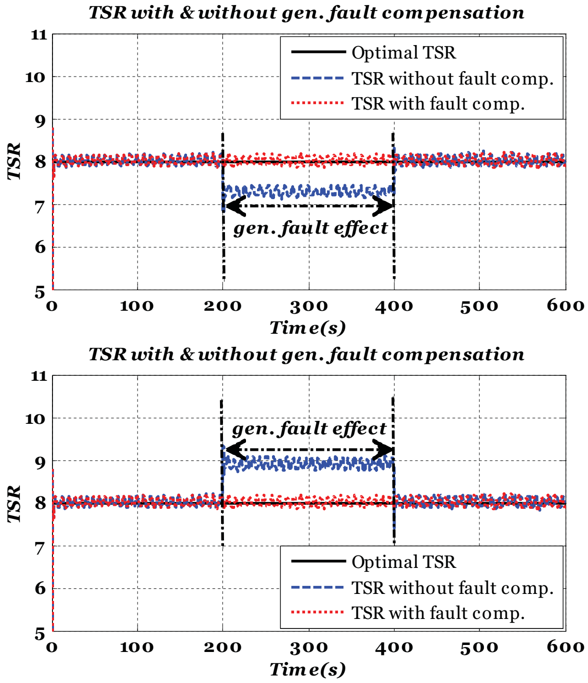


**Figure 3:** Effect of 1.1 (upper) and 0.9 (lower) sensor bias faults with(out) fault compensation

It is clear that the 1.1 bias sensor fault causes a deceleration of $\omega_r$ & $\omega_g$. Based on the faulty measurement the controller forces the turbine to reduce the rotational speed by increasing the reference generator torque which in turn increases the drive train load. This fault scenario is shown in Fig. 5 without sensor fault compensation.

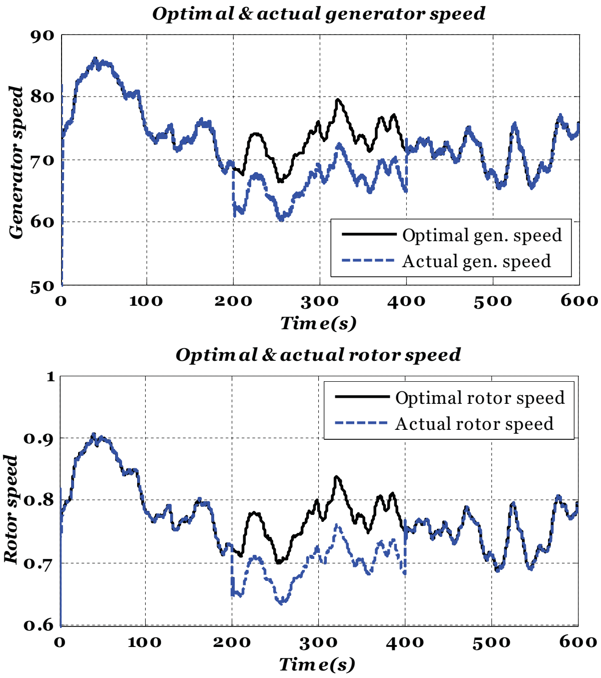


**Figure 4:** 1.1 sensor bias fault decelerate $\boldsymbol{\omega_r}$ & $\boldsymbol{\omega_g}$

On the other hand Fig. 6, shows the time variations of $\omega_r$ & $\omega_g$ in response to the proposed sensor FTC strategy.

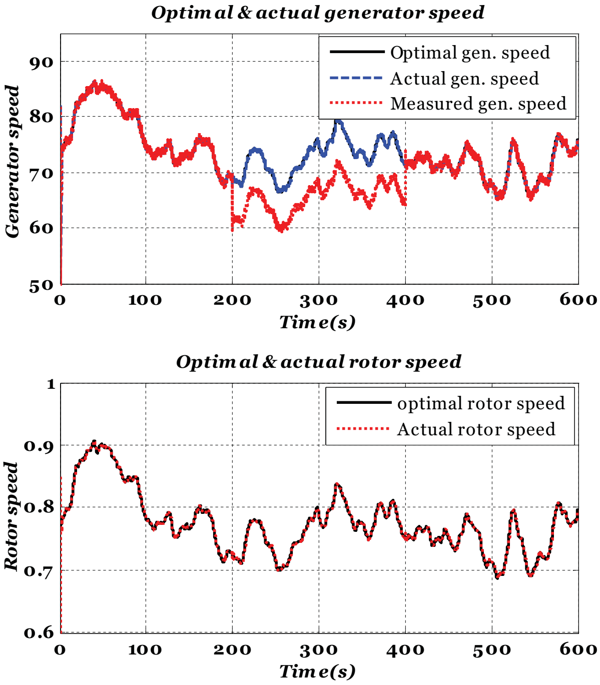


**Figure 5:** Actual and optimal $\boldsymbol{\omega_r}$ & $\boldsymbol{\omega_g}$ using the proposed sensor FTC strategy

Conversely, the 0.9 bias sensor fault causes acceleration of $\omega_r$ & $\omega_g$ since, based on faulty measurement; the controller releases the aerodynamic subsystem to rotate according to the available wind speed. Fig 7 shows the effect of the 0.9 sensor fault without compensation.

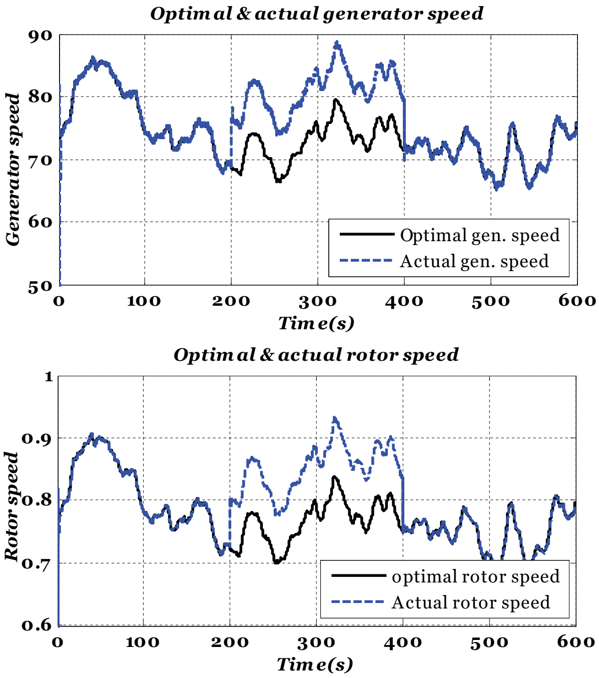


**Figure 6:** 0.9 sensor bias fault accelerate $\boldsymbol{\omega_r}$ & $\boldsymbol{\omega_g}$

The fault estimation signals for both sensor fault scenarios are shown in Fig.7.

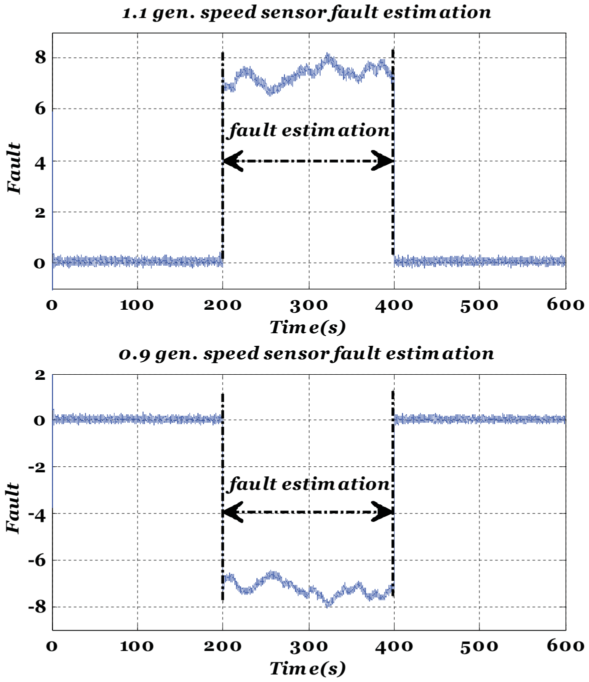


**Figure 7:** Estimation of 1.1 (upper) and 0.9 (lower) sensor bias faults

The use of the PMIO can also help to produce information about the severity of each fault. This is achieved through taking the ratio between the measured generator speed and the estimated signal. Hence, if there are no faults the ratio should be 1 otherwise any deviation from unity indicates the occurence of the fault and the magnitude of the deviation represents the fault severity. Fig. 8 shows the fault evaluation signal for both fault scenarios.

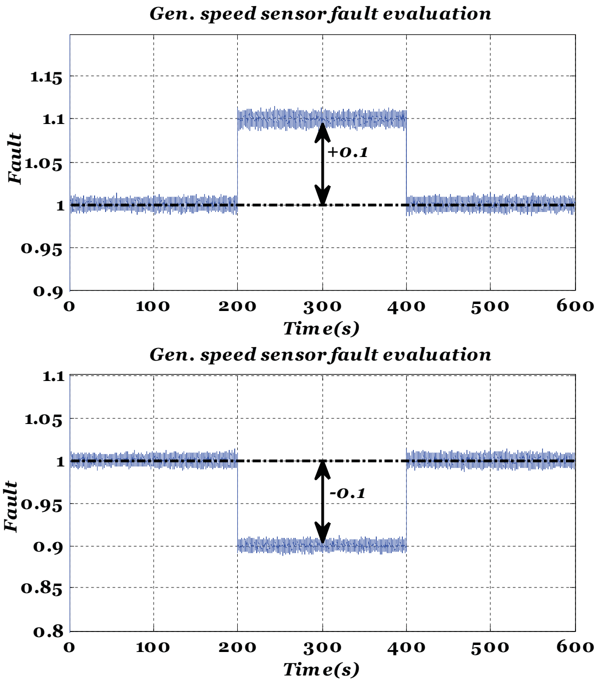


**Figure 8:** Deviation of 1.1 (upper) and 0.9 (lower) sensor measurement from unity

**Remark4:** Note that maintaining state estimation without changes during the whole range of operation is due to the fact that the PMIO perform *implicit fault estimation and compensation* of sensor fault from the input of PMIO. This fact is clearly interpreted from the error signal $(y_a - C_a\hat{x}_a)$ which can be rewritten as $(\bar{C}\bar{x} + \bar{D}f_s - \bar{C}\hat{\bar{x}} - \bar{D}\hat{f}_s)$, then as long as there are no sensor faults, $\hat{f}_s = 0$. However, once a sensor fault occurs the fault estimation $\hat{f}_s$ compensates the effect of the fault signal $f_s$ and hence the observer always receives a fault-free error signal.

## V. CONCLUSION

OWTs are complex and nonlinear systems that are driven by a stochastic and uncontrollable wind force and require a high degree of reliability and availability (sustainability). OWTs are also characterised by expensive and/or safety critical maintenance work. The main challenges for the deployment of wind turbine systems are to maximise the amount of good quality electrical power extracted from wind energy over a significantly wide range of weather conditions and minimise both manufacturing and maintenance costs. This paper has shown that active FTC can be of potential benefit as a very suitable solution for ensuring wind turbine reliability and sustainability requirements, particularly for offshore wind farms. This is illustrated through the use of the T-S extension to the well known PMIO fault estimation method in the observer based control strategy. However, OWTs are derived by uncontrollable signal in the form of effective wind speed which is not precisely measured due to the large vertical profile of blade swept area. Therefore, robustness of the FTC against this uncertain measurement is the direction of future work. Moreover, tolerate the effect of simultaneous actuator and sensor faults is the other direction of research in this field.

## REFERENCES

[1] G. J. W. van Bussel and M. B. Zaaijer, "Reliability, Availability and Maintenance Aspects of Large-Scale Offshore Wind Farms, a Concepts Study," in *MAREC 2001 Marine Renewable Energies Conference*, Newcastle, 2001, pp. 119-126.

[2] P. F. Odgaard, J. Stoustrup, and M. Kinnaert, "Fault Tolerant Control of Wind Turbines: a Benchmark Model," presented at the 7th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes *Safeprocess 2009*, Barcelona, Spain, 2009.

[3] Y. Amirat, M. E. H. Benbouzid, E. Al-Ahmar, B. Bensaker, and S. Turri, "A brief status on condition monitoring and fault diagnosis in wind energy conversion systems," *Renewable and Sustainable Energy Reviews,* vol. 13, pp. 2629-2636, 2009.

[4] Z. Hameed, Y. S. Hong, Y. M. Cho, S. H. Ahn, and C. K. Song, "Condition monitoring and fault detection of wind turbines and related algorithms: A review," *Renewable and Sustainable Energy Reviews,* vol. 13, pp. 1-39, 2009.

[5] C. Sloth, T. Esbensen, and J. Stoustrup, "Robust and fault-tolerant linear parameter-varying control of wind turbines," *Mechatronics,* vol. 21, pp. 645-659, 2011.

[6] E. Kamal, A. Aitouche, R. Ghorbani, and M. Bayart, "Robust Fuzzy Fault-Tolerant Control of Wind Energy Conversion Systems Subject to Sensor Faults," *IEEE Transactions on Sustainable Energy,* vol. 3, pp. 231-241, 2012.

[7] R. J. Patton, P. M. Frank, and R. N. Clark, *Fault Diagnosis in Dynamic Systems: Theory and Application*: Hemel Hempstead, UK, Prentice Hall International, 1989.

[8] Z. Gao, S. X. Ding, and Y. Ma, "Robust fault estimation approach and its application in vehicle lateral dynamic systems," *Optimal Control Applications and Methods,* vol. 28, pp. 143-156, 2007.

[9] T. Esbensen, B. T. Jensen, M. O. Niss, C. Sloth, and J. Stoustrup, "Joint Power and Speed Control of Wind Turbines," Aalborg University, Aalborg 120, 2008.

[10] S. X. Ding, *Model-based Fault Diagnosis Techniques Design Schemes, Algorithms, and Tools*: Springer-Verlag Berlin Heidelberg, 2008.

[11] T. M. Guerra, A. Kruszewski, L. Vermeiren, and H. Tirmant, "Conditions of output stabilization for nonlinear models in the Takagi-Sugeno's form," *Fuzzy Sets and Systems,* vol. 157, pp. 1248-1259, 2006.

[12] B. Mansouri, N. Manamanni, K. Guelton, and M. Djemai, "Robust pole placement controller design in LMI region for uncertain and disturbed switched systems," *Nonlinear Analysis: Hybrid Systems,* vol. 2, pp. 1136-1143, 2008.

[13] K. E. Johnson, "Adaptive torque control of variable speed wind turbines," NREL/TP-500-36265, National Renewable Energy Laboratory NREL/TP-500-36265, 2004.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Stability control of a railway vehicle using absolute stiffness and inerters

Alejandra Z. Matamoros-Sanchez,
Roger M. Goodall
and Argyrios C. Zolotas
School of Electronic, Electrical and
Systems Engineering. Loughborough University
Loughborough, United Kingdom LE11 3TU
Email: {A.Z.Matamoros-Sanchez,
R.M.Goodall, A.C.Zolotas}@lboro.ac.uk

Jason Zheng Jiang
and Malcolm C. Smith
Department of Engineering
University of Cambridge
Cambridge, United Kingdom CB2 1PZ
Email:{z.jiang, mcs}@eng.cam.ac.uk

***Abstract*—Work presented in this paper studies the potential of employing inerters –a novel mechanical device used successfully in racing cars– in active suspension configurations with the aim to enhance railway vehicle system performance. The particular element of research in this paper concerns railway wheelset lateral stability control. Controlled torques are applied to the wheelsets using the concept of absolute stiffness. The effects of a reduced set of arbitrary passive structures using springs, dampers and inerters integrated to the active solution are discussed. A multi-objective optimisation problem is defined for tuning the parameters of the proposed configurations. Finally, time domain simulations are assessed for the railway vehicle while negotiating a curved track. A simplification of the design problem for stability is attained with the integration of inerters to the active solutions.**

***Index Terms*—Railway vehicles, wheelsets stability control, absolute stiffness, active suspensions, inerter.**

## I. Introduction

A range of challenges is present in the design of modern railway vehicles, which are rather complex mechanical systems. Over the past thirty years, the use of sensors, electronic controllers, and actuators technology has enabled important changes in modern trains to deal with some of those difficulties hence enhancing vehicle system performance capabilities. In this context, the introduction of "active suspension" solutions, either in complementing or replacing conventional passive suspensions, did play a major role [1]. Conventional passive suspensions, via the vehicle geometry, both define important vehicle dynamics characteristics and are subject to operational constraints and limitations mainly due to the mechanical setup (e.g. combinations of springs, dampers, or pneumatic suspensions, and related mechanical linkages).

Admitting the role of passive suspensions as mechanical compensators of the vehicle's bodies dynamic interaction, a dominant factor in the overall performance results is in the type of structures that can be synthesised using simple elements like springs and dampers. Moreover, the nature of passive suspensions constrains the set of variables available for mechanical feedback. This highly determines the kind of resultant forces applied on the bodies and may cause some undesirable coupling between the system modes.

The use of active suspensions, i.e. usually the introduction of a controlled force-element (actuator), does offer enhanced suspension performance and some flexibility in the way the suspension structure evolves (actuator location and characteristics). However, this approach introduces the issue of energy consumption of the controlled element. Dynamic performance encompasses the problems associated to ride quality, dynamic stability and response to track features (e.g. curved sections of the rail track), for which active suspensions have been shown to facilitate higher speed and/or reduced track interaction [2]. Notwithstanding that active technologies cover a wide range of solutions for railway vehicles dynamic, passive compensation possibilities have grown with the introduction of the *inerter* [3]. The inerter is a novel concept and mechanical device with kinetic energy storage capabilities. This concept has been employed in automotive applications (being very successful in Formula 1 $^{\text{TM}}$ cars [4]), in motorcycles [5] and in building suspension control [6]. Also, there have been recent discussions on the use of inerters in pure passive railway suspension performance enhancement (see for example [7]).

This paper studies an approach that is different to the ones arising from current literature on inerters for railway vehicles. In particular, it investigates the performance merits of introducing the inerter together with the concept of active stability control of wheelsets via the absolute stiffness approach [8].

## II. Railway vehicle dynamics

Motions in a railway vehicle normally accept a certain degree of separation for suspension design and performance assessment purposes (for a typical railway vehicle schematic refer to Fig. 1). The study in this paper concerns wheelset stability control, and hence the modelling refers to only the lateral and yaw vehicle dynamic characteristics. Lateral and yaw motions are usually controlled by primary suspensions,

A. Z. Matamoros-Sanchez gratefully acknowledges the support of Universidad de Los Andes, Venezuela and Loughborough University, UK.
The work of J. Z. Jiang and M. C. Smith was supported by the Engineering and Physical Sciences Research Council grant number EP/G066477/1.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

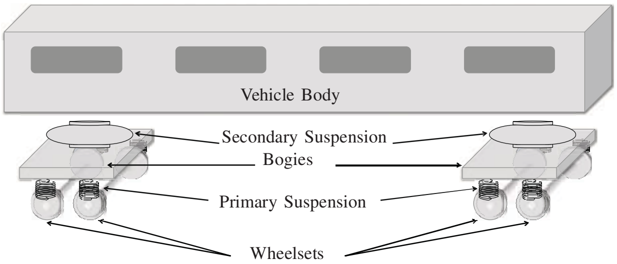


Fig. 1. Bodies and suspension stages of a railway vehicle.

with the main concern being the vehicle running behaviour at the wheel-rail interface, for which a design conflict between the vehicle stability and curving performance exists. Bruni et al. [2] presented a survey on the concepts and trends on active suspensions, where a detailed description of the control strategies for active primary suspensions is provided. They explain the essential kinematic instability of the solid axle wheelsets of the railway vehicle. Active solutions based on either lateral or yaw actuation controlled by feedback of yaw angles or lateral velocities re-locate the unstable poles of the wheelsets to the stable semi-plane. Yaw damping, both relative and absolute yaw stiffness, and lateral damping are some of the previously studied and widely known stabilising solutions in this field. The current study considers the application of relative yaw stiffness with the use of novel mechanical devices, in the search for complementing the active control strategy based on absolute yaw stiffness. The design of both structures is integrated so that they provide a synergetic solution.

Besides stability, the forces emerging from the wheel-rail interface when wheelsets negotiate curved tracks impose additional requirements to the design, albeit the control solutions for this aim are normally separated from the stability solutions. Steering control is dedicated to the ‘perfect curving’ [2], targeting: a) equal lateral creep forces on all the wheelsets to minimise track shifting forces, and b) zero steady-state longitudinal creep forces on all wheelsets. Some lateral creep is required in order to provide the necessary force to counterbalance the centrifugal force on curves [9], while longitudinal creep forces are associated to wear and noise and thus the requirement design is to significantly reduce them. Some steering solutions are also summarised in [2].

This paper assesses the curving behaviour as an outcome of the synergy between novel passive and active solutions designed for stability. Although some conditions in the overall design are established in regard to the steering problem, it is not defined as control objective for the active solution.

## III. Mathematical model of the system

The mathematical model approximating the lateral and inherent dynamics of a railway vehicle can be described by linear differential equations. The modelling is based upon the plan-view schematic representation presented in Fig.1. The use of linearised models is a normal practice in the design of active controllers, justified by the fact that active solutions reasonably overcome the effects of nonlinearities associated with the rail-wheel contact effects and other sources of nonlinear behaviour [10], in particular avoiding significant contact between the wheel flange and the rail. The mathematical description employed here for the assessment of a railway vehicle stability control corresponds to half of the vehicle comprising a two-axle bogie and half vehicle body. Primary and secondary suspensions in the vehicle provide the vertical linkages between the solid-axle wheelsets and the bogie frame, and the bogie frame and the vehicle body, respectively (Fig. 1). In the model, the lateral and longitudinal stiffness of the primary suspension are considered. Although the lateral primary damping is also modelled, the longitudinal is neglected. Likewise, the lateral stiffness and damping of the secondary suspension are included. The wheelsets consist of two wheels with conical profile rigidly joined together through a common axle, for which a nominal —ideally invariable— conicity coefficient, $\lambda$, and radius, $r_0$, are assumed. Rail/wheels interaction is taken into account through the longitudinal and lateral creep forces with ideally constant creep coefficients, $f_{11}$, and $f_{22}$. The creep forces arise at the contact points as consequence of the difference in strain rates of the wheels and the track (Fig. 2) –see Garg and Dukkipati [11] for an extended explanation. For every wheelset, those are given by:

$$F_x = 2 f_{11} l_{wy} \left( \frac{1}{R} - \frac{1}{V} \dot{\theta}_w \right) - \frac{2 f_{11} \lambda}{r_0} \left( y_w + y_t \right) \qquad (1)$$

$$F_y = -2 f_{22} \left( \frac{\dot{y}_w}{V} - \theta_w \right) \qquad (2)$$

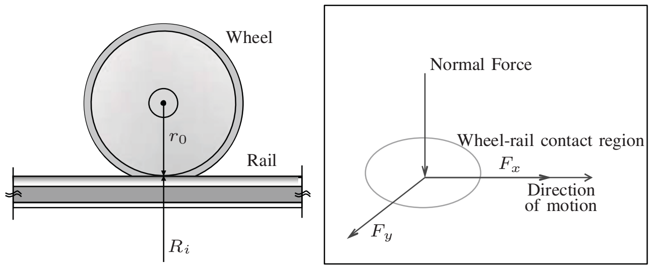


Fig. 2. Left: principal wheel and rail radii of curvatures. Right: Longitudinal and lateral creep forces.

Table I lists the parameters and the corresponding numerical values, while the schematic diagram of a wheelset is shown in Fig. 3. The differential equations (3)-(8) represent the dynamics of the seven degrees-of-freedom of the vehicle, namely: the lateral displacement and yaw angle of the wheelsets, respectively $y_{wi}$ and $\theta_{wi}$ for $i$=1, 2 (1: front, 2: rear), and the bogie, respectively $y_b$ and $\theta_b$, and also the vehicle body lateral motion $y_v$. This model with the default parameters values is equivalent to that used by Jiang et al. [12], except for the disregarded longitudinal primary damping and that the steering

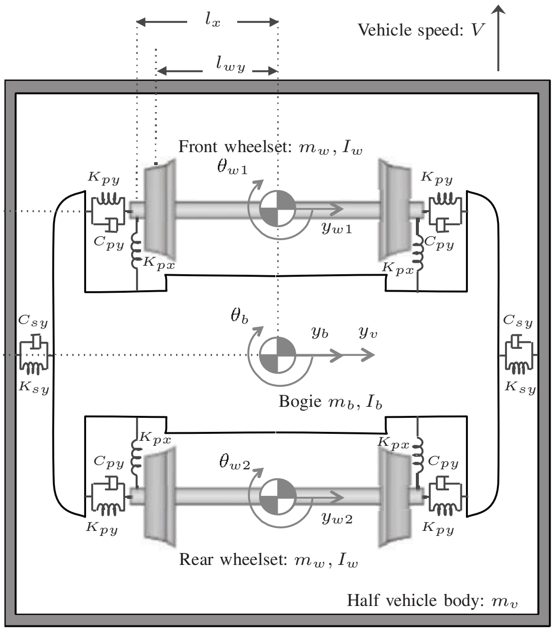


Fig. 3. Plan-view of the railway vehicle.

TABLE I
NOMINAL PARAMETERS VALUES OF THE HALF-VEHICLE MODEL

| Symbol | Description |
|---|---|
| $V$ | Vehicle speed ($55\text{ms}^{-1}$) |
| $m_w$ | Wheelset mass (1376 kg) |
| $I_w$ | Wheelset yaw inertia (766 kgm$^2$) |
| $m_g$ | Bogie frame mass (3477 kg) |
| $I_g$ | Bogie frame yaw inertia (3200 kgm$^2$) |
| $m_v$ | Half vehicle body mass (17230 kg ) |
| $r_0$ | Wheel radius (0.445 m) |
| $\lambda$ | Wheel conicity (0.3) |
| $f_{11}$ | Longitudinal creepage coefficient ($10^7$ N) |
| $f_{22}$ | Lateral creepage coefficient ($10^7$ N) |
| $l_{wx}$ | Semi-longitudinal spacing of wheelsets (1.225 m) |
| $l_{wy}$ | Half gauge of wheelset (0.75 m) |
| $l_x$ | Semi-lateral spacing of steering linkages and primary longitudinal suspension (1.2 m) |
| $K_{px}$ | Primary longitudinal stiffness per axle box ($10^6$ Nm$^{-1}$) |
| $K_{py}$ | Primary lateral stiffness per axle box ($4.71 \times 10^6$ Nm$^{-1}$ ) |
| $C_{py}$ | Primary lateral damping per axle box ($1.2 \times 10^4$ Nsm$^{-1}$ ) |
| $K_{sy}$ | Secondary lateral stiffness per axle box ($2.45 \times 10^5$ Nm$^{-1}$) |
| $C_{sy}$ | Secondary lateral damping per axle box ( $2 \times 10^4$ Nsm$^{-1}$) |
| $R_i$ | Radius of the curved track (1000 m) |
| $\theta_{c,i}$ | Cant angle of the curved track ( $6 \times \pi/180$ rad) |
| $g$ | Gravity (9.81ms$^{-2}$) |

linkage is not included.

$$\begin{aligned}
m_w\ddot{y}_{w,i} =& 2K_{py}\left(y_b + (-1)^{i-1}\, l_{wx}\theta_b - y_{w,i}\right) + \ F_{y,i} \\
& + 2C_{py}\left(\dot{y}_b + (-1)^{i-1}\, l_{wx}\dot{\theta}_b - \dot{y}_{w,i}\right) \qquad (3)\\
I_w\ddot{\theta}_{w,i} =& \ 2K_{px}l_x^2\left(\theta_b - \theta_{w,i}\right) + l_{wy}F_{x,i} \\
& + 2\, l_x^2 F_{pn,i} + \tau_{u,i} \qquad (4)\\
m_b\ddot{y}_b =& \ 2K_{py}\left(y_{w1} + y_{w2} - 2y_b\right) + 2K_{sy}\left(y_v - y_b\right) \\
& + 2C_{py}\left(\dot{y}_{w1} + \dot{y}_{w2} - 2\dot{y}_b\right) + 2C_{sy}\left(\dot{y}_v - \dot{y}_b\right) \\
& + \frac{m_b V^2}{2}\left(\frac{1}{R_1} + \frac{1}{R_2}\right) - m_b g\left(\frac{\theta_{c1} + \theta_{c2}}{2}\right) \qquad (5)\\
I_b\ddot{\theta}_b =& \ 2K_{py}l_{wx}\left(y_{w1} - y_{w2} - 2l_{wx}\theta_b\right) \\
& + 2K_{px}l_x^2\left(\theta_{w1} + \theta_{w2} - 2\theta_b\right) \qquad (6)\\
& + 2C_{py}l_{wx}\left(\dot{y}_{w1} - \dot{y}_{w2} - 2l_{wx}\dot{\theta}_b\right) \\
& - 2\, l_x^2\left(F_{pn1} + F_{pn2}\right) - \tau_{u1} - \tau_{u2} \qquad (7)\\
m_v\ddot{y}_v =& \ 2K_{sy}\left(y_b - y_v\right) + 2C_{sy}\left(\dot{y}_v - \dot{y}_v\right) \\
& + \frac{m_v V^2}{2}\left(\frac{1}{R_1} + \frac{1}{R_2}\right) - \frac{m_v g}{2}\left(\theta_{c1} + \theta_{c2}\right) \qquad (8)
\end{aligned}$$

It is noticeable from the model above that the suspensions in the longitudinal direction affect the wheelsets and bogie yaw mode only. This is explained by the symmetric allocation of those suspensions in the plane with respect to the centre of gravity of the bogie and every wheelset. That partly justifies the structure of the active control configuration exposed later.

For simulations and optimisation purposes, the state-space representation of this open-loop model was derived as

$$\begin{aligned}
\dot{x} =& A\,x + B\,u + B_\eta\,\eta \\
y_o =& C_o x \\
y_m =& C_m x \qquad (9)
\end{aligned}$$

with

$$\begin{aligned}
x =& [\dot{y}_{w1},\, y_{w1},\, \dot{\theta}_{w1},\, \theta_{w1},\, \dot{y}_{w2},\, y_{w2},\, \dot{\theta}_{w2},\, \theta_{w2},\, ... \\
& \dot{y}_b, y_b, \dot{\theta}_b, \theta_b, \dot{y}_v, y_v]^T \\
u =& [\tau_{u1},\, \tau_{u2},\, F_{pn1},\, F_{pn2})^T \\
\eta =& [{}^1\!/R_1,\, \theta_{c1},\, y_{t1},\, {}^1\!/R_2,\, \theta_{c2},\, y_{t2}]^T \\
y_o =& [F_{x1},\, F_{x2},\, F_{y1},\, F_{y2}]^T \\
y_m =& [\theta_{w1},\, \theta_{w2},\, \dot{\theta}_b - \dot{\theta}_{w1},\, \dot{\theta}_b - \dot{\theta}_{w2}]^T,
\end{aligned}$$

where $u$ is the vector of the active and passive control inputs applied to the system, $\eta$ are the exogenous inputs from the railway track, $y_o$ are the outputs related to the performance of the excited system, and $y_m$ are the variables/measurements to be fed back through the passive and active control configurations described in this paper.

## IV. ABSOLUTE STIFFNESS FOR STABILITY CONTROL

Highly stiff passive connections in the longitudinal direction between every axle and the bogie frame of a railway vehicle are known to provide the required levels of wheelset stability at high speeds on straight tracks [1]. However, since the bogie is relatively light, the two wheelsets and bogie become strongly coupled. The consequence is that an even higher longitudinal stiffness is necessary to achieve satisfactory stability levels, but because this stiffness also affects curving performance there is an increase in wheel and rail wear.

Mei and Goodall in [8] presented a more effective implementation for stiffness-based stabilisation in contrast with the conventional passive stiffness, the absolute stiffness *–skyhook stiffness* approach. This is an active configuration in which a longitudinal/yaw force/torque proportional to the high-pass filtered absolute wheelset yaw angle is applied between every wheelset and the bogie frame. The strategy decouples the wheelsets dynamic, preventing their stability being affected by the bogie dynamic and hence overcoming the complexity of the wheelsets stabilisation due to the wheelsets and bogie interaction. As pointed out by Mei and Goodall [8] the use of pure absolute yaw stiffness would cause problems for curving, so the high-pass filtering is necessary. In this manner absolute stiffness is an appealing solution to the stability problem.

The implementation of the strategy in this approach consists of the measurement of the wheelsets' yaw angles, which are filtered using individual second order high-pass filters of the same cut-off frequency, and are fed-back through a proportional controller to the respective actuators mounted between the wheelsets axles and the bogie frame applying the control torques, as represented in Fig. 4 and Eq. (10):

$$\hat{\tau}_{u,i}{=}-K_u \left( \frac{s^2/(2\pi f_c)^2}{s^2/(2\pi f_c)^2 + 1.414\, s/2\pi f_c + 1} \right) \hat{\theta}_{w,i} \qquad (10)$$

The controllable realisation of (10) is obtained in $A_\tau$, $B_\tau$, $C_\tau$, $D_\tau$ for a definition of the individual sub-systems (i.e. front—$i = 1$ and rear—$i = 2$ controllers) in the state-space as

$$\begin{aligned} \dot{x}_{\tau,i} &= A_\tau\, x_{\tau,i} + B_\tau\, \theta_{w,i} \\ \tau_{u,i} &= C_\tau\, x_{\tau,i} + D_\tau\, \theta_{w,i} \end{aligned} \qquad (11)$$

Variants of the implementation would be the estimation of the wheelsets yaw angle if the physical measurements are difficult to obtain, and/or to use linear actuators in the longitudinal direction instead of rotational ones. In this paper, the availability of the measurements and ideal actuators were supposed. Note, however, that Mei and Goodall [8] explain how a robust yaw gyro might effectively be used to derive the high-pass filtered yaw angle. Also in [8] Mei and Goodall point out the analogy between absolute stiffness and a "skyhook spring", i.e. a spring virtually connected between the wheelset and the "sky". As the authors noted in their paper [8], the analogy is not exact since in any practical implementation through a linear or rotational actuator there must be an equal and opposite reaction/force between the wheelset and bogie. The true reaction forces/torques are shown in Fig. 4 and implemented in the dynamic equations (3)-(8).

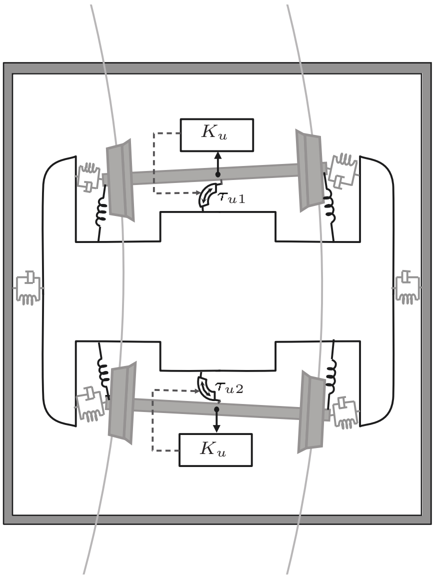


Fig. 4. Absolute stiffness stability control.

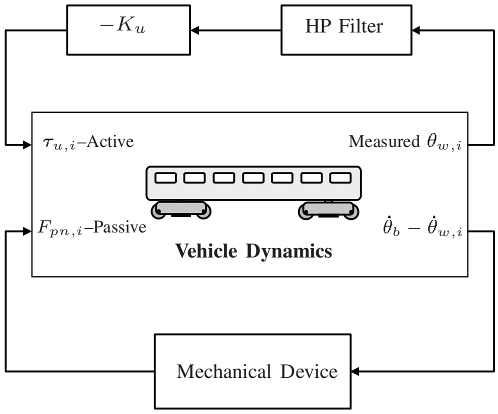


Fig. 5. Block diagram representation of absolute stiffness and novel mechanical devices integration.

## V. Absolute stiffness control and inerters

In this research, there is an interest on studying the potential of creating a synergy between passive and active solutions for railway vehicles dynamics, especially with the use of inerters, a recent concept in mechanical devices. Thereby, in combination with the active stability control strategy exposed above, different passive suspensions allocated in the longitudinal direction are proposed as represented in Fig. 5. The objective is to enhance the active solution through arbitrary passive networks, for which a cooperation between the two schemes is desirable to increase the possibilities of guaranteeing stability using absolute stiffness.

In the previous section, a description of two solutions for the bogie stability problem was provided, based on the use of passive and active stiffness. After having exposed that one of the major difficulties of applying either yaw or longitudinal stiffness alone between the wheelsets and the bogie is the response during curving, a frequency-based focus is provided

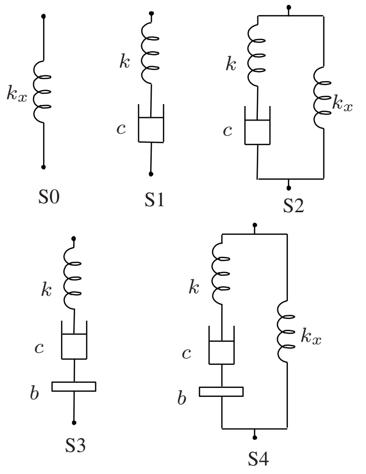


Fig. 6. Candidate suspension layouts.

in this section to introduce the role of more extensive passive suspensions.

In general, rail geometry characteristics in the lateral direction can be separated into high frequency for stochastic irregularities (predominant in straight tracks), and low frequency for curves. In fact, for the application of absolute stiffness it was already mentioned that high-pass filtering would alleviate the issues arising on curving. From the control perspective, the frequency response of mechanical suspensions is clearly a distinctive fact. For example, a linear spring has a flat frequency response to relative displacements between its terminals; this has a counterpart for the vehicle's behaviour which was identified earlier in this paper. In that sense, a non arbitrary range of passive networks can be designed using springs and dampers only to 'relax' the longitudinal stiffness at certain frequencies. However, the possibilities are reduced in the design of convenient configurations. A wider freedom on the design of passive suspensions using simple mechanical elements properties was given by the introduction of the inerter concept.

An inerter is a two-terminal mechanical device analogous to an ungrounded capacitor in an electric circuit, according to the force-current analogy, as described by Smith in [3]. It develops a linear force at its terminals which is proportional to the relative acceleration across them. It was termed by Chen et al. [4] to be 'the missing mechanical element' required to complete the referred analogy with electric networks using resistors, inductors and capacitors only, for which arbitrary passive impedances can be synthesised. In this manner, the inerter extends the possibilities in the synthesis of mechanical impedances and equivalently allows the formulation of arbitrary frequency characteristics with the unique constraint being on positivity requirements [3].

In this study, simple layouts using springs, dampers and inerters are examined. The candidate layouts basically consider the introduction of a novel mechanical device in the longitudinal plane with different stiffness according to the excitation frequency, augmenting the longitudinal stiffness already provided by the primary suspensions. Those are represented in Fig. 6, for which the parameter $b$ is the inertance property of the inerter, with units given in kilograms. $c$, $k$, and $k_x$ stand for normal damping and stiffness design parameters. The passive forces applied on the wheelsets and the bogie are characterised by the candidate layouts complex mechanical admittance $\mathrm{Y}_{pn}$(see [3]) as follows

TABLE II
COMPLEX ADMITTANCE OF THE CANDIDATE LAYOUTS IN FIG. 6

| Layout | Complex Admittance, $\mathrm{Y}_{pn}(s)$ |
|---|---|
| S0 | $\frac{k_x}{s}$ |
| S1 | $\frac{c}{c/k\ s+1}$ |
| S2 | $\frac{k_x}{s}+\frac{c}{c/k\ s+1}$ |
| S3 | $\frac{b\ s}{b/k\ s^2+b/c\ s+1}$ |
| S4 | $\frac{k_x}{s}+\frac{b\ s}{b/k\ s^2+b/c\ s+1}$ |

$$\hat{F}_{pn,i}=\mathrm{Y}_{pn}s\left(\theta_b-\theta_{w,i}\right) \tag{12}$$

The complex admittance of every layout is given in Table II. For every passive layout, the controllable canonical form was obtained in $A_{pn}$, $B_{pn}$, $C_{pn}$, $D_{pn}$, which is common to the two subsystems generating the longitudinal forces $F_{pn1}$ and $F_{pn2}$ in response to $\dot{\theta}_b-\dot{\theta}_{w1}$ and $\dot{\theta}_b-\dot{\theta}_{w2}$, respectively, and with states vectors $x_{pn1}$ and $x_{pn2}$ .

Although this reduced selection of passive structures does not allow to generalise on the effect of the inerter in passive longitudinal suspensions combined with active yaw control, it identifies the potential of the inerter in this proposed synergy.

## VI. Optimisation of parameters and performance assessment

Once the structure of the controllers has been established, both active and passive, a multi-objective optimisation problem was formulated on the closed-loop system represented in the state-space as

$$\begin{aligned}\dot{x}_{cl}&=\mathcal{A}x_{cl}+\mathcal{B}_\eta\eta\\ y_{cl}&=\mathcal{C}x_{cl}\end{aligned} \tag{13}$$

for the augmented state vector

$$x_{cl}=\left[x^T\middle|x_{\tau1}^T\middle|x_{\tau2}^T\middle|x_{pn1}^T\middle|x_{pn2}^T\right]^T, \tag{14}$$

and the output vector $y_{cl}=y_o$ re-written accordingly. The closed-loop system's matrices are defined as

$$\begin{aligned}\mathcal{A}&=\left[\begin{array}{c|c}A+B\,\Lambda_D^uC_m & B\,\Lambda_C^u\\ \hline \mathcal{A}_{21} & \Lambda_A^u\end{array}\right],\\ \mathcal{B}&=\left[\begin{array}{c|c|c|c}B_\eta^T & 0 & 0 & 0\end{array}\right]^T,\text{and}\\ \mathcal{C}&=\left[\begin{array}{c|c|c|c}C_o & 0 & 0 & 0\end{array}\right],\end{aligned}$$

with

$$\begin{aligned}\Lambda_A^u =& \text{diag}\left(A_\tau,\, A_\tau,\, A_{pn},\, A_{pn}\right)\\ \Lambda_C^u =& \text{diag}\left(C_\tau,\, C_\tau,\, C_{pn},\, C_{pn}\right)\\ \Lambda_D^u =& \text{diag}\left(D_\tau,\, D_\tau,\, D_{pn},\, D_{pn}\right),\end{aligned}$$

and

$$\mathcal{A}_{21}=\left[\, C^T_{m(1,*)}B^T_\tau \middle| C^T_{m(2,*)}B^T_\tau \middle| C^T_{m(3,*)}B^T_{pn} \middle| C^T_{m(4,*)}B^T_{pn} \,\right]^T.$$

The "zero" entries in the matrix definitions are zero vectors/matrices of appropriate dimension, and $C^T_{m(i,*)}$ corresponds to the $i$-th row of the matrix $C_m$ in (9).

In the stability problem devised for optimisation, two additional design objectives compromise the optimum stability: the minimisation of the active control gain and the minimisation of the fixed parallel stiffness $k_x$ for the layouts S0, S2 and S4. With this, one aims to determine whether or not the use of passive suspensions with inerters allows the reduction of the control gain and to improve the vehicle's performance while curving. The objective for an optimum stability was defined as the maximisation of the least damping ratio among the kinematic modes of the vehicle at a particular travelling speed of V=55ms$^{-1}$ (approx. 200km/h).

The Matlab® genetic algorithms (GA) toolbox was utilised to solve the trade-off emerging from this multiple goals formulation. The tuning parameters in the problem were: the control gain, $K_u$, the cut-off frequency of the high-pass filter, $f_c$, and the stiffness, damping and inertance parameters according to each individual candidate layout. Reasonable values for the physical constraints on the parameters were supplied to the algorithm. Because the optimum stability definition for the nominal speed does not necessarily account for lower speeds stability, a penalty was included in the stability cost function based on dismissing the parameter set for which the vehicle was unstable at the particular speed of 5ms$^{-1}$. Even though it does not avoid instability for all the speeds below the nominal, it does provide certain level of control on the stability along the evaluated range of speeds without significantly affecting the optimal tuning for the nominal speed. Indeed, with this procedure, levels of instability –possibly unrealistic– occur only for very low speeds which are transitional for a high-speed train running at normal conditions.

### A. Optimal stability

For the stability control design, the vehicle was considered to travel on a straight track, i.e. the curved track radius $R_i \rightarrow \infty$, and consequently also zero cant angle, $\theta_{c,i}$, for $i$=1, 2 (radius and cant angle measured at the front and rear wheelsets positions). The 3D plot in Fig. 9 shows the best population attained with genetic algorithms for those candidate layouts including $k_x$. It evidences the conflict between the three objectives defined in the study. The 3D plot reveals how the inerter simplifies the optimisation problem resulting with longitudinal stiffness values, $k_x$, between $0 - 2\times10^6$Nm$^{-1}$, and control gain values, $K_u$, between $(10-20)\times10^6$. Conversely, for the suspensions S0 and S1 the algorithm encountered different settings providing the same stability level with greater variability in the components parameters values. Decisions on the optimisation results are compiled in Table III, indicating also the best value for the least damping ratio, $\zeta_c$, at the design speed. The choice of the best set of parameters was arbitrarily done on the highest stability index value, $\zeta_c$. For those suspensions without the inerter the optimum set was chosen to be the one with the highest level of stability and lowest stiffness. For S3 and S4 the best stability resulted for a single set of parameters in every case, thus those were the chosen for comparisons. Because the values of $k_x$ and $K_u$ are reasonable, this method of choosing the best settings focused mostly on the degree of stability. Achievements are contrasted with the default passive configuration in Fig. 3 (equivalent to S0 with $k_x = 0$) with active stability control.

TABLE III
PARAMETERS VALUES FROM THE OPTIMISATION

| **Configuration** | $\zeta_c$ [%] | **Active Optimal parameters** | **Passive Optimal parameters** |
|---|---|---|---|
| Default Active | 20.6 | $K_u = 1.37\times10^7$<br>$f_c = 0.001$ Hz | N/A |
| Active and S0 | 20.6 | $K_u = 2.28\times10^7$<br>$f_c = 0.70$ Hz | $k_x = 6.39\times10^6$ Nm$^{-1}$ |
| Active and S1 | 20.7 | $K_u = 1.57\times10^7$<br>$f_c = 0.03$ Hz | $k_x = 0$,<br>$c = 6.42\times10^5$ Nsm$^{-1}$,<br>$k = 2.20\times10^6$ Nm$^{-1}$ |
| Active and S2 | 20.8 | $K_u = 4.94\times10^7$<br>$f_c = 0.76$ Hz | $k_x = 7.16\times10^6$ Nm$^{-1}$,<br>$c = 58.83\times10^5$ Nsm$^{-1}$,<br>$k = 2.55\times10^6$ Nm$^{-1}$ |
| Active and S3 | 21.2 | $K_u = 1.89\times10^7$<br>$f_c = 0.34$ Hz | $k_x = 0$,<br>$c = 6.13\times10^5$ Nsm$^{-1}$,<br>$b = 9.98\times10^4$ kg,<br>$k = 2.55\times10^6$ Nm$^{-1}$ |
| Active and S4 | 23.2 | $K_u = 1.91\times10^7$<br>$f_c = 0.28$ Hz | $k_x = 0.78\times10^6$ Nm$^{-1}$,<br>$c = 5.83\times10^5$ Nsm$^{-1}$,<br>$b = 6.84\times10^5$ kg,<br>$k = 2.77\times10^6$ Nm$^{-1}$ |

A classification of the candidate structures in Fig. 6 was done in two groups, A and B, for the presentation of the results: group A comprising novel devices without a fixed stiffness in parallel, i.e. S1 and S3, and group B consisting of novel devices with additional stiffness in parallel, i.e. S0, S2, and S4. The plots for $\zeta_c$ versus travelling speed with the optimum tuning obtained for the nominal speed $V$ are depicted in Fig. 7 and Fig. 8, providing also a comparison with the stability curve for the passive conventional system. The passive conventional setting is for a steering linkage stiffness of $k_x = 3.666\times10^7$ Nm$^{-1}$, guaranteeing the bogie stability, and $\tau_{u1}$ and $\tau_{u2}$ equal to zero.

From Fig. 7–8, improved stability achieved with all the configurations including active control can be observed. Particularly, those with inerters were found to provide, at some

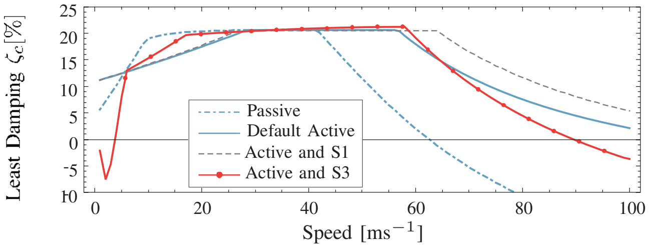


Fig. 7. Least damping ratio for the kinematic modes vs. speed (group A).

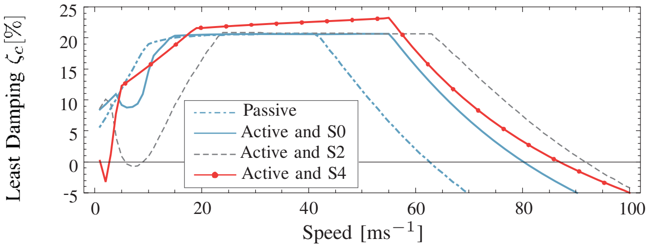


Fig. 8. Least damping ratio for the kinematic modes vs. speed (group B).

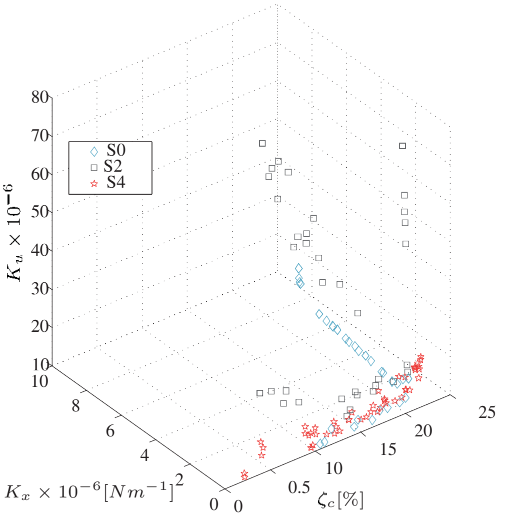


Fig. 9. Conflict 3D plot obtained from multi-objective optimisation for the active control combined with layouts in group B.

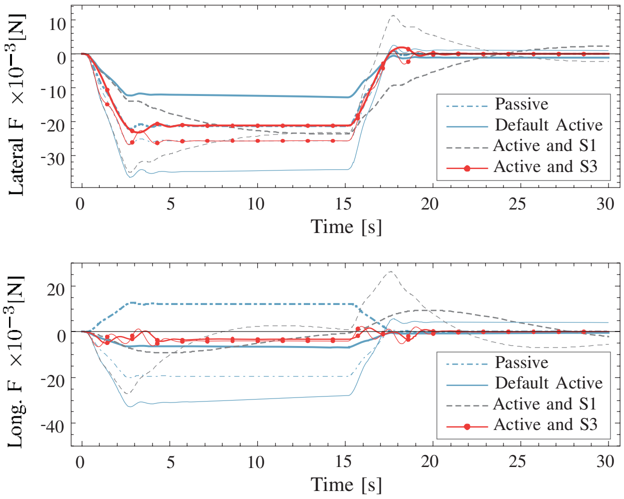


Fig. 10. Creep forces: Top- Lateral, Bottom- Longitudinal (group A).

extent, a better improvement on the vehicle's stability at the nominal speed over the other configurations: about 3% with S3, and 12.6% with S4 compared with achievements with S0. Also for low speeds better stability is attained in contrast with the configurations using the "relaxed suspensions" S1 and S3, albeit the irrelevant instability obtained for speeds below 5ms$^{-1}$. For speeds above the design speed, the other configurations result with better stability. In this regard, other exercises were done on optimising for higher speeds and the tendency is to obtain a slightly better stability with S3 and S4 for the design speed in contrast to the results for the structures without the inerter.

### B. Curving

With the optimum sets of parameters in Table III, curving behaviour in the time domain was also assessed. For this end, a track radius of 1000m and a cant angle of 6° were assumed, with a transition time of 3 seconds for a nominal travelling speed of V=55ms$^{-1}$ as in Jiang et al.'s paper [12].

The system excited by these transitional inputs develops longitudinal and lateral forces as displayed in Fig. 10 and Fig. 11, with applied control torques as shown in Fig. 12 and Fig. 13. In Figs. 10-13, bold curves correspond to the front wheelset responses, while normal curves are for the rear wheelset. If one refers back to the complex admittance of the layouts in Fig. 6 (i.e. Table II) and the parameter values from the optimisation (Table III), one can find that the strength of the layouts S3 and S4 is the better attenuation of the low frequency content of the relative displacements between wheelsets and bogie in the development of the resultant passive forces. S3 and S4, i.e. the structures with an inerter, provide an attenuation of +40dB/dec in contrast with the +20dB/dec provided by S1 and S2, and the all-pass characteristic of S0. It causes the longitudinal suspensions to behave even softer while curving, and with a flat stiffness for high frequency components, e.g. track irregularities. The filter effect of every suspension system is encountered at the following cut-off frequencies: 3.4 rads$^{-1}$ for S1, 0.4 rads$^{-1}$ for S2, 5.1 rads$^{-1}$ for S3, and 2 rads$^{-1}$ for S4. It reveals lowered cut-off frequencies for the structures with a fixed stiffness placed in parallel to the softer structures (or equivalently, frequency-dependent stiffness).

Results from the optimisation disclose that the active control technique based on absolute position of the wheelsets feedback can be further improved by feeding back the position of the wheelsets relative to the bogie position with adequate compensation. Analysing the simulations for the creep forces

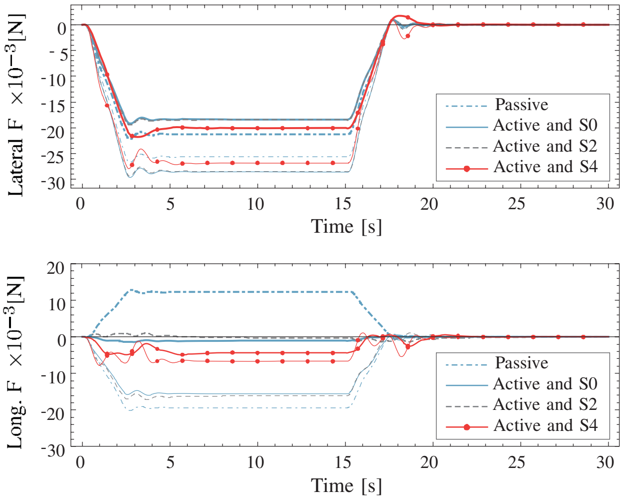


Fig. 11. Creep forces: Top- Lateral, Bottom- Longitudinal (group B).

and the active torques, the following benefits were obtained:

- a difference between the front and rear lateral creep forces for both S3 and S4 conveniently close to that obtained with the pure passive stabilisation configuration (for which $\zeta_c$ is approximately 7% only);
- a significant and favourable reduction of the front and rear longitudinal creep forces for both S3 and S4, when compared with those configurations without inerter, and
- lower active torque for S3, while higher for S4 –although perhaps better damped.

At higher nominal speeds, e.g. 83 ms$^{-1}$ (300km/h), suspensions as S1 and S3 will not guarantee a high degree of stability; results are not included here for brevity. In those cases, a fixed stiffness is needed for all the frequencies and therefore S2 and S4 are more appropriate. In fact, Fig. 10 and Fig. 12 show that even at 55 ms$^{-1}$, S1 is not adequate.

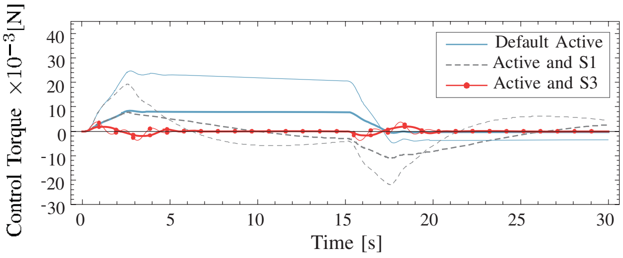


Fig. 12. Time-response for the applied control torque (group A).

## VII. Conclusion

Including inerters into the longitudinal suspensions simplifies the issues arising in stability control of a railway vehicle using absolute stiffness. A best compromise was achieved for the three objectives formulated: maximum stability for the nominal travelling speed, and reduction of the longitudinal stiffness and the control gain via GA optimisation. Time domain simulation results illustrated the usefulness of adding the inerter to the active control solution, i.e. some improvement in the wheelsets stability was attained while also longitudinal and lateral creep forces are reduced. This study presented possibilities of enhancing railway suspension behaviour via active control integrated with a novel mechanical element, the inerter.

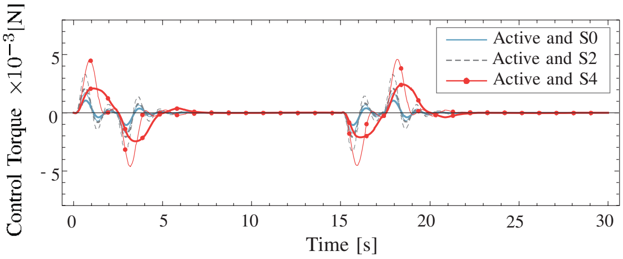


Fig. 13. Time-response for the applied control torque (group B).

## References

[1] S. Iwnicki, *Handbook of railway vehicle dynamics.* United States of America: CRC Press, 2006.

[2] S. Bruni, R. Goodall, T. X. Mei, and H. Tsunashima, "Control and monitoring for railway vehicle dynamics," *Vehicle System Dynamics*, vol. 45, no. 7-8, p. 743, 2007.

[3] M. C. Smith, "Synthesis of mechanical networks: The inerter," *IEEE Transactions on Automatic Control*, vol. 47, no. 10, p. 1648, 2002.

[4] M. Z. Q. Chen, C. Papageorgiou, F. Scheibe, F. cheng Wang, and M. C. Smith, "The missing mechanical circuit element," *Circuits and Systems Magazine, IEEE*, vol. 9, no. 1, pp. 10–26, 2009.

[5] S. Evangelou, D. J. N. Limebeer, R. S. Sharp, and M. C. Smith, "Mechanical steering compensators for high-performance motorcycles," *Journal of Applied Mechanics- Transactions of the ASME*, vol. 74, no. 2, p. 332, 2007.

[6] F. C. Wang, C.-W. Chen, M.-K. Liao, and M.-F. Hong, "Performance analyses of building suspension control with inerters," in *Decision and Control, 2007 46th IEEE Conference on*, 2007, p. 3786.

[7] F. C. Wang and M.-K. Liao, "The lateral stability of train suspension systems employing inerters," *Vehicle System Dynamics.*, vol. 48, no. 5, p. 619, 2009.

[8] T. X. Mei and R. M. Goodall, "Stability control of railway bogies using absolute stiffness: sky-hook spring approach," *Vehicle System Dynamics*, vol. 44, no. sup1, p. 83, 2006.

[9] S. Shen, T. X. Mei, R. M. Goodall, and J. T. Pearson, "A novel control strategy for active steering of railway bogies," in *UKACC Control 2004*, 2004.

[10] T. X. Mei and R. M. Goodall, "Recent development in active steering of railway vehicles," *Vehicle System Dynamics*, vol. 39, no. 6, p. 415, 2003.

[11] V. K. Garg and R. V. Dukkipati, *Dynamics of Railway Vehicle Systems.* United Kingdom: Accademic Press, 1984.

[12] J. Z. Jiang, A. Z. Matamoros-Sanchez, R. M. Goodall, and M. C. Smith, "Passive suspensions incorporating inerters for railway vehicles," *Vehicle System Dynamics*, 2012, in press.

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# A Multiobjective Trajectory Optimisation Method for Planning Environmentally Efficient Trajectories

Quintain McEnteggart
Cranfield University
Cranfield
Bedfordshire, MK43 0AL
Email: q.mcenteggart@cranfield.ac.uk

James Whidborne
Cranfield University
Cranfield
Bedfordshire, MK43 0AL
Email : j.f.whidborne@cranfield.ac.uk

*Abstract*—**This paper proposes a multi-objective trajectory optimization method to be used in the planning of environmentally efficient commercial aircraft trajectories. The problem of finding environmentally efficient trajectories is treated as an optimal control problem that is solved by applying a direct method of trajectory optimisation. The method involves an inverse trajectory parameterisation technique, methods for the calculation of environmental objectives, and the use of a multi-objective version of the Differential Evolution algorithm. The principal benefit of the method relative to previous work is that it allows the fast generation of Pareto optimal fronts between several competing objectives. This allows the decision maker to make informed decisions about potential trade-offs between different environmental goals.**

## I. Introduction

Over the last decade global passenger air traffic has increased by more than 45%, with similar levels of growth projected for the coming decade [1]. It has long been recognised that aviation benefits society, as a generator of wealth, and as an enabler to the exchange of ideas and culture between nations. However, increasing levels of air traffic will continue to impact the environment, with rising levels of aircraft emissions and higher numbers of people exposed to significant levels of aircraft noise [2].

The Advisory Council for Aviation Research in Europe (ACARE) is a group of representatives from the European commission, member states, industry and academia tasked with influencing the direction of European aviation research and development. The council, recognising the impact of aviation on the environment, has proposed to meet the challenges of sustainable aviation through the application of research and technology. To achieve this, it has created a strategic agenda for European aviation research with associated goals to be achieved by the year 2020. The environmental targets proposed by ACARE are (from a 2000 baseline) [3]

- Reduce fuel consumption and carbon dioxide (CO2) emissions by 50% per passenger kilometre.
- Reduce NOx emissions by 80%.
- Reduce perceived noise by 50% .

Of the targets, ACARE has proposed in its roadmap that 5-10% of the CO2 reduction be achieved through improved aircraft operations and air traffic management. This target is in line with the Single European Sky Air Traffic Management Research (SESAR) programme goal of reducing Air Traffic Management (ATM) related CO2 emissions by 10% per flight (from a 2005 baseline) [4]. Although no quantitative targets have been set for ATM related noise reduction, both initiatives recognise the role that improving aircraft operations has to play in reducing noise impact on communities around airports.

A particular focus of Air Traffic Management research related to delivering the ACARE goals is the design and delivery of trajectories that minimise environmental impact, referred to as green trajectories. Green trajectories, in the form of Continuous Descent Approaches (CDA), have already shown promise in delivering emissions and noise reductions to airport terminal areas [5]. The planning and optimization of green trajectories has been the subject of a number of theoretical studies. The Sourdine project, using expert analysis, developed a series of recommended noise abatement procedures for airport arrival and departures [6]. The trajectories were optimized to reduce noise under the flight-path for representative medium narrow-body and large wide-body aircraft. Simulations subsequently showed that large scale adoption of the Sourdine procedures could lead to significant reductions in noise footprints relative to conventional trajectories, although runway rates were adversely affected [7], [8]. Visser et al [9], [10] posed the problems of finding green arrival and departure trajectories as optimal control problems. This work used the direct collocation technique proposed by Hargreaves and Paris [11] to calculate optimal trajectories in terms of fuel burn and awakenings. Hebly et al [12] also used a collocation method and a weighted-sum cost function based on fuel burn and awakenings to optimise a Required Navigation (RNAV) departure procedure. Prats et al [13], again using a collocation method, recognised that the calculation of green trajectories can require the consideration of several conflicting optimisation criteria. To account for this, a lexiographic method for multi-objective optimisation was implemented. For the method, a hierarchy of importance for the objectives was established prior to the simulation. In this case, the method finds the minimum for the first objective and then seeks reductions in subsequent objectives providing

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

they do not increase the values of the objectives higher in the hierarchy. What results is a single Pareto optimal point that lies at the extreme of the first objective in the hierarchy.

When multiple conflicting objectives exist for an optimization problem, a single Pareto optimal point, or the minimum of a single objective, offer the Decision Maker (DM) very little information about potential tradeoffs between the optimization objectives. In environmental trajectory optimization, it is usually desirable to have multiple Pareto optimal points, where the tradeoffs in objectives, such as noise and fuel burn, can be assessed relative to each other.

The aim of this paper then is to solve the general multi-objective trajectory optimization problem for several types of environmental objectives, and to find the Pareto optimal set between minimised objectives. To do this, it is proposed to convert the optimal control problem to a Non Linear Programming (NLP) problem using a direct method of trajectory optimization that can be combined with aircraft emission and noise methods for the calculation of objectives. It is then proposed to solve the NLP problem for a Pareto optimal set using a stochastic evolutionary algorithm as the NLP solver. It is intended that the proposed method be useful to air traffic route designers and to airline flight planners as an approach that can be used to predict and optimize the environmental impact of commercial aircraft trajectory operations.

## II. Problem

Stated generally, the multi-objective trajectory optimization problem can be stated as the problem of minimising an array of scalar objective functions

$$\min_{\mathbf{z},\mathbf{u}}[f_1(\mathbf{z},\mathbf{u}), f_2(\mathbf{z},\mathbf{u}), \ldots, f_q(\mathbf{z},\mathbf{u})]^T \tag{1}$$

where, in this work, the aircraft states are $\mathbf{z} = [x(t), y(t), h(t), v(t), \gamma(t), \chi(t)]^T$, the aircraft position is $\mathbf{r} = [x(t), y(t), h(t)]^T$, $v(t)$ is the airspeed, $\gamma(t)$ is the flight path angle and where $\chi(t)$ is the heading angle. The aircraft controls are then $\mathbf{u} = [T(t), n(t), \phi(t)]^T$ where $T(t)$ is thrust, $n(t)$ is load factor and $\phi(t)$ is the bank angle. The individual objectives of the array are then minimised subject to the inequality constraints to be satisfied

$$\mathbf{c}_i(\mathbf{z},\mathbf{u}) \leq 0 \tag{2}$$

When solving a multiobjective optimization problem, where n is the number of objectives, the objective vector $\mathbf{s}^*$ is Pareto optimal, if there does not exist another objective vector $\mathbf{s} \in S$ where $s_i \leq s_i^*$ for all $i = \{1, \ldots, n\}$ and where $s_j^* < s_j$ for at least one index of $j$, $j \in \{1, \ldots, n\}$. A Pareto front is a set of Pareto optimal solutions. A set is Pareto optimal if each solution in the set is Pareto optimal. For conflicting objectives, the Pareto optimal set should identify the extremes of the objectives, most fuel efficient and shortest path for instance, and allow for tradeoffs between objectives to be examined. This facilitates the search for solutions that offer the best balance between objectives.

## III. Transcription

To solve the general optimal control problem it may be converted to a finite dimensional numerical optimization problem. Methods for transcription to a numerical problem are classified in Betts [14] as indirect and direct. Indirect methods involve forming the Hamiltonian of the system, estimating the costate variables, and finding a root of the two point boundary value problem. Direct methods involve discretising the optimal control problem and solving for the states and controls at a series of dividing nodes. Direct methods are preferred for this work because they do not require the definition of costate variables or constrained arcs and are therefore easier to apply to the problem.

Direct collocation methods use polynomials to parameterise the states and controls of the aircraft. Hargreaves and Paris [11] used piecewise cubic polynomials, while Fahroo et al [15], [16] and Benson et al [17] have used orthogonal polynomials with several different forms of collocation points. A significant drawback of collocation methods however, is that they can require large numbers of varied parameters. When applied to problems with large numbers of Air Traffic Control (ATC) constraints on aircraft position, speeds and rates of climb/descent, collocation methods may be very slow to evolve, first from infeasible to feasible solutions and then from feasible to a point along the global Pareto front. Yakimenko proposed an inverse method where the position states of the aircraft and their derivatives are parameterised using 7th degree polynomials [18]. Controls are then determined by inverting the state equations. The method significantly reduces the number of optimization variables required by analytically determining the polynomial coefficients from the prescribed states and controls at the boundaries ($t = 0$ and $t = t_f$). Instead of parameterising by time, Yakimenko adopted Taranenko's method of parameterising the polynomials by $\tau$, creating a virtual arc $\tau \in [0, \tau_f]$. The relationship between time $t$ and $\tau$ is defined as $\lambda = d\tau/dt$. The use of the relationship parameter $\lambda$ allows the definition of aircraft velocity using a separate reference function, enabling the creation of a virtual speed profile along the trajectory path of the aircraft. The method, termed the Inverse Dynamics in the Virtual Domain (IDVD) method, is a fast trajectory optimization method and has been considered for real time implementation [18], [19]. The IDVD method was adopted for this work because the small number of varied parameters allowed the Differential Evolution (DE) algorithm used in this work to quickly evolve decision vectors through the differential mutation and crossover mechanisms. The DE algorithm was chosen because it had the potential, when combined with the IDVD method, to efficiently converge on detailed global Pareto fronts.

For the IDVD method, the flat earth Cartesian coordinates $r_j (j = 1, 2, 3)$ and their derivatives are parameterised from the reference function (3) and its derivatives,

$$r(\tau)_j = \sum_{k=0}^{7} \frac{a_j k \tau^k}{\max(1, k(k-1))} \tag{3}$$

The coefficients of the polynomials are determined analytically from the coordinates and their derivatives at the boundaries ($\tau = 0$ and $\tau = \tau_f$) by making the coefficients the subjects of the following set of linear equations:

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & \tau_f & \frac{\tau_f^2}{2} & \frac{\tau_f^3}{6} & \frac{\tau_f^4}{12} & \frac{\tau_f^5}{20} & \frac{\tau_f^6}{30} & \frac{\tau_f^7}{42} \\ 0 & 1 & \tau_f & \frac{\tau_f^2}{2} & \frac{\tau_f^3}{3} & \frac{\tau_f^4}{4} & \frac{\tau_f^5}{5} & \frac{\tau_f^6}{6} \\ 0 & 0 & 1 & \tau_f & \tau_f^2 & \tau_f^3 & \tau_f^4 & \tau_f^5 \\ 0 & 0 & 0 & 1 & 2\tau & 3\tau^2 & 4\tau^3 & 5\tau^4 \end{bmatrix} \begin{bmatrix} a_{j0} \\ a_{j1} \\ a_{j2} \\ a_{j3} \\ a_{j4} \\ a_{j5} \\ a_{j6} \\ a_{j7} \end{bmatrix} = \begin{bmatrix} r_{j0} \\ r'_{j0} \\ r''_{j0} \\ r'''_{j0} \\ r_{jf} \\ r'_{jf} \\ r''_{jf} \\ r'''_{jf} \end{bmatrix} \tag{4}$$

To modify the polynomials, the variables iterated by the solver are then $\Xi = [x'''_{0,f}, y'''_{0,f}, h'''_{0,f}, v''_{0,f}, \tau_f]$. To transform the polynomials to the system dynamics, a point mass model is used. Therefore the state and controls are determined by inverting the following equations:

$$\begin{aligned} \dot{x} &= v \cos\gamma \cos\chi, & \dot{v} &= \frac{T - D}{m} - g \sin\gamma \\ \dot{y} &= v \sin\chi \cos\gamma, & \dot{\chi} &= \frac{gn \sin\phi}{v \cos\gamma} \\ \dot{h} &= v \sin\gamma, & \dot{\gamma} &= \frac{g}{v}(ncos\phi - \cos\gamma) \end{aligned} \tag{5}$$

The drag $D$ is modelled with the aid of the BADA drag polars [20], aircraft mass is $m$, $g$ is gravitational acceleration and $n$ is the load factor. Conversions between the virtual and time domain are achieved by:

$$\begin{gathered} \dot{r} = \lambda r'; \quad \ddot{r} = \lambda(r''\lambda + r'\lambda') \\ \dddot{r} = \lambda^3 r''' + 3\lambda^2 \lambda' r'' + (\lambda^2 + \lambda\lambda'^2) r' \end{gathered} \tag{6}$$

Once the state and control histories are determined from $\tau_o$ to $\tau_f$, the following constraints are applied to ensure that they lie between defined limits determined with the aid of the BADA database [20]:

$$\begin{gathered} T \in [T_{min}, T_{max}] \quad n \in [n_{min}, n_{max}] \quad |\phi| \le |\phi_{max}| \\ v \in [1.2 v_{stall}, v_{max}] \quad \dot{v} \in [\dot{v}_{min}, \dot{v}_{max}] \quad \gamma \in [\gamma_{min}, \gamma_{max}] \end{gathered}$$

For the BADA parameters, Eurocontrol compensate for inaccuracies in the parameter estimation process by defining limits for both aircraft thrust and acceleration.

## IV. Waypoint Constraints

For commercial aircraft trajectory optimization, constraints imposed by the operating environment must be considered. Calculated trajectories must be able to adhere to ATC constraints imposed by airspace sectorization, procedures and traffic flow corridors. Typically, operating environment and procedural restrictions manifest as constraints on the height, speed or path of the flight, or some combination of the three. Therefore a simple five dimensional constraints model is developed.

Waypoint fixes are defined by the user in 2 dimensions as $\mathbf{W}_r = [W_x, W_y]$. The aircraft's trajectory path can then be constrained to fly over the 2D fix. For each waypoint fix, the minimum distance between the trajectory path and the fix position is calculated as a 2D distance $d_{min}$, with a corresponding minimum time $t_{d_{min}}$,

$$d_{min} := \min_{t \in [t_0, t_f]} d(t) \;\text{ where }\; d(t) = \|r_{2D}(t) - \mathbf{W}_r\|$$

$$t_{d_{min}} := \min t \;\; s.t. \;\; d(t) = d_{min}$$

The aircraft is then constrained to fly within a distance radius of the centre point of the fix, where $\hat{d}$ is the upper constraint on $d_{min}$,

$$\mathbf{c}_1(d_{min}) = d_{min} - \hat{d}$$

The aircraft can also be constrained to cross the fix at a specified height, speed and arrival time. The cross above constraints, $\underline{h}, \underline{v}, \underline{t}$, and cross below constraints, $\overline{h}, \overline{v}, \overline{t}$, constrain the minimum and maximum heights, speeds and time of the aircraft crossing the waypoint. The minimum and maximum can be constrained simultaneously to create height,speed and time windows at the waypoint.

$$\mathbf{c}_2(\mathbf{h}, \mathbf{v}, \mathbf{t}, t_{d_{min}}) = \begin{bmatrix} h(t_{d_{min}}) - \overline{h}_{t_{min}} \\ \underline{h}_{t_{min}} - h(t_{d_{min}}) \\ v(t_{d_{min}}) - \overline{v}_{t_{min}} \\ \underline{v}_{t_{min}} - v(t_{d_{min}}) \\ t_{d_{min}} - \overline{t}_{min} \\ \underline{t}_{min} - t_{d_{min}} \end{bmatrix}$$

## V. Environmental models

For the calculation of civil aircraft noise impact local to airports, the most commonly used method in the field is the Noise Power Distance (NPD) method [21]. The NPD method utilises, for a number of noise metrics, tables of empirical data that relate the noise level calculated on the ground to the power utilised by the aircraft and the distance from the aircraft to the noise assessment point. Specifically, the noise level at a point is calculated as

$$noise\,level = f_n(P(t), d(t), \boldsymbol{\beta}), \quad \forall t \in [t_o, t_f] \tag{7}$$

where $P(t)$ is power, $d(t)$ is slant distance between the aircraft and the assessment point, $\boldsymbol{\beta}$ is a set of segment level correction terms, and $t_o$ and $t_f$ are respectively the start and end times of the trajectory. Power, in this instance, is corrected net thrust. The noise model chosen for use in this work is the Integrated Noise Model (INM). INM is a model developed by the Federal Aviation Administration (FAA) to assess the impact of civil aircraft noise on communities local to airports. INM version 7 [22] is fully compatible with ECAC Doc 29, [21] guidance that provides a standardized methodology for the computation of noise contours around civil airports. INM is able to calculate several types of aircraft noise impact metrics, including the maximum A-weighted sound pressure level $L_{Amax}$, the single event exposure metrics Sound Exposure Level $(SEL)$ and The Effective Perceived Noise Level $(EPNL)$. INM also allows the calculation of noise metric contours that can be used to calculate population exposures and awakenings.

For the calculation of aircraft emissions, the emissions CO2, water vapor (H2O) and sulphur oxides (SOx) are calculated using direct multipliers on fuel burn, such that their rate of emissions is:

$$\dot{e}(t) = f_e(F(t), \boldsymbol{\alpha}) \tag{8}$$

where $F(t)$ is the rate of fuel burn and $\boldsymbol{\alpha}$ is a set of fuel burn multipliers. For the work presented in this document, rate of fuel burn is calculated using the BADA fuel flow model [20], where $F(t)$ is a function of thrust and the thrust specific fuel consumption factor $\eta$ such that:

$$F(t) = f_F(P(t), \eta) \tag{9}$$

The emissions for hydrocarbons (HC), carbon (CO), and oxides of nitrogen (NOx) are calculated using the Boeing Fuel Flow Methodology (BFFM) [23]. The ICAO emissions databank contains empirical information for certified engines that relate fuel burn to emissions indices at 4 different engine thrust settings. The BFFM offers a procedure for correcting the data for atmospheric conditions and for interpolating the data to calculate emissions of HC, CO, and NOx from fuel flow such that their rate of emissions is:

$$\dot{e}(t)_{BFFM} = f_{BFFM}(F(t), \mathbf{EI}) \tag{10}$$

where EI is a set of of emissions indices.

## VI. Stochastic solver

NLP algorithms can be applied in the iterative solution of a wide array of optimization problems. NLP algorithms can be generally classified as direct search (no derivatives), gradient search or stochastic. Derivative based gradient solvers require the definition or estimation of the derivatives of the objective function, while derivative free optimisers compare only the objective function value. To reach a global optimum it helps if gradient and derivative free optimisers are initialised where the objectives and constraints are convex. Stochastic solvers however tend to find it easier to escape local minima. Like derivative free algorithms, stochastic algorithms compare only objective values between iterations but use probability functions in guiding the search for the optimum solution. More details on stochastic optimization methods can be found in [24]. Evolutionary algorithms are a popular type of stochastic method and are well suited to solving multi objective optimization problems. Evolutionary algorithms, at each step of an optimization, maintain a population of solutions. This, allows the algorithms to simultaneously explore different parts of the solution space to identify decision vectors that provide the minimum for each objective and also the Pareto optimal points that form a front between the minimums.

The stochastic solver chosen for this work was Differential Evolution [25]. The DE algorithm utilises the mechanism of differential mutation. Differential mutation is a self adaptive mechanism where 3 population vectors are randomly selected from the parent generation and the scaled difference between 2 of the vectors is added to the third. The DE algorithm has proved to be a simple yet effective method for handling global optimization problems [26]. DE requires only 3 configuration parameters for calibration and has shown promising results when used with the inverse dynamics method chosen for use in this research. Drury [27] tested the performance of the Inverse Dynamics method with a number of popular Non Linear Programming (NLP) algorithms. For 2000 different sets of boundary values, the DE algorithm achieved a convergence of 99.8% with a relative optimality score of 94%, outperforming all of the other NLP algorithms. A multiobjective variant of DE is defined in [25] where Lampinen's dominance based method for constrained optimization [28] is used to drive solutions towards a Pareto front. Madavan [29] subsequently showed that multiobjective DE could be supplemented with a nondominated sorting procedure and the crowding distance metric developed by Deb et al [30] for the Nondominated Sorting Genetic Algorithm (NSGAII). The result was a fast and powerful method that combined the self organising mechanism of differential mutation with the elitism and the diversity preservation of the NSGAII algorithm.

### A. Main

The variant of DE chosen for this work was DE/rand/1/Bin [25]. The algorithm initialises by generating a population of random individuals between the user specified upper $b_U$ and lower $b_L$ parameter bounds. For each individual in this population, a candidate vector is created that is the algebraic combination of 3 parent vectors, further modified by crossover between the candidate and the target vectors. Each candidate vector is placed in an offspring population. Once a candidate population has been generated that is the same length as the parent population, it is appended to the parent population. The combined population is then sorted into fronts using nondominated sorting. For the production of the new population for the next generation, the size of the combined population is reduced to the initial population size by truncation. Each solution in the solution set is sorted into fronts $(F)$ depending on the number of other solutions in the set each is dominated by. All fronts from the first nondominated front $(F_1)$ upwards , whose combined length is less than or equal to the initial population size are preserved for the next generation. The individuals in the final front are sorted according to their crowding distance value. Individuals with large crowding distance values are moved toward the start of the front and the individuals with

Algorithm 1 A fast, self-adaptive, elitist, diversity preserving, multi-objective evolutionary algorithm

```
$G = 1$                                                First generation
$P = \emptyset$                                        Initialise initial population
for $i := 1$ to $NP$                                   Create $NP$ real valued vectors
    $\mathbf{p}_i = \emptyset$                         Initialise population vector
    for $j := 1$ to D                                  Each population vector contains $D$ real parameters
        $x_{j,i,G} = rand(0,1)(b_{j,U} - b_{k,L}) + b_{j,L}$   Create parameters within bounds
        $\mathbf{p}_i = \cup\{x_{j,i,G}\}$             Add the parameter to the population vector
    end
    $P = \cup\{\mathbf{p}_i\}$                         Add the population to the set of populations
end


while $G < G_{max}$                                    While current generation is less than final generation
        $Q = \emptyset$                                Initialise offspring population
        for each $\mathbf{p}_{i,G} \in P_G$            For each population vector in the population
        $r_1, r_2, r_3 \in \{1, ..., NP\}$             Select 3 random indexes
        $F_S \in [0,1]$                                Scale factor
        $\mathbf{v}_{i,G} = \mathbf{p}_{r_1,G} + F_S(\mathbf{p}_{r_2,G} - \mathbf{p}_{r_3,G})$   Create mutant population with Differential mutation
        $CR \in [0,1]$                                 DE crossover parameter $CR$
        $k \in \{1, ..., D\}$                          Random parameter index
        $\mathbf{q}_i = \emptyset$                     Initialise candidate vector
        for each $x_{j,G} \in \mathbf{p}_{i,G}$        For each parameter in the vector
            $r = rand(0,1)$
            if $r <= CR \vee j == k$                   Crossover between parent and the mutant vectors
                $u_{j,G} = v_{j,G}$
            else
                $u_{j,G} = x_{j,G}$
            end
            $\mathbf{q}_{i,G} \cup \{u_{j,G}\}$        Add parameter to candidate vector
        end
        $Q_G \cup \{\mathbf{q}_{j,G}\}$                Add candidate vector to offspring population
        end
        $R_G = P_G \cup \{Q_G\}$                       Combine parent and offspring population
        $F = fast\ nondominated\ sort(R_G)$
        $P_{G+1} = \emptyset\ and\ l := 1$             All nondominated fronts of $R_G$
        while $|P_{G+1}| + |F_l| \leq |P_G|$           Until the new population $P_{G+1}$ is filled
                $crowding\ distance\ assignment(F_l)$  Calculate the crowding distance in $F_l$
                $P_{G+1} = P_G \cup \{F_l\}$           Include the $l^{th}$ nondominated front in the new population
                $l = l + 1$                            Check the next front for inclusion
        end
        $sort(F_l : F_{end}, \prec)$                   Sort final front in descending order using domination
        $P_{G+1} = P_{G+1} \cup F_L[1 : (N - |P_{G+1}|)]$   Truncate final front if required
        $G = G + 1$                                    Increment generation counter
end
```

small crowding distance values are moved towards the rear. If inclusion of the final front results on a population size greater than the initial population size, then the final front is truncated. This ensures that the solutions in the final front with the greatest solution diversity are retained for the next generation.

### B. Nondominated Sorting

In general, for 2 feasible solutions where $\mathbf{p}, \mathbf{q} \in S$, $\mathbf{p}$ dominates $\mathbf{q}$ $(\mathbf{p} \prec \mathbf{q})$ if $\forall k : f_k(\mathbf{p}) \leq f_k(\mathbf{q}) \wedge \exists k : f_k(\mathbf{p}) < f_k(\mathbf{q})$ Nondominated sorting involves using domination to rank each solution into fronts that are sets of solutions with equal dominance ranking. Solutions that are not dominated by any other solutions are assigned to the first front $F_1$, Solutions that are dominated by 1 other individual will appear on the next front $F_2$, and so on until all solutions are assigned to the appropriate front.

### C. Crowding Distance

The crowding distance metric is used to measure the distance along the same nondominated front from one solution to the 2 adjacent solutions. For each objective function, the greatest and smallest objective values are assigned an infinite crowding value, preserving the boundary value individuals. For each intermediate individual, its proximity to other individuals is determined by taking the normalised difference between the solutions either side of that solution. When this measure is summed over all individual's objective functions, a measure of the closeness between solutions is reached.

### D. Selection

Unlike many Evolutionary Algorithms, where parent and child populations are compared to each other for fitness, for the NSGA-II method, the parent and child populations are appended to each other, and all selection occurs within the same population. Fitter individuals as determined by their nondomination rank and crowding distance are moved to the front of the population, while less fit individuals are moved to the back of the population and may be subjected to truncation. As shown in Algorithm 2, domination is determined by the constraints, rank, and crowding distance of each individual. In Algorithm 2, the solution $\mathbf{i}$ dominates the solution $\mathbf{j}$ if both are feasible and $\mathbf{i}$ has either a lower nondominated rank ($rank$) or has a greater crowding distance value ($distance$) at the same nondominated rank. If $\mathbf{i}$ is feasible and $\mathbf{j}$ is infeasible then $\mathbf{i}$ dominates. If both solutions are infeasible, then the individual with the lowest overall constraint violation dominates.

Algorithm 2 Selection

$$\zeta(\mathbf{i}) = \sum_{k}^{m} max[0, g_k(\mathbf{i})]$$

$$\zeta(\mathbf{j}) = \sum_{k}^{m} max[0, g_k(\mathbf{j})]$$

$$\mathbf{i} \prec \mathbf{j} \;\; \textit{if} \begin{cases} \begin{cases} \zeta(\mathbf{i}) \leq 0 \wedge \zeta(\mathbf{j}) \leq 0 \\ \wedge \\ i_{rank} < j_{rank} \\ \vee \\ [i_{rank} = j_{rank}] \wedge [i_{distance} > j_{distance}] \end{cases} \\ \vee \\ \begin{cases} \zeta(\mathbf{i}) \leq 0 \\ \wedge \\ \zeta(\mathbf{j}) > 0 \end{cases} \\ \vee \\ \begin{cases} \zeta(\mathbf{i}) > 0 \\ \wedge \\ \zeta(\mathbf{i}) \;\; < \zeta(\mathbf{j}) \end{cases} \end{cases}$$

## VII. Results

A departing aircraft scenario was created to demonstrate the multi objective trajectory optimization method. In the scenario, a commercial aircraft is required to climb from an initial climb point below 500ft at the west of a large population centre, to an en-route connection point lying at 20,000ft on the far side of the population centre. The commercial aircraft simulated was the medium narrow-body Airbus A321 aircraft with twin International Aero Engine V2530 engines. The population was artificially created for the scenario, and consisted of 1.5 million people evenly distributed over an area of 45000 hectares. In addition to the constraints from Section III, the bank angle $\phi$ and the minimum climb gradient below 1000ft were constrained to $0(rad)$ and $12\%$ respectively. For the first scenario the objectives chosen were the greenhouse gas Carbon Dioxide and the population enclosed within the 70 dB(A) Sound Equivalent Level footprint contour. SEL was used here as it forms the basic 'building block' of the Lden (Day-Evening-Night Average Sound Level) and Ldn (Day-Night Average Sound Level) contour calculations used to asses the community impact of aircraft noise.

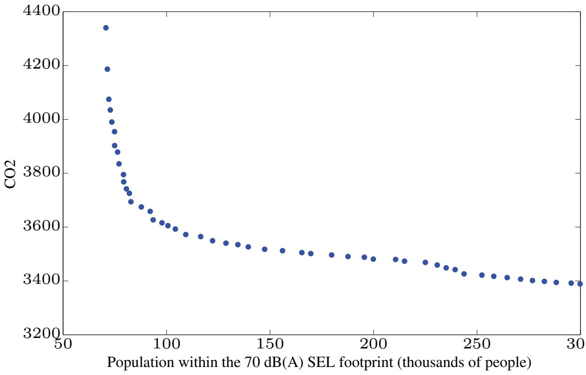


Fig. 1. Pareto front between CO2 emissions and the population within the 70dB(A) footprint

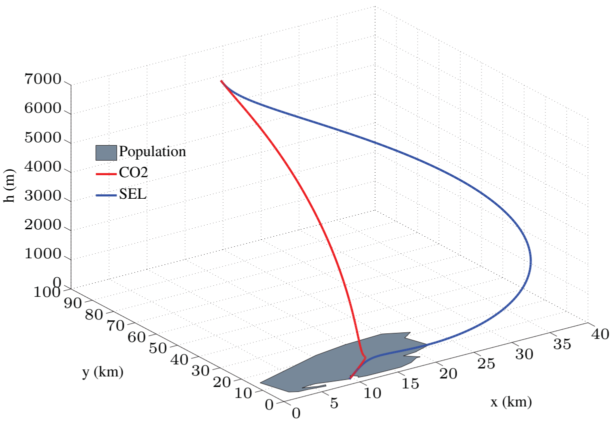


Fig. 2. Minimum noise and CO2 trajectories

Fig. 1 shows a Pareto front between the minimums of the two objectives. It can be seen from the front that there is a trade-off of approximately 900kg of CO2 between the most CO2 optimal trajectory and the most noise optimal trajectory. Similarly there is a trade-off of approximately 300,000 exposed people between the most noise optimal trajectory and the most CO2 optimal trajectory. Fig. 2 shows the trajectories for the two minimums. The aircraft trajectory for the minimum CO2 objective, after climbing out to 1000ft takes a direct route over the population area to the target end point, minimising excess track miles, fuel burn and therefore CO2. The trajectory for the minimum noise objective, initially progresses directly to the east, avoiding over-flying the majority of the population area and therefore minimising the population exposed to noise. It can be seen from Fig. 2 and from Fig. 3 that the noise optimised trajectory climbs to a height of 1500ft where it reduces thrust and accelerates to zero flap speed while passing over a population region near to the airport. On clearing the population region, the aircraft is turned to the target end point while thrust is increased gradually to maximum climb thrust and the aircraft accelerates to en-route climb speed. The trajectory produced mimics closely the Sourdine close-in noise abatement procedure as both involve an initial climb at full thrust followed by acceleration at reduced thrust and a gradual power increase.

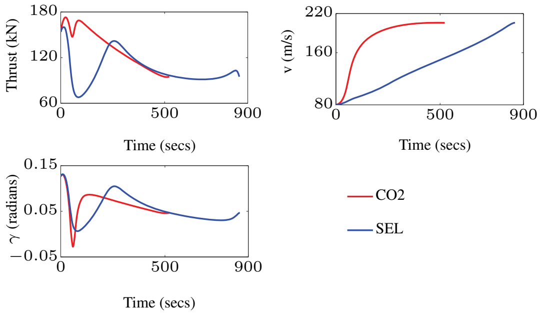


Fig. 3. Thrust, speed and flight path angle histories for the minimum noise and CO2 trajectories

A second scenario was created to examine the trade-offs between minimising for noise near to the airport and minimising for noise farther away from the airport. Boundary values remained as in the first scenario, however, all solution trajectories were constrained via the constraints in Section IV to lie along a common $x, y$ ground path shown in Fig. 4.

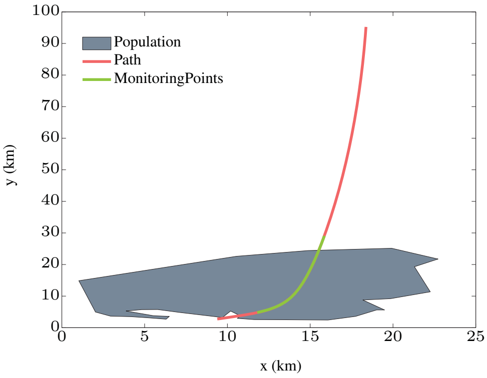


Fig. 4. Constrained $x, y$ path with noise monitoring region

All improvements in objectives were therefore achieved through changes to the vertical trajectory. A region of noise monitoring points, shown in green in Fig. 4, were then placed at 1000m intervals from a distance of 3000m to 30,000 metres from start of roll. The region of 3000m to 15000m was defined as the close-in region and the region of 15,000 to 30,000 metres was defined as the far region. Two objectives were then the subject of the optimization, Average EPNL in the close-in region, and average EPNL in the far region. Fig. 5 shows the Pareto front between the average EPNL in the near region and average EPNL in the far region. It can be seen that there is a 2-3 EPNdB average EPNL tradeoff between optimising trajectories for the different regions.

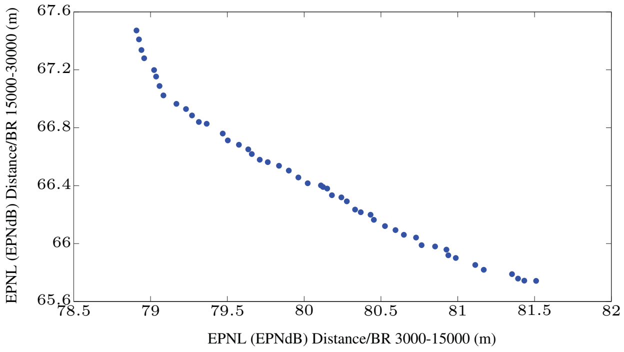


Fig. 5. Pareto front between close-in and far region average EPNL

It can be seen from Fig. 6 and from Fig. 7 that the near region optimization causes the aircraft to cutback its thrust and to assume a shallower flight path angle climb sooner than the far region optimization. As can be seen in Fig. 8, this results in a reduction in EPNL noise levels along the trajectory centreline at distances running from 3000 to 11000 metres. By contrast the far region optimization results in the aircraft climbing as high as possible using maximum thrust and flight path angle prior to the beginning of the far region. Once the far region is reached, the aircraft reduces thrust and climb angle, but the greater altitude attained allows the aircraft to fly higher over the far region increasing the noise attenuation distance and reducing noise levels on the ground. Once the far region has been passed, thrust is increase to accelerate the aircraft to enroute speed. Fig. 8, shows the corresponding higher close-region noise levels and lower far region noise levels that result from this trajectory.

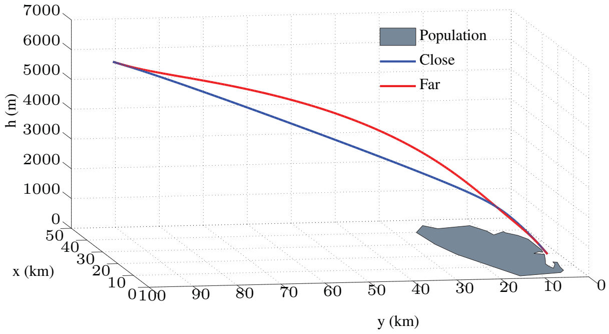


Fig. 6. Minimum close-in noise and far region noise trajectories

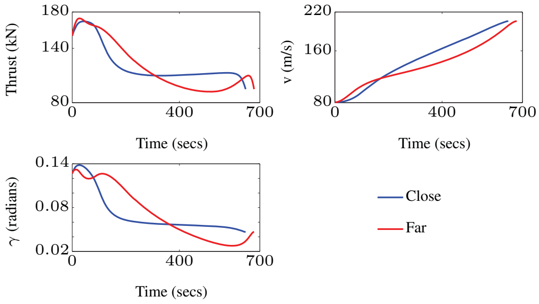


Fig. 7. Thrust, speed and flight path angle histories for the minimum close-in and far noise trajectories

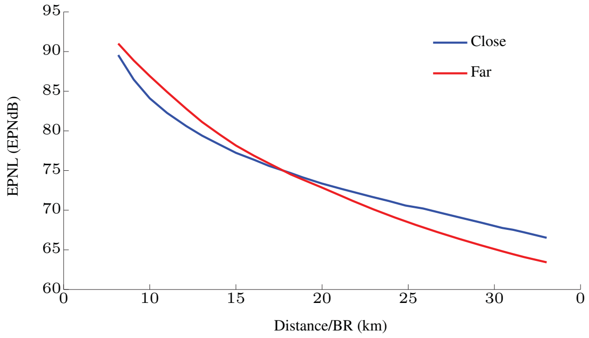


Fig. 8. Minimum close-in noise and far region noise trajectories

## VIII. Conclusion

As can be seen from Section VII, under many circumstances, there is no one trajectory that minimises all environmental costs. Therefore when planning environmentally efficient trajectories, the trade-offs between the objectives must be considered. This work has proposed a multi-objective trajectory optimisation method that may be used to analyse the trade-offs between environmental objectives that arise when operating commercial aircraft in different ways. The work assumes the availability of an advanced Flight Management System (FMS) with auto-throttle capable of tracking the detailed trajectory solutions. However, each trajectory may be segmented into a smaller number of operational steps that would be suitable for a pilot to execute. Further work will involve the automated analysis of the Pareto fronts so that certain solutions from the front can be recommended to the user to aid their decision making. This is intended to be especially useful when analysing Pareto fronts between more than two objectives.

## Acknowledgment

This work was completed thanks to funding support from Cranfield university. The authors would also like to thank Ken Lai from Cranfield University, Peter Hullah from Eurocontrol and Rick Drury from Flight Applications Limited for their help and guidance.

## References

[1] Airbus, “Delivering the future: Global market forecast 2011-2030,” 2011.

[2] ICAO, “Environmental report 2010,” 2010.

[3] ACARE, “Strategic research agenda. volume 2,” October 2002.

[4] European Commission, “European air traffic management master plan,” March 2009.

[5] Committee on Aviation Environmental Protection, “Review of continuous descent approach (CDA) implementation and associated benefits,” *Working Paper CAEP/7-WP/26*, 2006.

[6] Sourdine Consortium, “Study of optimisation procedures for decreasing the impact of noise (sourdine),” *Report D5*, 2001.

[7] NLR, “SourdineII noise results amsterdam schiphol,” *D4-1-2b*, 2006.

[8] INECO, “SourdineII noise results madrid barajas,” *D4-1-3b*, 2006.

[9] R. A. Visser, Hendrikus G. Wiljnen, “Optimization of noise abatement departure trajectories,” *Journal of Aircraft*, vol. 38(4), p. 620627, 2001.

[10] ——, “Optimization of noise abatement arrival trajectories,” *The Aeronautical Journal*, vol. 107(1076), p. 607615, 2003.

[11] C. Hargraves and S. Paris, “Direct trajectory optimization using nonlinear programming and collocation,” in *Astrodynamics 1985*, vol. 1, 1986, pp. 3–12.

[12] S. Hebly and H. Visser, “Advanced noise abatement departure procedures: custom optimized departure profiles,” in *AIAA Guidance, Navigation and Control Conference and Exhibit, Honolulu, HI*, 2008.

[13] X. Prats, V. Puig, J. Quevedo, and F. Nejjari, “Lexicographic optimisation for optimal departure aircraft trajectories,” *Aerospace Science and Technology*, vol. 14, no. 1, pp. 26–37, 2010.

[14] J. Betts, “Survey of numerical methods for trajectory optimization,” *Journal of guidance, control, and dynamics*, vol. 21, no. 2, 1998.

[15] F. Fahroo and I. Ross, “Advances in pseudospectral methods for optimal control,” in *AIAA Guidance, Navigation and Control Conference and Exhibit*, 2008.

[16] ——, “Direct trajectory optimization by a Chebyshev pseudospectral method,” in *Proceedings of the 2000 American Control Conference, 2000.*, vol. 6. IEEE, 2000, pp. 3860–3864.

[17] D. Benson, G. Huntington, T. Thorvaldsen, and A. Rao, “Direct trajectory optimization and costate estimation via an orthogonal collocation method,” *Journal of Guidance Control and Dynamics*, vol. 29, no. 6, pp. 1435–1440, 2006.

[18] O. Yakimenko, “Direct method for rapid prototyping of near-optimal aircraft trajectories,” *AIAA Journal of Guidance, Control, and Dynamics*, vol. 23, no. 5, 2000.

[19] C. Lai and J. Whidborne, “Real-time trajectory generation for collision avoidance with obstacle uncertainty,” in *AIAA Guidance, Navigation and Control Conference, AIAA 2011-6598*, Portland OR, August 2011.

[20] A. Nuic, “User manual for the base of aircraft data,” *(BADA) revision 3.7*, vol. 2010, p. 001, 2010.

[21] ECAC, “Standard method of computing noise contours around civil airports,” *ECAC. CEAC Doc. 29*, 2010.

[22] J. Olmstead, G. Fleming, J. Gulding, C. Roof, P. Gerbi, and A. Rapoza, “Integrated noise model (inm) version 7.0 technical manual,” *Report FAA-AEE-02-01, Office of Environment and Energy, Federal Aviation Administration*, 2002.

[23] S. Baughcum, T. G. Tritz, S. C. Henderson, and D. C. Pickett, *Scheduled civil aircraft emission inventories for 1992: Database development and analysis*. National Aeronautics and Space Administration, Langley Research Center, 1996.

[24] D. Goldberg, *Genetic algorithms in search, optimization, and machine learning*. Addison-wesley, 1989.

[25] K. Price, R. Storn, and J. Lampinen, *Differential evolution: a practical approach to global optimization*. Springer-Verlag New York Inc, 2005.

[26] R. Storn and K. Price, “Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces,” *Journal of Global Optimization*, vol. 11, no. 4, pp. 341–359, 1997.

[27] R. Drury, “Performance of NLP algorithms with inverse dynamics for near-real time trajectory generation,” in *AIAA Guidance, Navigation, and Control Conference*, August 2011.

[28] J. Lampinen, “A constraint handling approach for the differential evolution algorithm,” in *Proceedings of the 2002 Congress on Evolutionary Computation, 2002. CEC’02.*, vol. 2. IEEE, 2002, pp. 1468–1473.

[29] N. Madavan, “Multiobjective optimization using a pareto differential evolution approach,” in *Proceedings of the 2002 Congress on Evolutionary Computation, 2002. CEC’02.*, vol. 2. IEEE, 2002, pp. 1145–1150.

[30] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, “A fast and elitist multiobjective genetic algorithm: Nsga-II,” *Evolutionary Computation, IEEE Transactions on*, vol. 6, no. 2, pp. 182–197, 2002.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# A Formulation for Globally Optimal Controlled Variable Selection

Lingjian Ye
Ningbo Institute of Technology
Zhejiang University
315100, Ningbo, Zhejiang, China
Email: lingjian.ye@gmail.com

Yi Cao
School of Engineering
Cranfield University
Cranfield, Bedford MK43 0AL, UK
Email: Y.Cao@cranfield.ac.uk

***Abstract*—Self-optimizing control (SOC) is a powerful tool to select controlled variables (CVs) so that when these variables are maintained at constant set-points, the entire process operation is automatically optimal or near optimal (self-optimizing) in spite of the presence of various uncertainties. Over a decade development, many SOC theories and methods have been developed to select optimal CVs. However, all these methods are based on local linearization of the process model at a nominally optimal operating point, hence referred to as local methods.**

**Due to the nature of locality, existing SOC methods may cause a large performance loss when the feasible operation region is large and the process is highly nonlinear. In this paper, we propose a global approach to select optimal CVs for nonlinear processes so that the average loss over the entire feasible operation region is minimized. Firstly, the globally average loss minimization problem is formulated and a toy example is solved analytically to explain the difference between the global approach and other local methods. For more complex processes where an analytical solution is not tractable, a numerical approach is proposed to minimize the average loss globally. In the new approach, optimal CV selection is found by solving a regression problem to approximate the necessary conditions of optimality of the objective function. A case study on an exothermic reactor demonstrates the effectiveness of the new approach.**

## I. Introduction

Traditionally, CVs are selected from a list of available or inferred measurements based on heuristic experiences and understanding of the whole process from an engineering perspective. For example, the variables related to safety regulations and product qualities usually need to be actively controlled and they are naturally chosen as CVs, thus consuming most of the process degrees of freedom (DOF). In many cases, these active constraints will dominate the process operation. However, for processes with extra DOF more than active constraints, self-optimizing control (SOC) [1] showed that the remain DOF can be used for optimization purpose by selecting appropriate CVs. When the selected CVs are maintained at predetermined constant set-points, the entire process operation is automatically optimal or near optimal (self-optimizing) in spite of the presence of various uncertainties, disturbances and measurement errors.

Over a decade development, many SOC theories and methods have been developed to select optimal CVs. Halvorsen et al. [2] derived simple singular value rule and local exact method for CV selection. Recent works have been engaged in finding proper combination matrix $H$ of measurements as CVs to reduce the economic cost. Kariwala [3] minimized the local worst-case loss via singular value and eigenvalue decompositions. Later Kariwala et al. [4] derived optimal $H$ with average loss minimization. Alstad and Skogestad [5] presented a null space method to minimize the loss caused by disturbances. Furthermore, Alstad et al. [6] extended null space method using extra measurements to minimize the loss caused by implementation error. Hori and Skogestad [7] compared maximum gain rule and local exact method and found the former one should be used with care for ill-conditioned plants. All these SOC methods were derived based on linearization of the process model around a normally optimal operating point, hence are referred to as local methods. This means the CVs selected by using these methods may only be optimal in a small neighborhood around the nominal point.

To address the locality issue, this paper aims to select CVs for nonlinear processes to be globally optimal by minimizing the average loss across the entire operation region. The remaining of this paper is organized as follows: Section II briefly reviews the local SOC methods, and Section III presents the formulation of globally average loss minimization problem for CV selection, together with a toy example to explain the difference between the global approach and existing local approaches. In Section IV, a CV selection procedure through regression for more general processes is proposed. The effectiveness of proposed solution is further demonstrated through an exothermic reactor case study in Section V. Finally, Section VI concludes the work together with some suggestions for future works.

## II. A Brief Review of Local Methods based Self-Optimizing Control

Consider a generalized static optimization problem for continuous processes, which is given as

$$\begin{aligned} &\min_u J(u,d) \qquad (1)\\ &\text{s.t. } g(u,d) \le 0 \end{aligned}$$

with available measurements

$$y = f(u,d) \qquad (2)$$

where $J$ is the scalar objective function; $u \in \mathbb{R}^{n_u}$, $d \in \mathbb{R}^{n_d}$ and $y \in \mathbb{R}^{n_y}$ are manipulated, disturbance and measurement variables, respectively; $g : \mathbb{R}^{n_u \times n_d} \Rightarrow \mathbb{R}^{n_g}$ and $f : \mathbb{R}^{n_u \times n_d} \Rightarrow \mathbb{R}^{n_y}$ are the operational constraints and measurement equations, respectively.

Let $c$ represent the CVs with set-points at $c_s$. SOC [1] demonstrated that if $c$ are properly selected, then when these variables are perfectly maintained at their optimal values, $c_s = c_{opt}(d)$ in the presence of disturbance, $d$, manipulated variables, $u$ will approach to their optimal values $u_{opt}(d)$ through feedback control without re-optimizing $c_s$. To select CV properly, let measurements $y$ at the nominal point be linearized as

$$y = G_y u + G_{yd} W_d d + W_n n \tag{3}$$

where $G_y$ and $G_{yd}$ are the steady state gain matrices of $y$ with respect to $u$ and $d$, respectively; $n$ is the implementation error due to measurement noise and/or control errors associated with individual measurement; $W_d$ and $W_n$ are diagonal matrices representing magnitudes of $d$ and $n$ respectively.

The selected CVs as linear combinations of full measurements set $y$ can be represented as

$$c = Hy \tag{4}$$

where $H$ is the combination matrix with full row rank of $n_u$ to square the control system. Especially, zero columns in $H$ imply a subset of full y is unused. The worst case loss [2] and the average loss [4] in objective function due to maintaining CVs at constant set-points for uniformly distributed $d$ and $n$ are given in (5) and (6), respectively.

$$L_{\text{wc}} = \frac{1}{2}\sigma^2_{\max}(M) \tag{5}$$

$$L_{\text{av}} = \frac{1}{6(n_d + n_y)}\|M\|^2_F \tag{6}$$

where $\sigma_{\max}(\cdot)$ and $\|\cdot\|_F$ are the maximum singular value and Frobenius norm of a matrix respectively; $M = \left[J_{uu}^{1/2}\left(J_{uu}^{-1}J_{ud} - G^{-1}G_d\right)W_d \quad J_{uu}^{1/2}G^{-1}HW_n\right]$ with $G = HG_y$ and $G_d = HG_{yd}$. Here, $J_{uu} = \partial^2 J/\partial u^2$ and $J_{ud} = \partial^2 J/(\partial u \partial d)$ are the diagonal and off-diagonal Hessian matrices of $J$ evaluated at the nominal point.

CV selection is then characterized as minimizing (5) or (6) with respect to $H$. Recently, several explicit expressions for $H$ have been reported [5], [3], [4], [6]. For example, if $n_y = n_u + n_d$, the combination matrix $H$ according to null space method proposed by Alstad and coworkers [5], [6] can be selected as

$$H = \left[J_{uu}^{1/2} \quad J_{uu}^{-1/2}J_{ud}\right]\left[G_y \quad G_{yd}\right]^{-1} \tag{7}$$

## III. Selection of Globally Optimal Controlled Variables

### A. Formulation

In the existing SOC theory, the CV selection problem is solved by assuming measurements in (2) are linearized in (3). This assumption restricts the solution to be local. To avoid this locality, the linear model assumption is discarded in the new global formulation to be presented below. Furthermore, in general, CVs can be either linear (as shown in (4)) or nonlinear combinations of all or a subset of available measurements. Therefore, in this work, the CVs are parameterized by $w \in \mathbb{R}^{n_w}$ as follows.

$$c = \phi(y, w) \tag{8}$$

For simplicity but without loss of generality, it is assumed that $c_s = 0$.

Let the entire operating range defined by the disturbance $d \in \mathcal{D}$ and all possible measurement noise represented by $n \in \mathcal{N}$. It is assumed that $d$ and $n$ are statistically independent. The minimum cost of the optimization problem in (1) is a function of $d$, $J_{\text{opt}}(d)$, while the actual cost when $c = 0$ is a function of $d$ and $n$ denoted as $J_w(d, n)$. Then, the operation loss in terms of the cost function for $c = 0$ with specific $d$ and $n$ is $L_w(d, n) = J_w(d, n) - J_{\text{opt}}(d)$, whilst the average loss across the entire operation range is a function of design parameters, $w$ as follows.

$$\begin{aligned} L_{\text{gav}}(w) &= E[L_w(d,n)] && (9)\\ &= \int_{n\in\mathcal{N}, d\in\mathcal{D}} \rho(d)\rho(n)L_w(d,n)\mathrm{d}n\mathrm{d}d && (10)\end{aligned}$$

where $E[\cdot]$ and $\rho(\cdot)$ represent the expected value and the probability density of a random variable, respectively.

Then, the globally optimal CVs can be selected by designing $w$ to solve the following optimization problem

$$\begin{aligned} &\min_w L_{\text{gav}}(w) && (11)\\ &\text{s.t. } y = f(u, d) + n\\ &\quad\;\; 0 = \phi(w, y)\\ &\quad\;\; d \in \mathcal{D}\\ &\quad\;\; n \in \mathcal{N}\end{aligned}$$

The optimization problem in (11) is generally suitable for various $\mathcal{D}$, $\mathcal{N}$ and nonlinear combination, $\phi$.

It was shown in [8] that local SOC methods essentially capture necessary conditions of optimality (NCO) locally at the nominal point with a straight line as CV. The aim of this work is to search other curves as CVs, which have better average performance across the whole region. The differences of these methods are illustrated in Figure 1, where $c_{\text{loc}}$ is the CV obtained by local SOC method, $c_{\text{gav}}$ is the globally optimal CV. As shown in Figure 1(a), the vertical axis is the NCO value when $c = 0$, where the desired value is 0. The curve of NCO for $c_{\text{loc}}$ can only be maintained at the nominal point, $d = d^\star$, whereas the deviation goes large as the operation point drifts far away from the nominal point. The curve of NCO for $c_{\text{gav}}$ approaches the zero line more closely in an average sense to minimize the overall loss, which can be quantified by the area surrounded by the curve of $L$ and $d$ axis in the disturbance range $[\underline{d}, \bar{d}]$, as shown in Figure 1(b).

From Figure 1, it is clear that the price to gain better average performance across the entire region is the slightly increased loss around the nominal operating point. To help understanding of the general formulation, a toy example is to be solved as follows.

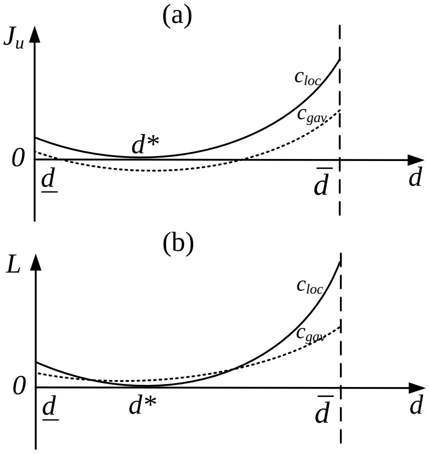


Fig. 1. Comparison of local SOC and globally optimal CV: (a) NCO against $d$; (b) Loss function $L$ against $d$

### *B. A toy example*

To illustrate the proposed new framework for globally optimal CVs selection, consider a problem to minimize the objective function

$$J = \frac{1}{2}(u-d)^2 \tag{12}$$

where both $u$ and $d$ are scalars. Two measurements are available as follows.

$$y_1 = u \tag{13}$$

$$y_2 = \frac{1}{4}u^2 + d \tag{14}$$

The nominal disturbance is $d^* = 0$. Correspondingly, the optimal point is then defined by $u^* = 0$, where $J^* = 0$, $y_1^* = 0$ and $y_2^* = 0$. The possible variation of $d$ is uniformly distributed between $-1$ and 1, *i.e.* $d \in \begin{bmatrix}-1 & 1\end{bmatrix}$.

It is clear that $u_{\text{opt}}(d) = d$ and correspondingly, $J_{\text{opt}}(d) = 0$. Therefore, for a non-optimal input, $u$, the loss, $L(u,d) = J(u,d) - J_{\text{opt}}(d) = J(u,d)$.

Firstly, the local SOC approach of null space method proposed in [5], [6] is applied to the problem. Since, $J_{uu} = 1$, $J_{ud} = -1$, $G_y = \begin{bmatrix}1 & 0\end{bmatrix}^T$ and $G_{yd} = \begin{bmatrix}0 & 1\end{bmatrix}^T$, the local optimal CV is then obtained using (7), which also minimizes the average loss locally for this problem[4]

$$c_{\text{lav}} = y_1 - y_2 \tag{15}$$

To maintain $c_{\text{lav}} = 0$ through feedback, the corresponding input has two solutions. The one satisfies the nominal operating condition, $u^* = 0$ is

$$u_{\text{lav}} = 2 - 2\sqrt{1-d} \tag{16}$$

Accordingly, the loss is then

$$L_{\text{lav}}(d) = \frac{1}{2}(2 - 2\sqrt{1-d} - d)^2$$

Hence, the expectation of the loss over the entire range is

$$\begin{aligned} E[L_{\text{lav}}(d)] &= \frac{1}{2}\int_{-1}^{1}\frac{1}{2}\left[8(1-d) + d^2 - 4(2-d)\sqrt{1-d}\right]\mathrm{d}d \\ &= 0.0183 \end{aligned}$$

Next, the global CV selection problem in (11) is to be solved for the toy example. For simplicity, consider the CV to be a linear combination of measurements. It can be parameterized as

$$c_{\text{gav}} = \phi(w,y) = y_1 + w_1 y_2 + w_2 \tag{17}$$

When $c_{\text{gav}} = 0$, the corresponding input, which is most close to $u^* = 0$ is

$$u_{\text{gav}} = \frac{2}{w_1}\left(-1 + \sqrt{1 - w_1(w_2 + dw_1)}\right) \tag{18}$$

Inserting (18) into (10) and solve it for optimization problem (11), we obtain the following results using Matlab Symbolic Math and Optimization Toolboxes:

$$\begin{aligned} w_1 &= -0.9231 \\ w_2 &= 0.0705 \end{aligned}$$

with the minimized average loss

$$\min_{w} L_{\text{gav}}(w) = 0.00278 \tag{19}$$

The result shows that the minimum global loss is much less than the local counterpart even though CVs of both methods are linear with the same number of parameters. The globally optimal linear CV is

$$c_{\text{gav}} = y_1 - 0.9231 y_2 + 0.0705 \tag{20}$$

which minimizes the average loss over the entire disturbance space.

## IV. A Regression Approach for Globally Optimal CV Selection

The toy example is solved analytically through finding the relationships between the loss expectation $E[L]$ and the CV parameters, $w$. However, finding analytical solution may not be tractable for general self-optimizing control problems. Therefore, it is necessary to develop a numerically effective approach to solve the globally optimal CV selection problem. In the following, we propose a regression approach to approximate the NCO of the optimization problem using CVs so that when these CVs are perfectly controlled at zero, the loss is proportional to the regression error. More specifically, let $u_w$ the control input corresponding to $c_w = \phi(w,y,n) = 0$. Then, the cost function $J(u,d)$ at a specific $d$ can be represented as the Taylor expansion around $u_w$,

$$J(u,d) = J_w + \eta_w^T \Delta u + 0.5\Delta u^T S_w \Delta u \tag{21}$$

where $\Delta u = u - u_w$, $J_w = J(u_w, d)$,

$$\eta_w = \left.\frac{\partial J(u,d)}{\partial u}\right|_{u=u_w} \tag{22}$$

and

$$S_w = \left.\frac{\partial \eta(u,d)}{\partial u}\right|_{u=u_w} \tag{23}$$

The gradient function can also be expanded similarly.

$$\eta(u,d) = \eta_w + S_w \Delta u \tag{24}$$

At the optimal point, $u_{\rm opt}(d)$, the above expansion becomes

$$J(u_{\rm opt}, d) = J_w + \eta_w^T \Delta u_{\rm opt} + 0.5 \Delta u_{\rm opt}^T S_w \Delta u_{\rm opt} \tag{25}$$

where $\Delta u_{\rm opt} = u_{\rm opt} - u_w$. Equivalently,

$$\begin{aligned} L_w(d,n) &= J_w - J(u_{\rm opt}, d) \\ &= -\eta_w^T \Delta u_{\rm opt} - 0.5 \Delta u_{\rm opt}^T S_w \Delta u_{\rm opt} \end{aligned} \tag{26}$$

Furthermore, at the optimal point, the NCO is

$$\eta(u_{\rm opt}, d) = \eta_w + S_w \Delta u_{\rm opt} = 0$$

Therefore,

$$\Delta u_{\rm opt} = -S_w^{-1} \eta_w \tag{27}$$

This leads to

$$L_w(d,n) = 0.5 \eta_w^T S_w^{-1} \eta_w \tag{28}$$

Note, a similar result on the loss has been derived in [9], however, for local loss around the nominally optimal point, whilst the result derived in (28) is globally valid.

Although loss derived in (28) can be directly used in (9) for global CV selection, the requirement of $S_w$ at each $d$ and $n$ is prohibitive. However, if CVs are represented as a summation of the NCO and an approximation error as follows,

$$c_w = \eta_w - \epsilon_w \tag{29}$$

then the approximation error, $\epsilon_w = \eta_w$ because $c_w = 0$ is perfectly controlled. Therefore, an upper bound of the global loss can be derived using the approximation error,

$$L_w(d,n) \le 0.5 M \|\epsilon_w\|_2^2 \tag{30}$$

where

$$M = \max_{d \in \mathcal{D}, n \in \mathcal{N}} \bar{\lambda}(S_w) \tag{31}$$

with $\bar{\lambda}(\cdot)$ denoting the maximum eigenvalue of a matrix.

To simplify the optimization problem in (11) further, the continuous operating region specified by $\mathcal{D}$ and $\mathcal{N}$ is discretized in $N$ sampling points, $d_i \in \mathcal{D}$ and $n_i \in \mathcal{N}$ for $i = 1, \ldots, N$. To avoid solving the perfect CV control equation, $\phi(w, y) = 0$, $u$ is also sampled at $N$ points within a feasible range, $\underline{u} \le u_i \le \bar{u}$, accordingly. With these three independent variable samples, the gradient, $\eta_i = \eta(u_i, d_i)$, $y_i = y(u_i, d_i) + n_i$ can be calculated from the process model. Then the following regression problem is to be solved in order to determine the optimal CV parameters,

$$\min_w \frac{1}{2} (\phi(w, y_i) - \eta_i)^T (\phi(w, y_i) - \eta_i) \tag{32}$$

This is a nonlinear least squares problem. The famous Levenberg-Marquardt algorithm [10] is available to solve this problem efficiently. Meanwhile, there are many nonlinear model structures are available in the literature as well, which can be adopted for $\phi(w, y)$, such as the polynomial and Gaussian kernel models. For simplicity, only linear regression is considered in the formulation bellow, for which an analytical solution can be derived.

The general form of linear $\phi(w, y)$ is given as follows.

$$c_w = Hy + b \tag{33}$$

with the parameter vector, $w = \begin{bmatrix} {\rm vec}(H)^T & b^T \end{bmatrix}^T$, where $H \in \mathbb{R}^{n_u \times n_y}$, $b \in \mathbb{R}^{n_u}$ and ${\rm vec}(\cdot)$ represents a matrix to be arranged in a vector.

Using the linear combination CVs given in (33), $n_u$ CVs can be calculated independently for each CV to approximate one of elements in the gradient vector, $\eta$. Therefore, for simplicity, in the following development, it is assumed that $n_u = 1$.

Denote

$$\eta = \begin{bmatrix} \eta_1 & \cdots & \eta_N \end{bmatrix}^T \tag{34}$$

$$Y = \begin{bmatrix} y_1 & \cdots & y_N \\ 1 & \cdots & 1 \end{bmatrix}^T \tag{35}$$

$$w = \begin{bmatrix} H & b \end{bmatrix}^T \tag{36}$$

Then the regression problem (32) is equivalent to a linear regression problem

$$\varepsilon = \min_w \frac{1}{2} (Yw - \eta)^T (Yw - \eta) \tag{37}$$

The least squares solution to the problem can be analytically obtained as

$$w_{\rm opt} = (Y^T Y)^{-1} Y^T \eta \tag{38}$$

The corresponding minimum total regression cost is

$$\varepsilon = \frac{1}{2} \eta^T (I - Y(Y^T Y)^{-1} Y^T) \eta \tag{39}$$

The corresponding global average loss by adopting this linear combination CV can be bounded by $L_{\rm gav} \le M\varepsilon$. Note although above solution is derived for linear combinations as CV, it can be easily extended to polynomial form by expanding matrix $Y$ with higher order terms, as illustrated in case studies below.

## V. Case Studies

### A. Toy example continued

Applying proposed regression method above, the variation ranges for $u$ and $d$ ($u, d \in \begin{bmatrix} -1 & 1 \end{bmatrix}$) are both discretized into 10 parts equally resulting total $N = 11^2 = 121$ sample points. At each sampling point, the measurements and gradient $J_u$ are calculated, then we obtained the matrix $Y$ and vector $\eta$. A linear LS regression is simply performed using (38), which results in the following CV

$$c_{w1} = y_1 - 0.9809 y_2 + 0.09809 \tag{40}$$

To calculate the average loss, the control input for $c_w$ is found to be

$$u_{w1} = 2.039 - 2.0389\sqrt{1.0962 - 0.9621 d} \tag{41}$$

Inserting (41) into (12) and (10) to yield the expectation of loss calculated as

$$E[L_{w1}(d)] = 0.00375 \tag{42}$$

The average loss for $c_{w1}$ has been significantly reduced as compared with local method and is only slightly bigger than optimal 0.00278, which is analytically searched in Section III. The convenience of proposed regression method is that it directly determines $w$ via numerical approach, avoiding the requirement of representing input $u$ in terms of $d$ and $w$, which is inevitable and hard to access in analytical way. To demonstrate the powerful usage of proposed method, a second order polynomial regression is further performed by adding terms $y_1y_2$, $y_1^2$ and $y_2^2$ into $Y$, the CV is found to be

$$c_{w2} = y_1 - y_2 + 0y_1y_2 + 0.25y_1^2 + 0y_2^2 + 0 \tag{43}$$

The corresponding control input $u_{w2}$ and the expected loss are interestingly found to be $u_{w2} = d$ and $\min_w L_{w2}(d) = 0$, respectively. It means that the process can achieve perfect self-optimizing control under any disturbance! Note, if we specify a quadratic form of $\phi(y, w)$ in the first place and solve it analytically, the solution process will be very complicated and almost prohibitive.

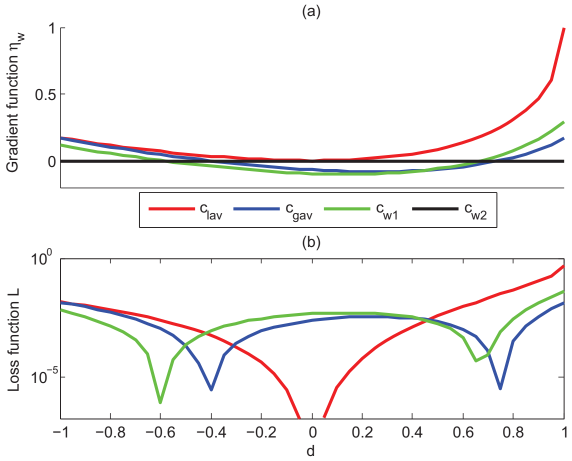


Fig. 2. Comparisons for different CVs for the toy example: (a) gradient function $\eta_w$ against $d$; (b) Loss function $L$ against $d$

Figure 2 (a) and (b) show the loss performances of various CVs when they are perfectly controlled at set-points 0. Compared with $c_{\text{lav}}$, globally optimal $c_{\text{gav}}$ is able to steer the gradient $\eta_w$ closer to 0 and minimize the loss over the entire disturbance range. $c_{w1}$ is suboptimal but its shape is approximately similar to $c_{\text{gav}}$, hence the loss is only slightly bigger. Moreover, $c_{w2}$ achieves perfect self-optimizing control and its curves overlap with desired 0 horizon line in the figures (not shown in semi-logarithmic plot (b)). Although for this problem $c_{w2}$ can also be derived through arrangements for model equations by eliminating $u$ and $d$, this may be hardly possible for other more complex problems, whereas regression method provides a very simple and efficient alternative, as illustrated in the reactor case study below.

### B. Exothermic Reactor

Self-optimizing control for the exothermic reactor has been previously studied by several researchers [3], [9], [11]. The reactant A is fed into a continuous stirred-tank reactor (CSTR) and undergoes a reversible exothermic reaction in the CSTR. The inlet temperature, concentrations of A and product B in the feed are denoted as $T_i$, $C_{Ai}$ and $C_{Bi}$ respectively, the outlet temperature, concentrations of unreacted A and product B in the outlet stream are denoted as $T$, $C_A$ and $C_B$ respectively. The schematic of exothermic reactor process is shown in Figure 3.

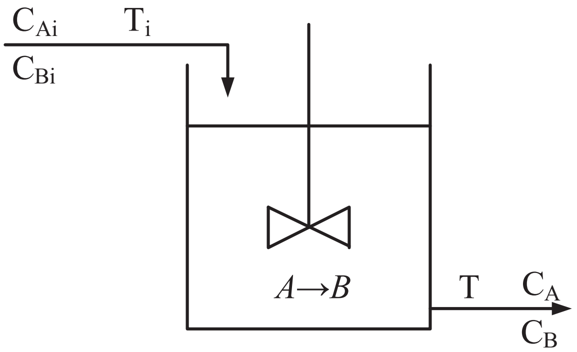


Fig. 3. Exothermic reactor process

The first principle models are composed of differential equations for mass and energy balances

$$\frac{dC_A}{dt} = \frac{1}{\tau}(C_{Ai} - C_A) - r \tag{44}$$

$$\frac{dC_B}{dt} = \frac{1}{\tau}(C_{Bi} - C_B) + r \tag{45}$$

$$\frac{dT}{dt} = \frac{1}{\tau}(T - T_i) + 5r \tag{46}$$

where $\tau = 60$s is the residence time, and $r$ is the rate of reaction which is dependent on process variables

$$r = 5000e^{-\frac{10000}{1.987T}}C_A - 10^5e^{-\frac{15000}{1.987T}}C_B \tag{47}$$

The classifications for manipulated variable, available measured variables and disturbances are given as

$$u = [T_i] \tag{48}$$

$$y = [C_A \quad C_B \quad T \quad T_i]^T \tag{49}$$

$$d = [C_{Ai} \quad C_{Bi}]^T \tag{50}$$

The anticipated noises for measured variables are ±0.01 mol/L for concentrations $C_A$ and $C_B$, ±0.5K for temperatures $T$ and $T_i$. The allowable sets for disturbances are considered as $0.5 \leq C_{Ai} \leq 1.5$ and $0 \leq C_{Bi} \leq 0.5$.

The operational objective of exothermic reactor process is to maximize the economic profit, which is equivalent to minimizing a cost function

$$J = -20090C_B + (0.1657T_i)^2 \tag{51}$$

where the first term of $J$ is the negative profit of product $B$ and the latter represents the cost of heating the input stream. The nominal values for process variables are given in Table 1. The operational degree of freedom for this case is 1, so only 1 CV is to be selected to square the control system.

TABLE I
PROCESS VARIABLES AND NOMINAL VALUES

| Variable | Physical description | Nominal value | Unit |
|---|---|---|---|
| $C_A$ | Outlet A concentration | 0.498 | $\mathrm{mol} \cdot \mathrm{L}^{-1}$ |
| $C_B$ | Outlet B concentration | 0.502 | $\mathrm{mol} \cdot \mathrm{L}^{-1}$ |
| $T$ | Outlet steam temperature | 426.803 | K |
| $T_i$ | Inlet steam temperature | 424.292 | K |
| $C_{Ai}$ | Inlet A concentration | 1.0 | $\mathrm{mol} \cdot \mathrm{L}^{-1}$ |
| $C_{Bi}$ | Inlet B concentration | 0 | $\mathrm{mol} \cdot \mathrm{L}^{-1}$ |
| $J$ | Economic objective | $-5149.3$ | \$ |

For this example, an analytical CV solution to minimize the globally overall loss is not available. Alternatively, the regression approach is use to select the optimal CV. Samples for regression are collected as follows: the possible variation range of each independent variable ( $C_{Ai}$, $C_{Bi}$ and $T_i$) is discretized equally into 10 parts, therefore, $11^3$ points of data are generated. Each point of data contains four measured variables and $dJ/du$, which is calculated using input perturbations and finite differences. The variation ranges for disturbances are defined earlier, the range for $T_i$ is considered as $\begin{bmatrix}380\mathrm{K} & 450\mathrm{K}\end{bmatrix}$, which is determined from later observations that the majority of optimum $u$ falls into this interval.

Least square regression is performed straightforward to get a combination model as CV with measurements as predictors. A linear and second order polynomial regression result in the following CVs,

$$c_1 = -772.2 - 184.3y_1 + 152.0y_2 - 7.4y_3 + 9.3y_4 \tag{52}$$

$$\begin{aligned} c_2 = {} & 1131.0 + 324.3y_1 - 1298.8y_2 - 105.5y_3 + 100.3y_4 \qquad (53) \\ & + 12.0y_1y_2 - 44.7y_1y_3 + 43.3y_1y_4 + 15.6y_2y_3 \\ & - 12.0y_2y_4 + 81.2y_1^2 - 30.5y_2^2 + 3.4y_4^2 \end{aligned}$$

with an $R^2$ regression index of 0.9464 and 0.9997, respectively. As a comparison, the methods proposed by Kariwala et al. [4] to minimize average local loss and Alstad et al. [6] using extended null space method are also applied to current example. These CVs are

$$\begin{aligned} c_{\mathrm{Kariwala}} = {} & 0.76y_1 - 0.65y_2 - 6.58 \times 10^{-5}y_3 \qquad (54) \\ & - 0.0051y_4 + 2.15 \\ c_{\mathrm{Alstad}} = {} & -171.76y_1 + 145.23y_2 + 0.0083y_3 \qquad (55) \\ & + 1.15y_4 - 479.18 \end{aligned}$$

A Monte Carlo experiment for 100 set of randomly generated disturbances within expected ranges is conducted and the results are shown in Table 2. The results in Table 2 show that self-optimizing performance can be significantly improved by using proposed method. Compared to $c_{\mathrm{Kariwala}}$ and $c_{\mathrm{Alstad}}$, the average losses (11.57 and 10.21) are furthermore effectively reduced by $c_1$ and $c_2$ (3.07 and 0.0896) because proposed method minimize the loss globally in the entire operation region.

TABLE II
AVERAGE ECONOMIC LOSSES WITH DIFFERENT COMBINATION CV

| CV | Average loss | Maximum loss | Standard deviation |
|---|---|---|---|
| $c_1$ | 3.07 | 15.39 | 3.26 |
| $c_2$ | 0.0896 | 0.97 | 0.15 |
| $c_{\mathrm{Kariwala}}$ | 11.57 | 52.78 | 12.11 |
| $c_{\mathrm{Alstad}}$ | 10.21 | 71.33 | 14.76 |

## VI. CONCLUSIONS

This paper presented a formulation for selecting CVs with globally average loss minimization in the context of SOC. A toy example is provided to illustrate the procedure of selecting globally optimal CVs for self-optimizing control. Compared with existing local SOC methods, which are only accurate in small neighborhood around the nominal point, new approach is advantageous that the global loss is minimized in an average sense. To circumvent the difficulty for solving the complicated and maybe non-convex optimization problem, a numerical solution is proposed alternatively to find optimal CVs, which is based on least squares regression to approximate the NCOs. The usage and effectiveness of the numerical method are demonstrated by the toy example and a more realistic exothermic reactor case study. The later shows that the loss over the entire operating range is significantly reduced comparing with the results obtained before using local SOC approaches.

## REFERENCES

[1] S. Skogestad, “Plantwide control: The search for the self-optimizing control structure.” *J. Proc. Control*, vol. 10, no. 5, pp. 487–507, 2000.

[2] I. J. Halvorsen, S. Skogestad, J. C. Morud, and V. Alstad, “Optimal selection of controlled variables,” *Ind. Eng. Chem. Res.*, vol. 42, no. 14, pp. 3273–3284, 2003.

[3] V. Kariwala, “Optimal measurement combination for local self-optimizing control,” *Ind. Eng. Chem. Res.*, vol. 46, no. 11, pp. 3629–3634, 2007.

[4] V. Kariwala, Y. Cao, and S. Janardhanan, “Local self-optimizing control with average loss minimization,” *Ind. Eng. Chem. Res.*, vol. 47, no. 4, pp. 1150–1158, 2008.

[5] V. Alstad and S. Skogestad, “Null space method for selecting optimal measurement combinations as controlled variables,” *Ind. Eng. Chem. Res.*, vol. 46, no. 3, pp. 846–853, 2007.

[6] V. Alstad, S. Skogestad, and E. S. Hori, “Optimal measurement combinations as controlled variables,” *J. Proc. Control*, vol. 19, no. 1, pp. 138–148, 2009.

[7] E. S. Hori and S. Skogestad, “Selection of controlled variables: Maximum gain rule and combination of measurements,” *Ind. Eng. Chem. Res.*, vol. 47, no. 23, pp. 9465–9471, 2008.

[8] L. Ye, Y. Cao, Y. Li, and Z. Song, “Approximating necessary conditions of optimality as controlled variables,” *Ind. Eng. Chem. Res.*, p. submitted, 2012.

[9] J. Jaschke and S. Skogestad, “NCO tracking and self-optimizing control in the context of real-time optimization,” *Journal of Process Control*, vol. 21, pp. 1407–1416.

[10] D. Marquardt, “An algorithm for least-squares estimation of nonlinear parameters.” *SIAM Journal of Applied Mathematics*, vol. 11, pp. 431–441, 1963.

[11] V. Alstad, “Studies on selection of controlled variables,” Ph.D. dissertation, Norwegian University of Science and Technology, 2005.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Branch and Bound Method for Globally Optimal Controlled Variable Selection

Vinay Kariwala
ABB Global Industries & Services Ltd
Mahadevpura, Bangalore 560048, India
Email: vinay.kariwala@in.abb.com

Lingjian Ye
Ningbo Institute of Technology
Zhejiang University
315100, Ningbo, Zhejiang, China
Email: lingjian.ye@gmail.com

Yi Cao
School of Engineering
Cranfield University
Cranfield, Bedford MK43 0AL, UK
Email: Y.Cao@cranfield.ac.uk

*Abstract*—**For selection of controlled variables (CVs) in self-optimizing control, various criteria have been proposed in the literature. These criteria are derived based on local linearization of the process model and the necessary conditions of optimality (NCO) at a nominally optimal operating point. Recently, a novel CV selection framework has been proposed by Ye et al. [1] by converting the CV selection problem into a regression problem to approximate the NCO globally over the entire operation region. In this approach, linear combinations of a subset of available measurements are used as CVs. The subset selection problem is combinatorial in nature redering the application of the globally optimal CV selection method to large-scale processes difficult. In this work, an efficient branch and bound (BAB) algorithm is developed to handle the computational complexity associated with the selection of globally optimal CVs. The proposed BAB algorithm identifies the best measurement subset such that the regression error in approximating NCO is minimized. This algorithm is applicable to the general regression problem. The efficiency and effectiveness of the proposed BAB algorithm is demonstrated through a binary disdillation column case study.**

## I. Introduction

The selection of controlled variables (CVs) from available measurements is an important task during the design of control systems. For CV selection, several methods have been proposed in the literature. Skogestad [2] introduced the concept of self-optimizing control for selection of CVs based on process economics. In this approach, CVs are selected such that in presence of disturbances, the loss incurred in implementing the operational policy by holding the selected CVs at constant setpoints is minimal, as compared to the use of an online optimizer. The advantages of self-optimizing control approach for CV selection has been demonstrated through several case studies; see *e.g.* [3] for an overview.

The choice of CVs based on the general non-linear formulation of self-optimizing control requires solving large-dimensional non-convex optimization problems [2]. To quickly pre-screen alternatives, exact local methods with worst-case [4] and average loss minimization [5] have been proposed. These local methods are useful for selecting a subset or linear combinations of available measurements as CVs, where the latter approach provides lower losses. Recently, explicit solutions to the problem of finding locally optimal measurement combinations have been proposed [6], [5], [7]. Hu *et al.* [8] have proposed a local method to explicitly handle the input and output constraints during CV selection.

The available CV selection criteria are derived based on local linearization of the process model. Recently, a globally optimal CV selection framework has been proposed by Ye *et al.* [1], [9]. In this framework, the CV synthesis problem is converted into a regression problem using CVs as measurement combinations to approximate the Necessary Conditions of Optimality (NCO) globally over the entire operation region. It has been proven that the average loss is globally minimized when the regression error is minimal over the entire operation region and the measurement combinations as CVs are perfectly controlled at zero. A number of linear and nonlinear regression models have been proposed to approximate the NCO. Case studies showed that all these models are able to significantly reduce the loss, as compared to using CVs designed by using existing local methods.

In general, CV selection is a combinatorial problem. For selection of individual measurements as CVs, a number of efficient branch and bound (BAB) approaches, called bidirectional BAB (B$^3$), have been developed for various local criteria to address the combinatorial issue [18], [10], [11]. These BAB algorithms are not required for the selection of individual measurements as globally optimal CVs, as the selection is not combinatorial any more because approximations of individual gradients are not correlated and can be solved separately. However, to select measurement combinations, the combinatorial difficulty still exists for the global CV selection problem.

It is known that the use of combinations of a few measurements as CVs often provide similar loss as the case where combinations of all available measurements are used [6], [5], [7], [9]. Though the former approach results in control structures with lower complexity, it gives rise to a combinatorial optimization problem involving selection of measurements, whose combinations can be used as CVs. For local self-optimizing control methods, partially bidirectional BAB (PB$^3$) methods have been proposed to solve this combinatorial problem efficienctly [10], [11]. In this work, the framework is further extended to measurement subset selection for synthesis of globally optimal CVs chosen as linear combinations of measurements. It is proven that the selection criterion is equivalent to a quadratic problem, for which a standard BAB algorithm [12] exists. The standard algorithm is improved into a downwards BAB algorithm. The efficiency and effectiveness of the proposed BAB algorithm is demonstrated through a distillation case study [13].

## II. Local Methods for Self-Optimizing Control

Consider that the steady-state economics of the plant is characterized by the scalar objective function $J(u, d)$, where $u \in \mathbb{R}^{n_u}$ and $d \in \mathbb{R}^{n_d}$ are inputs and disturbances, respectively. The optimal operation policy is to update $u$ according to $d$, which usually requires the use of an online optimizer. For this case, let the optimal value of the objective function be denoted as $J_{\text{opt}}(d)$. A simpler strategy involves indirect adjustment of $u$ using a feedback controller. In this case, the feedback controller manipulates $u$ to hold the CVs $c$ close to their specified setpoints. Here, in addition to $d$, $J$ is also affected by the error $e$ in implementing the constant setpoint policy, which results due to uncertainty and measurement noise. The suboptimal objective functional value under the second strategy is denoted as $J_c(e, d)$. Then, the worst-case and average losses due to the use of the suboptimal strategy are given as

$$\begin{aligned} \text{Worst-case loss} &= \max_{e\in\mathcal{E}} \max_{d\in\mathcal{D}} \left(J_{\text{opt}}(d) - J_c(e, d)\right) && (1) \\ \text{Average loss} &= E[J_{\text{opt}}(d) - J_c(e, d)] && (2) \\ & && (3) \end{aligned}$$

where $\mathcal{D}$ and $\mathcal{E}$ represent the sets of allowable disturbances and implementation errors, respectively, and $E$ is the expectation operator. Self-optimizing control is said to occur, when we can achieve an acceptable loss by holding the CVs close to their setpoints without the need to reoptimize when disturbances occur [2]. Based on this concept, the appropriate CVs can be selected by comparing the losses for different alternatives.

As mentioned earlier, the use of nonlinear formulation of self-optimizing control is difficult. Hence, some local methods were developed to estimate the losses defined in (1) and (2) by linearising the process model around the normally optimal operating point as follows:

$$y = G^y\, u + G^y_d\, W_d\, d + W_e\, e \tag{4}$$

where $y \in \mathbb{R}^{n_y}$ denotes the process measurements and $e \in \mathbb{R}^{n_y}$ denotes the implementation error, which results due to measurement and control errors. Here, the diagonal matrices $W_d$ and $W_e$ contain the expected magnitudes of disturbances and implementation error, respectively. The CVs $c \in \mathbb{R}^{n_u}$ are given as

$$c = H\, y = G\, u + G_d\, W_d\, d + H\, W_e\, e \tag{5}$$

where $H$ is a selection or combination matrix and

$$G = H\, G^y, \quad G_d = H\, G^y_d \tag{6}$$

It is assumed that $G \in \mathbb{R}^{n_u \times n_u}$ is invertible. This assumption is necessary for integral control. When $d$ and $e$ are assumed to be uniformly distributed over the set

$$\left\| \begin{bmatrix} d^T & e^T \end{bmatrix}^T \right\|_2 \leq 1 \tag{7}$$

the local worst-case and average losses are given as [4], [5]:

$$\begin{aligned} L_{\text{worst}}(H) &= 0.5\bar{\sigma}^2 \left( J_{uu}^{1/2} (H\, G^y)^{-1} H\, Y \right) && (8) \\ L_{\text{average}}(H) &= \frac{1}{6(n_y + n_d)} \left\| J_{uu}^{1/2} (H\, G^y)^{-1} H\, Y \right\|_F^2 && (9) \end{aligned}$$

where $\bar{\sigma}$ and $\|\cdot\|_F$ denote the maximum singular value and Frobenius norm, respectively, and

$$Y = \begin{bmatrix} (G^y\, J_{uu}^{-1}\, J_{ud} - G^y_d)\, W_d & W_e \end{bmatrix} \tag{10}$$

with $J_{uu} = \frac{\partial^2 J}{\partial u^2}$ and $J_{ud} = \frac{\partial^2 J}{\partial u \partial d}$, evaluated at the nominal operating point. In comparison with worst-case loss, the selection of CVs is preferred through minimization of average loss, as the worst-case may not occur frequently in practice [5].

When individual measurements are selected as CVs, $H$ can be considered to be a selection matrix. Instead of using individual measurements, it is possible to use combinations of measurements as CVs. For this case, Alstad et al. [7] has recently proposed an explicit expression for $H$, which minimizes the $L_{\text{average}}$ in (9) and is given as

$$H^T = (YY^T)^{-1} G^y ((G^y)^T (YY^T)^{-1} G^y)^{-1} J_{uu}^{1/2} \tag{11}$$

As shown by Kariwala et al. [5], the $H$ in (11) also minimizes $L_{\text{worst}}$ in (8). The locally optimal combinations of all the available measurements, which can be used as CVs can be found using (11).

## III. Globally Optimal Method

The local methods [4], [5] are based on linearization around the nominally optimal operating point. Therefore, the identified CVs are only locally optimal. To derive globally optimal solution CVs, it is assumed that the NCO is approximated by CVs and the CVs are perfectly controlled at zero. Then, the loss, $L(d)$ for a particular disturbances $d$, due to the approximation error, $\epsilon(d)$ can be expressed as [1], [9]:

$$L(d) = 0.5\epsilon(d)^T H(d)^{-1} \epsilon(d) \tag{12}$$

where $H(d)$ is the reduced Hessian of the cost function evaluated at point where the CV, $c(d)$ is perfectly controlled corresponding to particular disturbance, $d$, whilst $\epsilon(d) = g(d) - c(d)$, where $g(d)$ is the reduced gradient evaluated at the same point.

The average loss over the entire operation region, $D$ can be represented as,

$$\bar{L} = E_{d\in D} L(d) \approx \frac{1}{2N} \sum_{i=1}^{N} \epsilon(d_i)^T H(d_i)^{-1} \epsilon(d_i) \tag{13}$$

where $d_i \in D$, $i = 1, \ldots, N$ are $N$ samples of disturbances in $D$ and $E$ is the expectation operator.

According to (13), the loss minimization is equivalent to a least squares regression problem to minimize the weighted residual, $H^{-1/2}\epsilon$. However, due to the difficulty and reliability to evaluate the reduced Hessian for every $d_i \in D$, $H(d_i)$ can be replaced by a constant matrix, *e.g.* the identity matrix or $H$ evaluated at nominal value of $d$. Then the regression problem can be set up as discussed next.

Let the entire operation region be sampled by $N$ points for independent variables (input and disturbance), $u = \begin{bmatrix} u_1 & \cdots & u_N \end{bmatrix}$ and $d = \begin{bmatrix} d_1 & \cdots & d_N \end{bmatrix}$. The corresponding measurement values and the reduced gradient values are $y = \begin{bmatrix} y_1 & \cdots & y_N \end{bmatrix}$ and $g = \begin{bmatrix} g_1 & \cdots & g_N \end{bmatrix}$, respectively. The CV is parameterized by $\theta$ at each sampling point as:

$$c_i = f_\theta(y_i), \ i = 1, \ldots, N \tag{14}$$

where $f_\theta(\cdot)$ is the parameterized regression function of measurements, which can be either linear or nonlinear, such as polynomial or Gaussian. Then the optimal CV, $c^* = f_{\theta^*}(y)$ is determined by adjusting $\theta$ to minimize the regression error $\epsilon = f_\theta(y_i) - g_i$ as follows:

$$\min_\theta \ \frac{1}{2N} \sum_{i=1}^{N} (f_\theta(y_i) - g_i)^T H^{-1} (f_\theta(y_i) - g_i) \tag{15}$$

For linear regression, $f_\theta(y) = \theta y$, *i.e.* $g = \theta y + \epsilon$. Then the regression problem in (15) can be solved analytically; see Section IV-B for details. In principle, it is possible to parametrize the CVs in terms of all the available measurements. Control systems with lower complexity can be obtained by using a subset of available measurements to parametrize the CVs, which often provides similar loss as the case where CVs are chosen to be functions of all the available measurements [6], [5], [7], [9]. The selection of the subset of measurements is a combinatorial optimization problem, which makes the application of this method difficult to large-scale processes. The BAB framework used to overcome this difficulty is presented in the next section.

## IV. Branch and bound method

### A. General principle

Let $X_r = \{x_i | i = 1, 2, \cdots, r\}$, be an $r$-element set. A subset selection problem with the selection criterion $\phi$ involves finding the optimal solution, $X_n^*$, such that

$$\phi(X_n^*) = \max_{X_n \subset X_r} \phi(X_n) \tag{16}$$

For this problem, the number of alternatives is $\mathcal{C}_r^n = \frac{r!}{(r-n)!n!}$, which grows very quickly with $r$ and $n$ rendering exhaustive search unviable. A BAB approach can provide globally optimal solution for the subset selection problem in (16) without exhaustive search. In this approach, the original problem (node) is divided (branched) into several non-overlapping subproblems (sub-nodes). If any of the $n$-element solution of a sub-problem cannot lead to the optimal solution, the sub-problem is not evaluated further (pruned), else it is branched again. The pruning of sub-problems allows the BAB approach to gain efficiency in comparison with exhaustive search.

The available BAB methods for subset selection can be classified as downwards [12], [14], [15], [16], [17] and upwards [18], [10], [11] BAB methods based on the search direction. For the regression problem associated with globally optimal CV selection, the downwards BAB approach is applicable and is discussed next. The reader is referred to [12], [14], [17] for details on downwards BAB method.

In a downwards BAB approach, each node is represented by $X_s = F_f \cup C_c$, where $s = f + c$ and, $F_f$ and $C_c$ denote the fixed and candidate sets, respectively. Here, the subscript denote the size of the set. The relationship between the fixed and candidate sets of a node and its $i^{\text{th}}$ sub-node (branching rule) is given as follows:

$$F_{f_i}^i = F_f \cup \{x_1, \cdots, x_{i-1}\}; \ C_{c_i}^i = C_c \setminus \{x_1, \cdots, x_i\} \tag{17}$$

where $F_{f_i}^i$ and $C_{c_i}^i$ denote the fixed and candidate sets of the $i^{\text{th}}$ sub-node and $i = 1, 2, \cdots, n - f + 1$. An example of the solution tree obtained by recursively applying the branching rule in (17) is shown in Figure 1. For the root node in this solution tree, we have $F_f = \emptyset$ and $C_c = X_r$. The label of the nodes denote the element being removed from $X_s$. The solution tree has $\mathcal{C}_n^r$ terminal nodes, which represent different $n$-element subsets of $X_r$.

To describe the pruning principle, let $\mathcal{X}$ denote the ensemble of all $n$-element subsets, which can be obtained using (17), *i.e.*

$$\mathcal{X} = \{F_f \cup C_c / X_{f+c-n} | X_{f+c-n} \in C_c\} \tag{18}$$

and $\underline{\phi}(F_f \cup C_c)$ be the upper bound on $\phi$ computed over all elements of $\mathcal{X}$, *i.e.*

$$\underline{\phi}(F_f \cup C_c) = \max_{X_n \in \mathcal{X}} \phi(X_n) \tag{19}$$

Assume that $B$ is a lower bound of the globally optimal criterion, *i.e.* $B \leq \phi(X_n^*)$. Then,

$$\phi(X_n) < \phi(X_n^*) \quad \forall X_n \in \mathcal{X}, \quad \text{if} \ \ \underline{\phi}(F_f \cup C_c) < B \tag{20}$$

Hence, any $X_n \in \mathcal{X}$ cannot be optimal and can be pruned without further evaluation, if $\underline{\phi}(F_f \cup C_c) < B$.

Although pruning of nodes using (20) results in an efficient BAB algorithm, further efficiency can be gained by performing pruning on the sub-nodes directly. This happens as the lower bounds for different sub-nodes are related and can be computed together from $\underline{\phi}(F_f \cup C_c)$ resulting in computational efficiency. For $x_i \in C_c$, the $i^{\text{th}}$ sub-node can be pruned if

$$\underline{\phi}(F_f \cup C_c / x_i) < B \tag{21}$$

For a BAB method involving pruning of sub-nodes, branching needs to be carried on sub-node level as well, which requires choosing a decision element to branch upon. Here, the decision element is selected as the element with largest $\underline{\phi}(F_f \cup C_c / x_i)$ among all $x_i \in C_c$ (best-first search).

### B. Application to CV Selection using Regression

The linear regression model is given as:

$$b = A\theta + \epsilon \tag{22}$$

where $b$ are the observations or measurements, matrix $A$ contains the regressors, $\theta$ are the unknown parameters and $\epsilon$ is the noise. Under the assumption that $\epsilon$ is independently

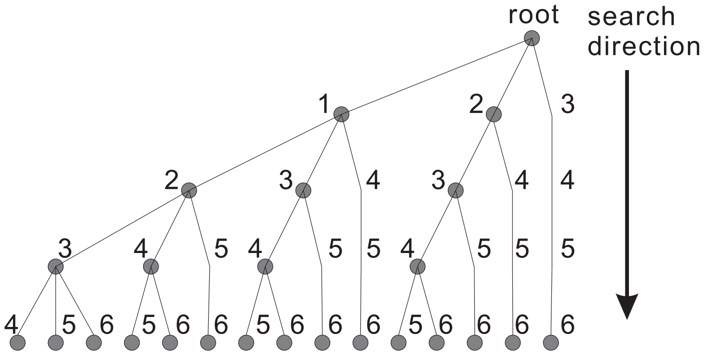


Fig. 1. Solution tree for selecting 2 out of 6 elements

and identically distributed (i.i.d.), it is well known that the unbiased estimate of $\theta$ is given as; see *e.g.* [19],

$$\hat{\theta} = \left(A^T A\right)^{-1} A^T b \tag{23}$$

With the estimate of $\theta$ given in (23), the predicted values of observations are $\hat{b} = A\hat{\theta} = A\left(A^T A\right)^{-1} A^T b$ and the prediction error $e$ is given as

$$e = b - \hat{b} = Pb \tag{24}$$

where $P = \left(I - A\left(A^T A\right)^{-1} A^T\right)$. Then, sums of squares of errors (SSE) can be computed as

$$SSE = e^T e = b^T P^T P b = b^T P b \tag{25}$$

where the last identity follows as $P$ is an idempotent matrix [19]. The SSE can be further expressed as

$$SSE = b^T b - b^T A\left(A^T A\right)^{-1} A^T b \tag{26}$$

As the first term in (26) is constant, the variables can be selected by minimizing SSE or equivalently maximizing $b^T A\left(A^T A\right)^{-1} A^T b$. Let us define $y = A^T b$ and $C = A^T A$. Now, the subset selection can be performed by solving the following optimization problem:

$$\max_{X_n \subset X_r} \phi(X_n) = \mathbf{y}_{X_n}^T (\mathbf{C}_{X_n, X_n})^{-1} \mathbf{y}_{X_n} \tag{27}$$

where $X_r = \{1, 2, \cdots, r\}$, $\mathbf{y}_{X_n}$ denotes the elements of $\mathbf{y}$ with indices in $X_n$ and $\mathbf{C}_{X_n, X_n}$ represents the principal submatrix of $\mathbf{C}$ with rows and columns indexed by $X_n$. Note that a similar combinatorial optimization problem is considered in [20], where the objective function instead needs to be minimized.

The use of BAB for solving the optimization problem in (27) requires an upper bound on the selection criteria, calculated over the ensemble $\mathcal{X}$ in (18). This upper bound is derived in the next proposition.

***Proposition 1:*** Consider a node with fixed set $F_f$ and candidate set $C_c$. For $\mathcal{X}$ in (18),

$$\phi(F_f \cup C_c) \geq \max_{X_n \in \mathcal{X}} \phi(X_n) \tag{28}$$

Proposition 1 implies that the non-optimal nodes can be pruned using $\phi(F_f \cup C_c)$ as the upper bound. To gain further efficiency by pruning the sub-nodes directly, we relate the selection criteria of a node with its sub-nodes in the next proposition.

***Proposition 2:*** Consider a node with fixed set $F_f$ and candidate set $C_c$. Let $S = F_f \cup C_c$. For $x_i \in C_c$, $i = 1, 2, \cdots, c$,

$$\phi(S \setminus x_i) = \phi(S) - \alpha_i^2 / \delta_i \tag{29}$$

where

$$\alpha_i \quad = \quad \mathbf{z}_i^T \mathbf{y}_S \tag{30}$$

whilst $\mathbf{z}_i^T$ and $\delta_i$ are the $i$th row and $(i, i)$th element of $\mathbf{C}_{S,S}^{-1}$ respectively.

The evaluation of (29) requires inversion of only one matrix $\mathbf{C}_{S,S}$, which is the same for all $x_i \in C_c$. Thus, the use of (29) to obtain the selection criteria for all sub-nodes together is computationally more efficient than directly evaluating the selection criteria for every node. In summary, the following BAB algorithm can be used for subset selection for regression.

***Algorithm 1:*** Initialize $f = 0$, $F_f = \emptyset$, $C_c = X_r$, $\phi(F_f) = 0$ and $B = 0$. Call the following recursive algorithm:

1) If $\phi(F_f \cup C_c) > B$, prune the current node and return, else perform the following steps.
2) Calculate $\alpha_i$ in (30) $\forall i \in C_c$. Prune the subsets with $\phi(F_f \cup C_c) - \alpha_i^2 < B$.
3) If $f = n$ or $f + c = n$, go to next step. Otherwise, generate the $c$ sub-nodes according to the branching rule in (17) and call the recursive algorithm in Step 1 for each sub-node. Return to the caller after the execution of the loop finishes.
4) Find $J_{\max} = \phi(F_f \cup C_c) - \max_{i \in C_c} \alpha_i^2$. If $J_{\max} > B$, update $B = J_{\max}$. Return to the caller.

## V. Binary Distillation Column Case Study

To evaluate the efficiency of proposed BAB algorithm for selecting globally optimal CVs, we test the performance of BAB algorithm on a binary distillation column [13]. All tests are conducted on a PC running Windows 7 SP1 with Intel Core i3-2100 3.10GHz processor, 8GB RAM using Matlab R2011a.

The objective is to minimize the relative steady-state deviations of the distillate ($z^L_{\text{top}}$) and bottoms ($z^H_{\text{btm}}$) compositions from their nominal values, *i.e.*

$$J = \left( \frac{z^H_{\text{top}} - z^H_{\text{top,s}}}{z^H_{\text{top,s}}} \right)^2 + \left( \frac{z^L_{\text{btm}} - z^L_{\text{btm,s}}}{z^L_{\text{btm,s}}} \right)^2 \tag{31}$$

where the superscripts $L$ and $H$ refer to the light and heavy components and the nominal steady-state values are $z^H_{\text{top,s}} = z^L_{\text{btm,s}} = 0.01$ (99% purity). The distillation column has 4 manipulated variables: reflux flow rate ($L$), vapor boilup ($V$), distillate flow rate ($D$) and bottoms flow rate ($B$). Because the levels of top condenser and bottom reboiler need to be stabilized, which consumes two degrees of freedom. We select $D$ and $B$ to control the levels, which is also referred as LV configuration for distillation column control, therefore, two degrees of freedom are remained for composition control. The main disturbances are feed flow rate ($F$), feed composition ($z_F$) and vapor fraction of feed ($q_F$), which are allowed to vary between $1 \pm 0.2$, $0.5 \pm 0.1$ and $1 \pm 0.1$, respectively. The top and bottom compositions are not measured online and thus two CV's needs to be identified for indirect control of the compositions. It is considered that the temperatures on 41 trays ($y_1, \ldots, y_{41}$, counting from bottom to top) are measured with an accuracy of $\pm 0.5°$C, whose combinations can be used as CVs for implementation of self-optimizing control strategy.

Data samples for NCO regression are generated as follows: each independent variable is evenly divided into 5 parts within its variation range. The variation range for disturbances are defined earlier. Variation range for reflux flow rate $L$ is chosen as $1 \pm 10\%$ at its nominal value and vapor boilup V is bounded within $(L-(1-q_F)F, L+q_F F)$ in order to let $0 < B,D < F$. For each scenario, temperatures at each tray are calculated and the two NCO components $J_L$ and $J_V$, which refer to the gradient of $J$ with respect to $L$ and $V$, respectively, are also obtained using finite difference method. Therefore, $6^5 = 7776$ samples are collected for regression. Because the number of candidate measurements for regression is large, we apply proposed BAB algorithm to choose an appropriate subset and determine globally optimal CVs. Similar computation performances are observed for regressing $J_L$ and $J_V$, as summarized in Figure 2.

Figure 2 (a) and (b) show that using full set of measurements as predictors for NCOs is not necessary. When $n > 5$, the SSE can only be slightly reduced. A trade-off between SSE, which is directly related to the overall economic loss, and the number of measurements used has to be made. Therefore, we can choose $n = 6$ and get CV models for self-optimizing control of this column. Figure 2 (c), (d), (e) and (f) show the computation time and number of node evaluations and demonstrate the usefulness and effectiveness of proposed BAB algorithm. Brute force cannot handle such a large problem, whereas proposed BAB algorithm solves it successfully. It takes about 1260s to complete all the selection tasks, which is reasonable for off-line computing. Largest computation time is seen for $n = 15$, which takes about 100s for selecting 15 out of 41 measurements. Overall, proposed algorithm is practically appealing, as the algorithm makes it possible to reduce the overall operation cost and meanwhile, reduces the investment for hardware sensors (*e.g.* temperature sensors for the column).

Furthermore, the economic losses associated with the CVs obtained by using the proposed BAB algorithm are evaluated and compared with those associated with CVs obtained in previous work [11] through local self-optimizing control. It is found that $n = 6$ results in a good trade-off between self-optimizing performance and investment cost of sensors, where both CVs are designed using 6 measurements.

The self-optimizing method using local average criterion [5] is adopted for comparison. To derive the CVs, linear model and Hessian matrices at the nominal point are obtained through finite difference. We also choose $n = 6$ and apply the PB$^3$ algorithm [11], [22] to get the best measurement subset.

A Monte Carlo experiment with 100 sets of randomly generated disturbances within their allowable ranges is used. The feedback control actions are implemented to maintain CVs at 0 and objective function $J$ is calculated, whose optimal value is expected to be 0. The average and maximal objective costs for regression method are 2.222 and 82.766, respectively, while the average and maximal objective costs for local method are 170.411 and 1577.474, respectively. This is because that the distillation process is strongly nonlinear, therefore, local method fails to characterize the process in wider operation range, whereas the regression method is advantageous in this respect resulting in significantly enhanced self-optimizing performance.

## VI. Conclusions

In the context of self-optimizing control, a novel branch and bound (BAB) algorithm is proposed for selecting globally optimal controlled variables (CVs) based on the method of regression for necessary conditions of optimality. The BAB algorithm aims to identify the best measurement set to minimize the sums of squares of errors (SSE) and use the set to construct the optimal CVs. Numerical tests using a practical binary distillation column case study show the efciency and effectiveness of this algorithm. It is pointed out that the proposed algorithm is applicable to the general linear regression problem as well as other statistical problems. The proposed downwards algorithm is efficient for problems where a few among many candidate variables need to be discarded. However, for problems where a few variables needs to be selected from many candidate variables, the computational expense incurred by the algorithm is still large. To this end, upwards and bidirectional BAB algorithms are currently being developed and will be presented in an extended paper.

## References

[1] L. Ye, Y. Cao, Y. Li, and Z. Song, "A data-driven approach for selecting controlled variables," in *to be presented at the 8th Intl. Symposium on ADCHEM*, Singapore, 2012.

[2] S. Skogestad, "Plantwide control: The search for the self-optimizing control structure." *J. Proc. Control*, vol. 10, no. 5, pp. 487–507, 2000.

[3] G. P. Rangaiah and V. Kariwala, *Plantwide Control: Recent Developments and Applications*. Chichester, UK: John Wiley & Sons, 2012.

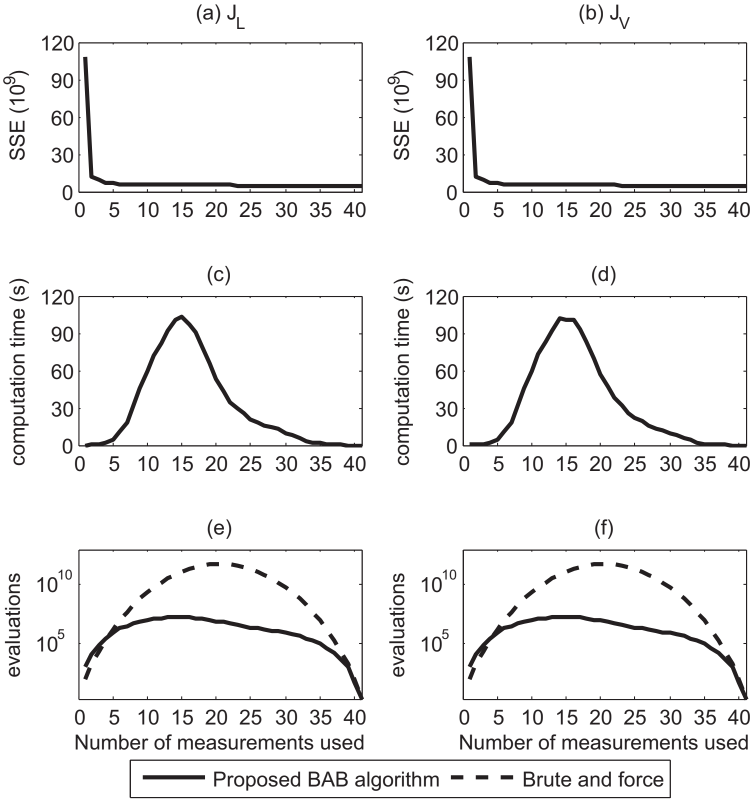


Fig. 2. BAB performance for the column case study: (a) regression SEE for $J_L$; (b) regression SEE for $J_V$; (c) computation time for $J_L$; (b) computation time for $J_V$; (e) Evaluations for $J_L$; (f) Evaluations for $J_V$

[4] I. J. Halvorsen, S. Skogestad, J. C. Morud, and V. Alstad, "Optimal selection of controlled variables," *Ind. Eng. Chem. Res.*, vol. 42, no. 14, pp. 3273–3284, 2003.

[5] V. Kariwala, Y. Cao, and S. Janardhanan, "Local self-optimizing control with average loss minimization," *Ind. Eng. Chem. Res.*, vol. 47, no. 4, pp. 1150–1158, 2008.

[6] V. Kariwala, "Optimal measurement combination for local self-optimizing control," *Ind. Eng. Chem. Res.*, vol. 46, no. 11, pp. 3629–3634, 2007.

[7] V. Alstad, S. Skogestad, and E. S. Hori, "Optimal measurement combinations as controlled variables," *J. Proc. Control*, vol. 19, no. 1, pp. 138–148, 2009.

[8] W. Hu, L. M. Umar, G. Xiao, and V. Kariwala, "Local self-optimizing control of constrained processes," *J. Proc. Contr.*, vol. 22, pp. 488–493, 2012.

[9] L. Ye, Y. Cao, Y. Li, and Z. Song, "Approximating necessary conditions of optimality as controlled variables," *Ind. Eng. Chem. Res.*, p. submitted, 2012.

[10] V. Kariwala and Y. Cao, "Bidirectional branch and bound for controlled variable selection: Part II. Exact local method for self-optimizing control," *Comput. Chem. Eng.*, vol. 33, no. 8, pp. 1402–1412, 2009.

[11] ——, "Bidirectional branch and bound for controlled variable selection: Part III. Local average loss minimization," *IEEE Trans. Ind. Informat.*, vol. 6, no. 1, pp. 54–61, 2010.

[12] P. Narendra and K. Fukunaga, "A branch and bound algorithm for feature subset selection," *IEEE Trans. Comput.*, vol. C-26, pp. 917–922, 1977.

[13] S. Skogestad, "Dynamics and control of distillation columns - A tutorial introduction," *Trans. IChemE Part A*, vol. 75, pp. 539–562, 1997.

[14] B. Yu and B. Yuan, "A more efficient branch and bound algorithm for feature selection," *Pattern Recognition*, vol. 26, pp. 883–889, 1993.

[15] P. Somol, P. Pudil, F. Ferri, and J. Kittler, "Fast branch and bound algorithm in feature selection," in *Proceedings of World Multiconference on Systemics, Cybernetics and Informatics*, B. Sanchez, J. Pineda, J. Wolfmann, Z. Bellahsense, and F. Ferri, Eds., vol. VII, Orlando, Florida, USA, 2000, pp. 1646–651.

[16] X.-W. Chen, "An improved branch and bound algorithm for feature selection," *Pattern Recognition Letters*, vol. 24, pp. 1925–1933, 2003.

[17] Y. Cao and P. Saha, "Improved branch and bound method for control structure screening," *Chem. Engg. Sci.*, vol. 60, no. 6, pp. 1555–1564, 2005.

[18] Y. Cao and V. Kariwala, "Bidirectional branch and bound for controlled variable selection: Part I. Principles and minimum singular value criterion," *Comput. Chem. Engng.*, vol. 32, no. 10, pp. 2306–2319, 2008.

[19] R. Johnson and D. Wichern, *Applied multivariate statistical analysis*. Prentice hall Upper Saddle River, NJ, 2002, vol. 4.

[20] V. Kariwala, P. Odiowei, Y. Cao, and T. Chen, "A branch and bound method for isolation of faulty variables through missing variable analysis," *Journal of Process Control*, vol. 20, no. 10, pp. 1198–1206, 2010.

[21] Y. Cao, D. Rossiter, and D. H. Owens, "Output effectiveness and scaling sensitivity for secondary measurement selection," *Transaction of IChemE, Part A*, vol. 76, pp. 849–854, 1998.

[22] Y. Cao and V. Kariwala, "B3AV," MATLAB File Exchange, November 2009, available at http://www.mathworks.com/matlabcentral/fileexchange/25870.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Control Structure Selection for Optimal Operation of a Heat Exchanger Network

Johannes Jäschke
Department of Chemical Engineering
Norwegian University of Science and Technology
NTNU, 7491 Trondheim, Norway
Email: jaschke@chemeng.ntnu.no

Sigurd Skogestad
Department of Chemical Engineering
Norwegian University of Science and Technology
NTNU, 7491 Trondheim, Norway
Email: skoge@chemeng.ntnu.no

***Abstract*—We consider the control structure design for a heat exchanger network (HEN), where a stream is split into parallel lines which are heated individually before they are merged together again. The objective is to find a control structure which maximizes the final temperature. We consider two scenarios, where (Scenario 1) the flow rates and the heat transfer coefficients are considered as disturbances, and (Scenario 2) where the hot stream temperatures are treated as additional disturbances. In both scenarios it is found that controlling linear measurement combinations gives very good performance, and that including flow measurements in the combinations gives little advantage over using only combinations of temperature measurements.**

***Index Terms*—Control structure selection, Self-optimizing control, Heat exchanger networks, Optimization**

## I. Introduction

With growing markets and limited natural resources, it is increasingly necessary to use the available resources in the best possible way. In many cases, this can be directly translated into re-using energy. Important tools for re-using energy are heat exchanger networks (HENs), which are operated such that a maximum amount of energy is transferred from one set of process streams to another.

It has been shown in [1], [2], that if there are no stream splits in the heat exchanger network, then optimal operation of the HEN can be considered and modelled as a linear program (LP), where the optimal operation point is always at constraints. In particular this means that the available degrees of freedom should be used for

1) keeping required target temperatures at their setpoints
2) meeting active constraints (i.e. fully opening or closing of some bypasses and utilities)

Since the stream parameters such as temperatures and flow rates change under operation, the task of achieving optimal operation can be re-formulated as finding the set of active constraints corresponding to the current operating conditions.

However, when streams are split, the optimal operating point will generally no longer be at constraints; and the optimal split will not remain constant because of disturbances such as changing temperatures of the streams, changing flow rates, or change of heat transfer properties due to fouling.

One approach for adjusting the split optimally is to periodically solve a numerical optimization problem to find the optimal setpoint values for some set of controlled variables (real-time optimization (RTO)) [3]. A good choice of controlled variables (CVs) will reduce the necessity to update the setpoints, while a poor choice will require frequent setpoint updates to remain close to optimality.

The RTO approach relies heavily on a good process model and the ability to solve the optimization problem online. Furthermore it is necessary to obtain good estimates of the model parameters. All these factors render the online-optimization approach relatively expensive.

Therefore, it is desirable to find controlled variables, which do not have to be updated often. A control structure which, in spite of varying disturbances, gives an acceptable loss without the need to adjust the setpoints for the controlled variables is called “self-optimizing” [4]. In such a control structure the setpoints of the controlled variables need to be updated very infrequently, or not even at all. For a recent overview of available methods for control structure design using the self-optimizing control ideas, we refer to [5] and the literature cited therein.

The contribution of this work is to study the effect of different self-optimizing control structures for a HEN with stream split, and in particular to evaluate the necessity of relative expensive flow measurements. Our paper is structured such that in the next section we present some basic concepts for finding self-optimizing variables. In Section III we introduce the heat exchanger network, and Section IV contains the simulation scenarios and results. Our paper is closed with a discussion and conclusions in Section V.

## II. Controlled variable selection

In this paper we apply local controlled variable selection procedures from self-optimizing control [4], [6], [7]. We assume that optimal operation corresponds to minimizing a scalar cost function, and for local optimality around the nominal optimum, it can be approximated [8] as

$$\min_{\Delta u} J(\Delta u, \Delta d) := [\Delta u^T \; \Delta d^T] \begin{bmatrix} J_{uu} & J_{ud} \\ J_{ud}^T & J_{dd} \end{bmatrix} \begin{bmatrix} \Delta u \\ \Delta d \end{bmatrix}. \tag{1}$$

Here, $J$ is the scalar approximation of the nonlinear cost function, $\Delta u \in \mathbb{R}^{n_u}$ and $\Delta d \in \mathbb{R}^{n_d}$ denote the inputs and the disturbances in deviation variables, respectively. $J_{uu} = \frac{\partial^2 J}{\partial u^2}$ denotes the positive definite Hessian of the cost function with respect to the degrees of freedom $u$, and $J_{ud} = \frac{\partial^2 J}{\partial u \partial d}$, $J_{dd} = \frac{\partial^2 J}{\partial d^2}$ denote second order partial derivatives with respect to $[u, d]^T$ and $[d, d]^T$, respectively.

The model is linearized around the nominal optimum, such that

$$\Delta y = G^y \Delta u + G^y_d \Delta d + \Delta n^y, \tag{2}$$

where $\Delta y$ denotes the $n_y$-dimensional measurement vector, $G^y = \frac{\partial y}{\partial u}$ and $G^y_d = \frac{\partial y}{\partial d}$ are gain matrices of appropriate sizes, evaluated at the nominal operating point. Finally, $\Delta n^y \in \mathbb{R}^{n_y}$ denotes the measurement noise or implementation error.

### A. Exact local method

The goal is to find a set of controlled variables (CV)

$$\Delta c = H \Delta y, \tag{3}$$

where $\Delta c \in \mathbb{R}^{n_u}$, and $H$ is a measurement selection or combination matrix of size $n_u \times n_y$, such that controlling the CVs at constant setpoints, $\Delta c = 0$, results in acceptable loss

$$L = J(H, \Delta d, \Delta n^y) - J_{opt}(d) \tag{4}$$

in spite of varying disturbances $\Delta d$ and noise $\Delta n^y$. In [6] it is shown that solving the convex optimization problem

$$\begin{aligned} &\min_H ||HY||_F \\ &\text{s.t.} \\ &HG^y = J_{uu}^{1/2} \end{aligned} \tag{5}$$

leads to a $H$ which locally minimizes the average and worst case loss. Here $||\cdot||_F$ denotes the Frobenius norm,

$$Y = [FW_d \quad W_{n^y}] \tag{6}$$

denotes the augmented optimal sensitivity matrix where $F = \partial y^{opt} / \partial d$, and the matrices $W_d$ and $W_{n^y}$ denote diagonal scaling matrices of appropriate dimensions, with

$$\Delta d = W_d d' \quad \text{and} \quad \Delta n^y = W_{n^y} n^{y'}, \tag{7}$$

such that

$$\left|\left| [d' \, n^{y'}]^T \right|\right|_2 \leq 1. \tag{8}$$

The local average loss resulting from a given control structure represented by $H$ can be evaluated using ([6], [7])[1]

$$L = \frac{1}{2} \left|\left| J_{uu}^{1/2} (HG^y)^{-1} HY \right|\right|_F^2 \tag{9}$$

**Remark 1.** *The matrices $F = \frac{\partial y^{opt}}{\partial d}$ and $G^y$ can be obtained by re-optimizing the process and a finite difference approximation, or as done in this work, by using the efficient sensitivity calculation routines in sIPOPT [9]. The software provides the sensitivity of the KKT solution with respect to parameters (disturbances d). As sIPOPT re-uses the matrix factorizations from the NLP solver IPOPT [10], the sensitivities are obtained by a simple backsolve. Similarily, the inverse of the reduced Hessian $J_{uu}$ can be easily obtained from sIPOPT.*

### B. Null-space methods

If there is no noise, and we have sufficient independent measurements ($n_y \geq n_u + n_d$), then H may be chosen in the left null space of $F$. In this case we have $HF = 0$, and since $W_{n^y} = 0$, we also have that $HY = 0$. Hence the local loss from (9) is zero. In practice, however, there will always be measurement noise ($W_{n^y} \neq 0$), and the CVs obtained from the null space method will cause a loss, which can be unacceptably high. Alstad et al. [6] present the extended null-space method, which selects $H$ in the null-space of $F$ (optimal disturbance rejection), and uses the remaining degrees of freedom to minimize the effect of noise. We, however, aim at minimizing the combined effect of disturbances and noise. Therefore, we use Eq. (5) for determining candidates for controlled variables.

### C. Selecting subsets of measurements

The above expressions for optimal measurement combinations are valid for controlled variables $c$ which contain a given set of measurements $y$. However in practice it is often desirable to use subsets of the available measurements, because the performance does not improve significantly beyond including a certain number of measurements, while installation and maintenance costs increase.

For finding the best subsets of measurements, branch-and-bound methods have been developed [11], [12], [13]. Recently [14] proposed a method to implement certain structural constraints on $H$.

In this paper we will not dwell on the available methods, but rather investigate some sets of controlled variables for a HEN with a stream split. For finding sets of measurements, we use the freely available software `b3av.m` [13] for MATLAB™.

## III. Heat exchanger network

### A. Process description

We consider the HEN in Fig. 1, where a feed stream with temperature $T_0$ and heat capacity $w_0$ is split into 6 lines, which are heated independently before they are merged again. The structure of this HEN is the same as the structure of the HEN for feed pre-heating of the crude oil distillation unit at the Mongstad refinery in Norway. The objective is to adjust the feed split such that the end temperature is maximized,

$$\min J = -T_{end}. \tag{10}$$

This corresponds to minimizing the energy required for the crude oil distillation unit. The heat exchanger network has 30 measurable temperatures and 14 flows. Thus, the potentially

[1] Our definition of the average loss is slightly different from [7], where it is defined as $L = \frac{1}{6(n_y+n_d)} \left|\left| J_{uu}^{1/2} (HG^y)^{-1} HY \right|\right|_F^2$. Although this definition is formally correct, one may reduce the loss arbitrarily by adding (unused) measurements, $n_y \to \infty$. To avoid this, we define the average loss as in (9).

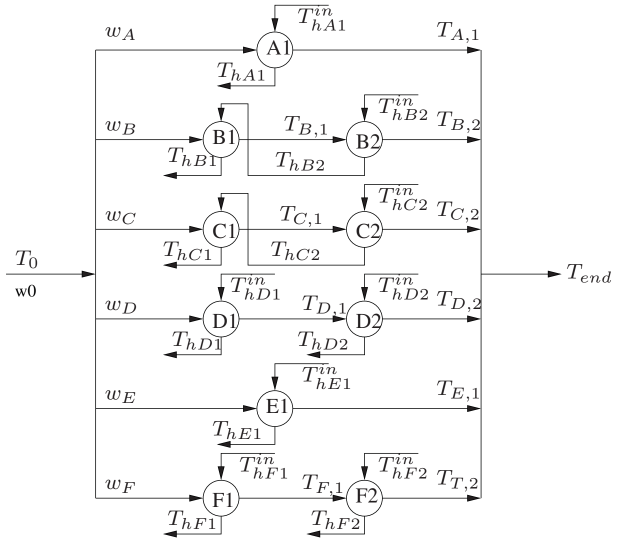


Fig. 1. Heat exchanger network

available measurement vector is

$$\begin{aligned} y = \big[ & T_0, T_{end}, T_{1A}, T_{h1A}, T^{in}_{h1A}, \\ & T_{2B}, T_{1B}, T_{h1B}, T_{h2B}, T^{in}_{h2B}, \\ & T_{2C}, T_{1C}, T_{h1C}, T_{h2C}, T^{in}_{h2C}, \\ & T_{2D}, T_{1D}, T_{h1D}, T_{h2D}, T^{in}_{12D}, T^{in}_{h2D}, \\ & T_{1E}, T_{h1E}, T^{in}_{h1E}, \\ & T_{2F}, T_{1F}, T_{h1F}, T_{h2F}, T^{in}_{h1F}, T^{in}_{h2F}, \\ & w_A, w_{h1A}, w_B, w_{h2B}, w_C, w_{h2C}, \\ & w_D, w_{h1D}, w_{h2D}, w_E, w_{h1E}, w_F, w_{h1F}, w_{h2F} \big]. \end{aligned} \tag{11}$$

We assume that all temperature measurements have an uncertainty (noise) of +/-1$°C$, that is $W_{n^{y_T}} = 1°C \; \mathrm{I}_{30\times30}$, where $\mathrm{I}_{30\times30}$ denotes the identity matrix of dimension 30 × 30.

The flow measurements are assumed to have an uncertainty of 2% of the nominal value, resulting in $W_{n^{y_w}} = diag(0.64, 0.96, 0.92, 1.85, 0.48, 0.72, 0.59, 2.37, 0.68, 0.77, 1.60, 1.11, 0.94, 3.50)$ kW/K. The resulting overall noise weighting matrix is

$$W_{n^y} = \begin{bmatrix} W_{n^{y_T}} & 0 \\ 0 & W_{n^{y_w}} \end{bmatrix}. \tag{12}$$

### B. Heat exchanger network model

The heat exchangers are modelled with simple energy balances. We present the model of the heat exchanger in the first line (A) in detail here; the heat exchangers on the other lines are modelled analogously. The main assumptions are incompressible fluids and constant specific heat capacities in all flows.

The energy balance around the cold and hot sides give

$$Q_{A1} = w_A(T_{A,1} - T_0) = w_{hA1}(T^{in}_{hA1} - T_{hA1}), \tag{13}$$

where $Q_{A1}$ denotes the heat, $w_A = m_A c_P$ is the heat capacity flow rate (product of the mass flow rate $m_{A,1}$ and the specific heat capacity $c_P$) of stream A, $w_{hA1}$ is the heat capacity flow rate of the hot stream $hA1$, and $T^{in}_X$ and $T_X$ denote the temperatures of stream $X$ at the inlet and outlet of the heat exchanger, respectively.

The transfered heat is calculated as

$$Q_{A1} = UA_{A1}\Delta T_{lm}, \tag{14}$$

where the logarithmic temperature difference is calculated using the Underwood approximation [15]:

$$(\Delta T_{lm})^{1/3} = \frac{1}{2}[(\Delta T_1)^{1/3} + (\Delta T_2)^{1/3}] \tag{15}$$

where $\Delta T_1 = T^{in}_{hA1} - T_{A,1}$ and $\Delta T_2 = T_{h,A1} - T_0$ denote the temperature differences on the two sides of the heat exchanger. This approximation is very close to the logarithmic mean temperature, while it has better numerical properties. In particular, it is still defined when $\Delta T_1 = \Delta T_2$, while the exact logarithmic mean temperature is not, which can cause problems during the optimization.

An energy balance around the mixer gives

$$T_{end} = \frac{1}{w_0} \sum_{i=A,B,\ldots,F} w_i T_i, \tag{16}$$

and a mass balance around the splitter:

$$w_0 = \sum_{i=A,B,\ldots,F} w_i. \tag{17}$$

### C. Degrees of freedom analysis and nominal optimum

Under operation, the hot flow rates and temperatures are given, as well as the feed temperature $T_0$. Since the total feed flow rate is fixed, the flow in 5 of 6 lines can be set freely. The sixth flow is given from the law of mass conservation. Thus, there are 5 degrees of freedom, and we need to find $n_c = 5$ controlled variables. The nominal parameter values are listed in Table I, and the corresponding optimal measurement values are given in Table II.

## IV. Simulation Scenarios

We consider two scenarios: In Scenario 1, the varying disturbances are assumed to be only the flow rates of the feed and the hot streams, and the heat transfer $UA$ in all heat exchangers. In Scenario 2 the temperatures of incoming streams are considered as additional disturbances.

For both scenarios, we consider different choices of CVs, including single measurements and selected measurement combinations. Their performance is compared in terms of the average loss $L$ calculated from equation (9).

### A. Scenario 1. Flow rate and heat transfer disturbances

The flow and heat transfer disturbances for Scenario 1 are given together with their numerical values in Table III. In the next sections, we consider the performance of some selected control structures.

TABLE I
NOMINAL PARAMETER VALUES

| Symbol | Value | Unit | Description |
|---|---|---|---|
| $T_0$ | 133 | °C | Feed temperature |
| $w_0$ | 225 | kW/K | Feed heat capacity |
| $T_{h1A}^{in}$ | 300 | °C | Hot stream 1A temperature |
| $w_{h1A}$ | 48 | kW/K | Hot stream 1A heat capacity |
| $UA_{A1}$ | 131 | kW/K | Heat transfer coefficient times area A1 |
| $wh1B$ | 92.5 | kW/K | Hot stream 1B heat capacity |
| $UA_{B1}$ | 102.7 | kW/K | Heat transfer coefficient times area B1 |
| $T_{h2B}^{in}$ | 270.3 | °C | Hot stream 2B temperature |
| $UA_{B2}$ | 88.64 | kW/K | Heat transfer coefficient times area B2 |
| $wh1C$ | 35.8 | kW/K | Heat capacity stream 1C |
| $UA_{C1}$ | 84 | kW/K | Heat transfer coefficient times area C1 |
| $T_{h2C}^{in}$ | 245 | °C | Hot stream 2C temperature |
| $UA_{C2}$ | 133.6 | kW/K | Heat transfer coefficient times area C2 |
| $T_{12D}^{in}$ | 226 | °C | Hot stream 1D temperature |
| $w_{h1D}$ | 118.5 | kW/K | Heat capacity stream 1D |
| $UA_{D1}$ | 132.8 | kW/K | Heat transfer coefficient times area D1 |
| $T_{h2D}^{in}$ | 273.8 | °C | Hot stream 2D temperature |
| $w_{h2D}$ | 33.9 | kW/K | Heat capacity stream 2D |
| $UA_{D2}$ | 41.6 | kW/K | Heat transfer coefficient times area D2 |
| $T_{h1E}^{in}$ | 256.4 | °C | Hot stream 1E temperature |
| $w_{h1E}$ | 79.9 | kW/K | Heat capacity stream 1E |
| $UA_{E1}$ | 190.9 | kW/K | Heat transfer coefficient times area E1 |
| $T_{h1F}^{in}$ | 203 | °C | Hot stream 1F temperature |
| $w_{h1F}$ | 47.2 | kW/K | Heat capacity stream 1F |
| $UA_{F1}$ | 49.4 | kW/K | Heat transfer coefficient times area 1F |
| $T_{h2F}^{in}$ | 248 | °C | Hot stream 2F temperature |
| $w_{h2F}$ | 175.1 | kW/K | Heat capacity stream 2F |
| $UA_{F2}$ | 224.1 | kW/K | Heat transfer coefficient times area 2F |

TABLE II
NOMINAL OPTIMAL MEASUREMENT VALUES

| Variable | Value | Variable | Value |
|---|---|---|---|
| $T_{end}$ | 255.676 °C | $T_{h2D}$ | 247.350 °C |
| $T_{1A}$ | 283.102 °C | $T_{1E}$ | 251.430 °C |
| $T_{h1A}$ | 200.358 °C | $T_{h1E}$ | 199.429 °C |
| $T_{2B}$ | 261.281 °C | $T_{2F}$ | 244.393 °C |
| $T_{1B}$ | 231.824 °C | $T_{1F}$ | 164.622 °C |
| $T_{h1B}$ | 206.632 °C | $T_{h1F}$ | 165.818 °C |
| $T_{h2B}$ | 255.680 °C | $T_{h2F}$ | 222.716 °C |
| $T_{2C}$ | 243.184 °C | $w_A$ | 31.8637 kW/K |
| $T_{1C}$ | 213.075 °C | $w_B$ | 45.9091 kW/K |
| $T_{h1C}$ | 171.672 °C | $w_C$ | 23.8251 kW/K |
| $T_{h2C}$ | 224.962 °C | $w_D$ | 29.4667 kW/K |
| $T_{2D}$ | 254.175 °C | $w_E$ | 38.4364 kW/K |
| $T_{1D}$ | 223.745 °C | $w_F$ | 55.4990 kW/K |
| $T_{h1D}$ | 203.435 °C | | |

*1) Open loop operation:* Simply leaving the split at the nominal optimal values results in an average loss of

$$L_{OL1} = 0.3147°C \tag{18}$$

Although this value is quite low, it may be reduced further by controlling the right variables.

*2) Controlling a combination of all measurements:*

*a) Exact local method:* Using a combination $c = Hy$ of all measurements $y$, for our scenario the average loss is

$$L_{All1} = 0.0284°C \tag{19}$$

This is the locally best loss that can be achieved in this scenario using a linear measurement combination.

TABLE III
FLOW AND HEAT TRANSFER DISTURBANCES WITH MAGNITUDES FOR SCENARIO 1

| Variable | Weighting in $W_d$ | Description |
|---|---|---|
| $w_0$ | 11.25 kW/K | Feed heat capacity |
| $w_{h1A}$ | 2.40 kW/K | Hot stream 1A heat capacity |
| $UA_{A1}$ | 13.10 kW/K | Heat transfer coefficient times area A1 |
| $UA_{B1}$ | 10.20 kW/K | Heat transfer coefficient times area B1 |
| $w_{h2B}$ | 4.60 kW/K | Hot stream 2B heat capacity |
| $UA_{B2}$ | 8.90 kW/K | Heat transfer coefficient times area B2 |
| $UA_{C1}$ | 8.40 kW/K | Heat transfer coefficient times area C1 |
| $w_{h2C}$ | 1.80 kW/K | Hot stream 2C heat capacity |
| $UA_{C2}$ | 13.40 kW/K | Heat transfer coefficient times area C2 |
| $w_{h1D}$ | 5.90 kW/K | Heat capacity stream 1D |
| $UA_{D1}$ | 13.30 kW/K | Heat transfer coefficient times area D1 |
| $w_{h2D}$ | 1.70 kW/K | Heat capacity stream 2D |
| $UA_{D2}$ | 4.20 kW/K | Heat transfer coefficient times area D2 |
| $w_{h1E}$ | 4.00 kW/K | Heat capacity stream 1E |
| $UA_{E1}$ | 19.10 kW/K | Heat transfer coefficient times area E1 |
| $w_{h1F}$ | 2.40 kW/K | Heat capacity stream 1F |
| $UA_{F1}$ | 4.90 kW/K | Heat transfer coefficient times area 1F |
| $w_{h2F}$ | 8.80 kW/K | Heat capacity stream 2F |
| $UA_{F2}$ | 22.40 kW/K | Heat transfer coefficient times area 2F |

*b) Null space method:* If we neglect the noise and select $H$ in the left null space of $F$, the performance of the plant with noise is dependent on which basis vectors of the null space are chosen for $H$. For two different sets of basis vectors $H_1$ and $H_2$ we have

$$\begin{aligned} L_{All1}^{NullSpace}(H_1) &= 19.1711°C \\ L_{All1}^{NullSpace}(H_2) &= 2.0702°C \end{aligned} \tag{20}$$

The loss is very different for different choices of basis vectors of the left null space of $F$. If there was no noise, $W_d = 0$, we would have $L_{All1}^{NullSpace}(H_1) = L_{All1}^{NullSpace}(H_2) = 0$. We do not further consider the issue of selecting the best basis vectors using the extended null-space method [6], because this is beyond the scope of this paper, and because the exact local method (Eq. (5)) selects the best (optimal) $H$ in presence of noise and disturbances.

*3) Using subsets of measurements:* The average loss when using different subsets of measurements in linear variable combinations is shown in blue in Fig. 2. The largest reduction in the loss comes when increasing the number of measurements from 5 to 6. The reason for this is that the disturbances may enter the plant in all 6 lines. When using only 5 measurements from 5 lines, the disturbance in the 6th line is not detected. Adding a 6th measurement from the remaining line makes it possible to detect disturbances affecting this line, too, and therefore reduces the loss significantly.

Above a certain number of measurements the loss does not significantly decrease further. From Fig. 2 it is evident that the loss decreases very little for more measurements than 15, and above 20 measurements the decrease in loss is hardly observable[2]. The loss is very low for this scenario, and in

[2] It decreases slightly for some more measurements, but for the last 9 measurements it is completely flat. These last 9 measurements are the inlet temperatures of the feed and the hot streams, which are assumed to be constant. Thus they do not provide any further information for disturbance rejection in Scenario 1.

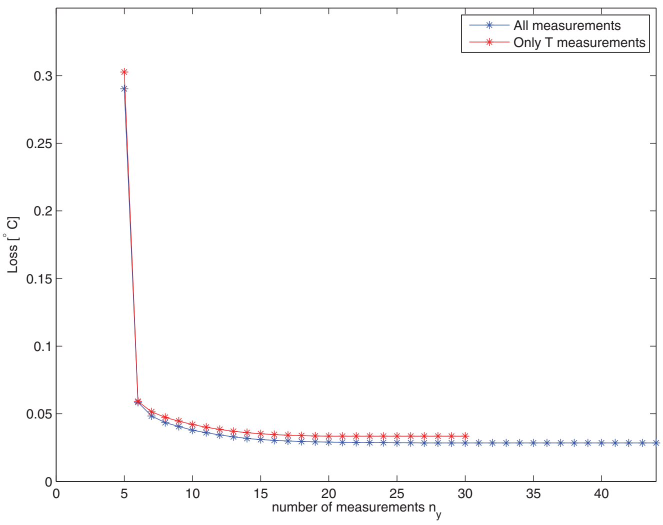


Fig. 2. Scenario 1 (Flow rate and $UA$ disturbances): Loss for best combinations of different number of measurements $n_y$

most practical cases, open loop operation will be sufficient. However, if one wants to optimize performance further, it does not seem reasonable invest in measurement equipment for more than 10-15 measurements.

*a) Only temperature measurements:* Typically, not all measurements are equally capital intensive. E.g. temperature measurements are generally much cheaper than flow measurements. Therefore we next consider cases with different sets of temperature measurements. The relationship between the average loss and the number of temperature measurements included in the CVs is given in red in Fig. 2. The trend is very similar to when flow measurements are used, too. After a sharp initial decrease, the curve flattens out quickly. The average loss resulting from using all temperature measurements is

$$L_{AllT1} = 0.0335°C. \tag{21}$$

This is only little higher than when using all measurements, including flows. Note that also here, the constant inlet temperatures do not contribute to minimizing the loss.

*b) Only specific temperature measurements:* In many practical cases we may have that the outlet temperatures of the cold streams are measured, i.e.

$$y = [T_{1A}, T_{1B}, T_{2B}, T_{1C}, T_{2C}, T_{1D}, T_{2D}, T_{1E}, T_{1F}, T_{2F}]^T \tag{22}$$

In this case the average loss using a linear combination is

$$L_{outTemp1} = 0.0639°C \tag{23}$$

*c) Single temperature measurements:* The simplest strategy (beside open loop operation) is to control as many single temperatures as there are degrees of freedom. The best set gives a loss of

$$L_{Single\,T1} = 0.3027°C \tag{24}$$

TABLE IV
TEMPERATURE DISTURBANCES WITH MAGNITUDES FOR SCENARIO 2

| Variable | Weighting in $W_d$ | Description |
|---|---|---|
| $T0$ | $13.0°C$ | Feed temperature |
| $T^{in}_{h1A}$ | $30.0°C$ | Hot stream 1A temperature |
| $T^{in}_{h2B}$ | $27.0°C$ | Hot stream 2B temperature |
| $T^{in}_{h2C}$ | $24.5°C$ | Hot stream 2C temperature |
| $T^{in}_{12D}$ | $22.6°C$ | Hot stream 1D temperature |
| $T^{in}_{h2D}$ | $27.4°C$ | Hot stream 2D temperature |
| $T^{in}_{h1E}$ | $25.6°C$ | Hot stream 1E temperature |
| $T^{in}_{h1F}$ | $20.3°C$ | Hot stream 1F temperature |
| $T^{in}_{h2F}$ | $24.8°C$ | Hot stream 2F temperature |

with the measurements

$$y = [T_{1A}, T_{2B}, T_{h2C}, T_{2D}, T_{1E}]^T. \tag{25}$$

### *B. Scenario 2. Flow, heat transfer, and temperature disturbances*

In this more general scenario, in addition to varying flow rates and heat transfer coefficients, as in the previous section, we assume that the stream temperatures are also varying in the ranges given in Table IV. We proceed as in the previous section with considering the different cases with different measurements involved.

*1) Open loop operation:* Operating the heat exchanger in open loop with the nominal values for the splits, results in a quite high average loss of

$$L_{OL2} = 2.2836°C. \tag{26}$$

*2) Controlling a combination of all measurements:*

*a) Exact local method:* Using all available measurements, the exact local method gives an average loss of

$$L_{All2} = 0.0318°C. \tag{27}$$

which is just a little bit higher than the loss in Scenario 1.

*b) Null space method:* If we neglect the noise and select $H$ in the left null space of $F$, such that $HF = 0$, the actual loss with noise is again dependent on the choice of basis vectors for $H$. Two examples are

$$\begin{aligned} L^{NullSpace}_{All1}(H_1) &= 19.1711°C \\ L^{NullSpace}_{All1}(H_2) &= 2.0702°C \end{aligned} \tag{28}$$

The loss is the same as in Scenario 1, since the same columns of the null space have been selected, and the left null space for Scenario 1 is contained in the null space for Scenario 2.

*3) Using subsets of measurements:* Plotting the number of measurements used in the controlled variable versus the loss, Figure 3 in blue, indicates that it is not necessary to include all measurements, since the loss is not reduced significantly when including more than 10 measurements. Moreover, we find that above 10 measurements including flow measurements does not give any significant advantage in terms of loss over using only temperature measurements.

*a) Only temperature measurements:* Using only temperatures, the best combination results in a very small a loss of

$$L_{AllT2} = 0.0438°C. \tag{29}$$

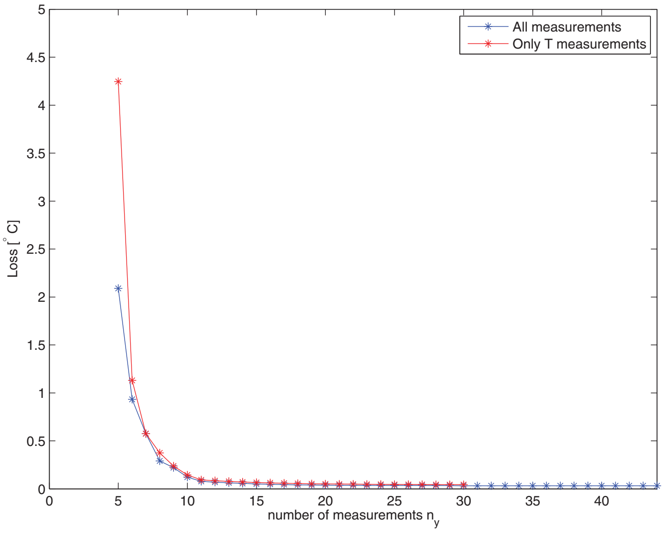


Fig. 3. Scenario 2 (Flow rate, Temperature and $UA$ disturbances): Loss for best combinations of different number of measurements $n_y$

*b) Only specific temperatures measurements:* Using only the outlet temperatures of the cold streams, as in Eq. (22) gives a loss of

$$L_{outTemp2} = 16.9079°C, \tag{30}$$

which is very high. However, if we add some disturbance measurements, namely the stream inlet temperatures,

$$\begin{aligned} y =[&T_0, T_{1A}, T_{h1A}^{in}, T_{2B}, T_{h2B}^{in}, T_{2C}, T_{h2C}^{in}, T_{2D}, T_{1D}, \\ &T_{h1D}^{in}, T_{h2D}^{in}, T_{1E}, T_{h1E}^{in}, T_{2F}, T_{1F}, T_{h1F}^{in}, T_{h2F}^{in}] \end{aligned} \tag{31}$$

the loss is reduced to

$$L_{outTemp2a} = 0.0961°C \tag{32}$$

Thus, measuring the disturbances (the stream inlet temperatures) reduces the loss orders of magnitudes.

*c) Single temperature measurements:* Controlling single temperature measurements, leads to a loss of

$$L_{SingleT2} = 4.2459°C \tag{33}$$

with the measurement set

$$y = [T_{h1A}, T_{1B}, T_{1C}, T_{2D}, T_{h1E}] \tag{34}$$

## V. Discussion and Conclusions

This HEN case study shows clearly that the control structure design has a strong impact on the performance. Especially in cryogenic processes and systems which process large quantities, like refineries, improving the average end temperature just 0.5-1°C leads to significant economics savings.

Although the disturbances in Scenario 1 are included in Scenario 2, we treated Scenario 1 separately, because it shows that the stream inlet temperatures are the disturbances, which have the strongest effect on the loss. If it could be guaranteed that the inlet temperatures remain constant (Scenario 1), then keeping the splits constant would most likely be sufficient.

The situation changes dramatically if the stream inlet temperatures are varying (Scenario 2). Here an open loop policy causes high loss, and controlling single measurements performs even worse. In this case, however, controlling good temperature measurement combinations can reduce the loss up to around 2 orders of magnitude.

The relationship between the number of measurements used and the loss, Fig. 2 and 3, shows that including flow measurements does not reduce the loss significantly, provided enough temperature measurements are available. Controlling combinations of about 10 or more temperature measurements results in a very small loss.

The null space method, which gives zero loss without noise, has been shown to give very poor performance, sometimes even worse than open loop operation, because the loss is dependent on which set of null space basis vectors is chosen as $H$. Thus, this case study clearly demonstrates how necessary it is to take noise into account when finding $H$.

## References

[1] N. Aguilera and J. L. Marchetti, “Optimizing and controlling the operation of heat-exchanger networks,” *AIChE Journal*, vol. 44, no. 5, pp. 1090–1104, 1998.

[2] V. Lersbamrungsuk, T. Srinophakun, S. Narasimhan, and S. Skogestad, “Control structure design for optimal operation of heat exchanger networks,” *AIChE Journal*, vol. 54, no. 1, pp. 150–162, 2008.

[3] T. Lid, S. Strand, and S. Skogestad, “On-line optimization of a crude unit heat exchanger network,” in *Proceedings of the 6th Conference on Chemical Process Control (CPC VI), Tucson Arizona, AIChE Symposia Series No. 326*, January 2001, pp. 476 – 480.

[4] S. Skogestad, “Plantwide control: The search for the self-optimizing control structure,” *Journal of Process Control*, vol. 10, pp. 487–507, 2000.

[5] L. M. Umar, W. Hu, Y. Cao, and V. Kariwala, “Selection of controlled variables using self-optimizing control method,” in *Plantwide Control:Recent Developments and Applications*, G. P. Rangaiah and V. Kariwala, Eds. John Wiley & Sons, 2012.

[6] V. Alstad, S. Skogestad, and E. S. Hori, “Optimal measurement combinations as controlled variables,” *Journal of Process Control*, vol. 19, no. 1, pp. 138–148, 2009.

[7] V. Kariwala, Y. Cao, and S. Janardhanan, “Local self-optimizing control with average loss minimization,” *Industrial & Engineering Chemistry Research*, vol. 47, pp. 1150–1158, 2008.

[8] I. J. Halvorsen, S. Skogestad, J. C. Morud, and V. Alstad, “Optimal selection of controlled variables,” *Industrial & Engineering Chemistry Research*, vol. 42, no. 14, pp. 3273–3284, 2003.

[9] H. Pirnay, R. López-Negrete, and L. T. Biegler, “Optimal sensitivity based on ipopt,” *Submitted to submitted to Math. Prog. Comp.*, 2011.

[10] A. Wächter and L. T. Biegler, “On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming,” *Mathematical Programming*, vol. 106, no. 1, pp. 25–57, 2006.

[11] Y. Cao and V. Kariwala, “Bidirectional branch and bound for controlled variable selection: Part i. principles and minimum singular value criterion,” *Computers & Chemical Engineering*, vol. 32, no. 10, pp. 2306 – 2319, 2008.

[12] V. Kariwala and Y. Cao, “Bidirectional branch and bound for controlled variable selection. part ii: Exact local method for self-optimizing control,” *Computers & Chemical Engineering*, vol. 33, no. 8, pp. 1402 – 1412, 2009.

[13] ——, “Bidirectional branch and bound for controlled variable selection part iii: Local average loss minimization,” *Industrial Informatics, IEEE Transactions on*, vol. 6, no. 1, pp. 54 –61, 2010.

[14] R. Yelchuru and S. Skogestad, “Convex formulations for optimal selection of controlled variables and measurements using mixed integer quadratic programming,” *Submitted to Journal of Process control*, 2012.

[15] L. T. Biegler, *Nonlinear Programming: Concepts, Algorithms, and Applications to Chemical Processes*. SIAM - MOS, 2010.

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# Plantwide Control of A Benchmark Bleach Plant

Fernando Suárez Antelo
School of Chemical Engineering and Advanced Materials
Newcastle University
Newcastle upon Tyne NE1 7RU

Jie Zhang
School of Chemical Engineering and Advanced Materials
Newcastle University
Newcastle upon Tyne NE1 7RU
e-mail: jie.zhang@newcastle.ac.uk

*Abstract*— **This paper presents a study on the plantwide control of a benchmark pulping process. Control structure determination was based on Relative Gain Array (RGA) and Relative Disturbance Gain (RGD) analysis. RGA and RDG were calculated based process a multiple-input multiple-output (MIMO) process transfer model identified from open-loop tests. Both static and dynamic RGA and RDG were calculated and analysed. These analyses confirmed the setting provided by the authors of the Benchmark, based on static RGA. Controllers in the individual control loops were tuned using Internal Model Control (IMC) tuning methodologies and the best set of controller parameters was chosen by evaluating the control system performance for set-point tracking and disturbance rejection in terms of the Integral of Absolute Error (IAE), setting time, time constant and percentage of overshoot. The Kappa factor, a key quality variable, was controlled by PI controllers combined with Smith-predictors and tuned with IMC. PI-only controllers tuned using IMC were used to control the secondary variables. With Smith predictors larger controller gains and smaller reset times can be used leading to faster control response. The proposed control strategy is successfully implemented on the benchmark simulation.**

***Keywords-component; formatting; style; styling; insert (key words)***

## I. INTRODUCTION

In a pulp mill, the main objective is to produce pulp of a certain Kappa number (measurement of lignin content) and brightness (traditional measurement of blue reflectance) using the minimum energy resources, utilities and chemicals. Castro and Doyle [1] presented a detailed control study of the fiberline of a pulp mill process. The process model consists of a set of equations with approximately 5000 states to capture the dynamics of the main unit operations: pulp digester, oxygen reactor, bleach towers, washers, and storage tanks. Heuristic methods were used to determine the primary control variables and relative gain array (RGA) analysis was performed to obtain the input-output pairings for decentralized control. The performance of model predictive control and decentralized single-input single-output control were compared, finding that the MPC offers a better framework when controlling the digester but no big differences between both techniques were found when controlling the bleach plant.

Vanbrugghe et al. [2] recalculated the RGA of the bleach plant and proposed a real-time optimization algorithm based on a modified version of an IMC-based optimization method. The performance was improved by on-line estimation of the process parameters and the total cost of the bleaching section was reduced by 10.6%. In 2004, Castro and Doyle [3] introduced a benchmark problem of a pulping process, including both the fiber line and the chemical recovery sections. The complete details of the pulp mill process were given, as well as the control objectives, modes of operation, process constraints, measurements and costs. The dynamic model, including the source/binary code of all the unit operations was made available to the process control academic community as a benchmark for its use in process system engineering studies. The benchmark also provides code for different controller structures, such as PID and MPC, including other decentralized advanced tools like feedforward controllers and Smith predictors.

Since its introduction, RGA has been a very important tool for determining the best input-output parings for decentralized control. However, there are many control practitioners who doubt about its usefulness in some control applications because it does not include the effects of disturbances. Stanley et al. [4] proposed a new measurement, called relative disturbance gain (RDG) in order to include disturbances when selecting the input-output pairings. There is not an explicit study of RGD in the bleach plant supporting and confirming the proposed input-output pairings obtained using RGA or dynamic RGA. It should be therefore necessary to demonstrate that the input-output pairings given by Castro and Doyle [1] are suitable in the bleach plant in terms of disturbance rejection performance.

The paper is organized as follows. Section II presents the benchmark simulated bleach plant. Control structure selection using RGA and RDG analysis is presented in Section III. Section IV gives the control system performance. Some concluding remarks are presented in Section IV.

## II. THE BLEACH PLANT

The main objective of the bleach plant is to remove lignin from the pulp and to obtain an appropriate brightness coefficient. This objective is achieved by using bleach towers, where the pulp is mixed together with oxidizing chemicals

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

which make the lignin to be soluble in water. Fig. 1 shows a schematic of the bleaching plant [3].

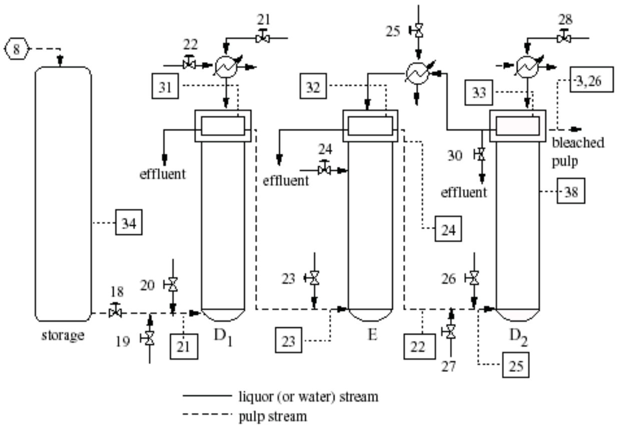


Fig. 1. Bleach plant flowsheet

After storage, the pulp is subjected to three bleach stages. The bleach towers are vertical cylindrical vessels. The pulp moves vertically in plug flow. Each tower has auxiliary equipment such as a chemical mixer, a washer and a seal tank.

The washer is mounted in the bleach tower exit and it is used to eliminate dissolved chemicals from the pulp. It is a rotary drum washer, where clear water or recycled washer effluent coming from the following bleach tower is used to eliminate dissolved chemicals. The pulp enters in the rotating drum under vacuum. The outer surface of the drum is at higher pressure than its internal part, so the pulp enters in the drum through its porous surface. This causes the formation of a mat of pulp in the surface of the drum. Then, wash showers are used to remove further dissolved solids which may still in the pulp. Additional water is mixed together with the pulp to achieve the desired consistency. The temperature of this water is controlled with heat exchangers. The seal tank acts as storage or buffer for compensating variations in the composition of the liquor. The first and third bleach towers use chlorine dioxide as chemical agent to remove lignin from the pulp while the second uses sodium hydroxide.

The benchmark bleach plant has 11 controlled variables (CV), 14 possible manipulated variables (MV) and 10 potential disturbances (DV). Tables I, II and III show, respectively, the controlled, manipulated and disturbance variables.

TABLE I.
CONTROLLED VARIABLES OF THE BLEACH PLANT

| CV | Description | CV | Description |
|---|---|---|---|
| 21 | Temperature of bleach tower $D_1$ | 31 | Washer 5 dilution factor |
| 22 | Bleach tower E Kappa no. | 32 | Washer 6 dilution factor |
| 23 | Temperature of bleach tower E | 33 | Washer 7 dilution factor |
| 24 | E washer $[OH^-]$ | 34 | Storage volume |
| 25 | Temperature of bleach tower $D_2$ | 38 | $D_2$ tower volume |
| 26 | $D_2$ tower brightness | | |

Controlled variables CV22 and CV26 are considered to be the quality variables of the bleach plant, so special care must be taken when designing a control system for these variables. Variable CV24 is not a quality variable, but it has a big influence in variable CV26. These three variables are assumed to be the primary controlled variables of the bleach plant. The rest of the controlled variables are the secondary controlled variables of the process.

TABLE II.
DISTURBANCE VARIABLES OF THE BLEACH PLANT

| DV | Description | DV | Description |
|---|---|---|---|
| 13 | $D_1$ $ClO_2$ stream temperature | 18 | $D_2$ $ClO_2$ stream temperature |
| 14 | $D_1$ $ClO_2$ stream composition | 19 | $D_2$ $ClO_2$ stream composition |
| 15 | E caustic temperature | 20 | $D_2$ caustic temperature |
| 16 | E caustic composition | 21 | $D_2$ caustic composition |
| 17 | E back-flush stream temperature | 22 | Wash washer temperature |

TABLE III.
MANIPULATED VARIABLES OF THE BLEACH PLANT

| MV | Description | MV | Description |
|---|---|---|---|
| 17 | O steam flow 3 | 24 | E back-flush flow |
| 18 | Storage exit flow | 25 | E steam flow |
| 19 | $D_1$ water flow | 26 | $D_2$ $ClO_2$ flow |
| 20 | $D_1$ $ClO_2$ flow | 27 | $D_2$ caustic flow |
| 21 | $D_1$ wash water flow | 28 | $D_2$ wash water flow |
| 22 | $D_1$ steam flow | 30 | Split fraction 4 |
| 23 | E caustic flow | 38 | $D_2$ exit flow |

## III. CONTROL STRUCTURE SELECTION

The RGA technique was developed by Bristol [5] and has become the most important technique for measuring interaction and a very useful tool for decentralized control design. It is a valuable technique for the selection of manipulated-controlled variable pairings and it can also be used to predict the behaviour of controlled responses [6]. Grosdidier et al. [7] provided a derivation of the properties of the RGA. Additional properties were presented by Hovd and Skogestad [8], who extend the rules to the frequency domain.

The RGA methodology requires the steady-state gains of the process to determine the best set of input-output pairings. The presence of the storage tank and the $D_2$ bleach tower makes the process to be open-loop unstable. This is caused by the integrating nature of level systems. Before proceeding with the open-loop tests, two loops were closed to control the level of these two vessels. The manipulated variable selected to control the level of the storage tank was the storage exit flow. The manipulated variable chosen to control the level of the $D_2$ bleach tower was the $D_2$ exit flow. In order to avoid excessive variations of these two manipulated variables, two proportional-only controllers with default settings were used to perform this task. After removing these two input-output variables from the steady-state gain matrix, twelve candidate manipulated variables were available to control the remaining nine output variables. The generalization of the RGA for non-square plants was used to perform the RGA analysis. Then, the RGA, $\Lambda$, is given by:

$$\Lambda = H * \left(H^T\right)^{-1} \tag{1}$$

where * represents element by element multiplication.

The gain matrix $G$ can be decomposed, using singular value decomposition, as:

$$G = UDV^T \tag{2}$$

where $U$ and $V$ are orthogonal matrices and $D$ is a diagonal matrix containing only the positive singular values. In Eq(1), $H$ is the pseudo-inverse of matrix $G$. It was observed that the sum of terms of the columns related to manipulated variables “MV24”, “MV19” and “MV19” were much lower than one, so these variables were removed from the RGA to obtain a square matrix. The resulted RGA is presented in Fig. 2. Due to the sequential nature of the bleach plant, the RGA is almost a diagonal matrix. All diagonal terms are close to one and off-diagonal terms are negligible.

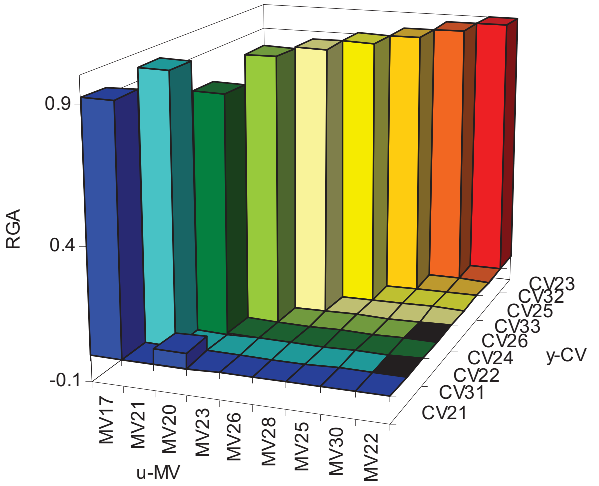


Fig. 2. Relative Gain Array of the bleach plant

After approximating the open-loop responses with continuous first-order-plus-time-delay transfer functions, the RGA was calculated as a function of frequency to find if the pairings calculated using static RGA are still appropriate from a dynamic point of view. All diagonal terms of the RGA remain closed to one in the whole frequency range, which indicates the proposed set of input-output pairings is also suitable from the dynamic point of view. Fig. 3 shows the first three diagonal elements of the dynamic RGA.

Since its introduction, RGA has been widely used in industry to configure multi-loop control systems [9]. McAvoy [10] has shown that there is a strong link between the RGA and the stability and design of a control loop. However, the RGA has been the subject of controversy [11]. Some engineers think that it does not have any use while others strongly rely on it.

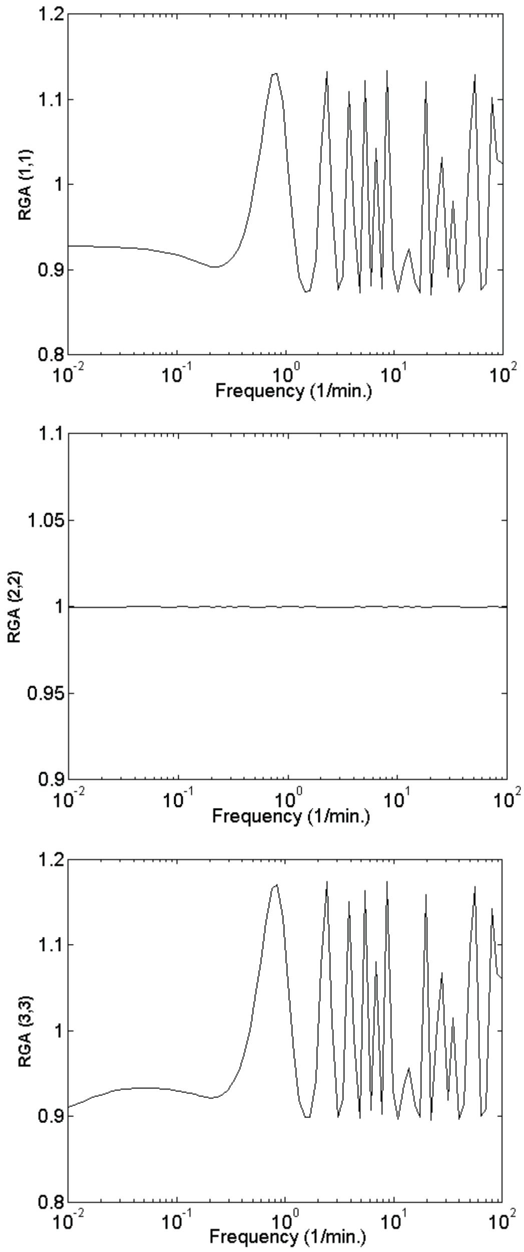


Fig. 3. The first three diagonal terms of the dynamic RGA

RGA is generally considered having two main disadvantages. Firstly, RGA is calculated from steady-state data only and therefore dynamic interactions may cause it to provide wrong conclusions. This problem was overcome by introducing dynamic interaction measures [12], [13]. Secondly, RGA is independent of load disturbances affecting the control loop. Stanley et al. [4] defined RDG in order to include disturbances in the analysis of the control loop performance. RGD is similar to RGA in the sense that it involves the ratio of two different steady-state gains: perfect control gains to open-loop gains. RDG can be calculated from steady-state information only but it can also be extended to take into account dynamic interactions. For simplicity, the definition of the RDG is presented for a 2×2 system. The extension for an $n \times n$ case is straightforward. The steady-state gain matrix for a 2×2 system including the gains of a load disturbance, $d$, is:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} K_{11} & K_{12} & K_{F1} \\ K_{21} & K_{22} & K_{F2} \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ d \end{bmatrix} \qquad (3)$$

The RGD, $\beta_1$, is defined as the ratio of changes in the controller output that is required to bring $x_1$ back to its desired set-point when the load disturbance, $d$, is introduced into the system under two situations: multi-loop control and single loop control. Mathematically, $\beta_1$ is defined as a ratio of two gains:

$$\beta_1 = \frac{\left.\dfrac{\partial m_1}{\partial d}\right|_{x_1, x_2}}{\left.\dfrac{\partial m_1}{\partial d}\right|_{x_1, m_2}} \qquad (4)$$

Thus, $\beta_i$ can also be interpreted as a comparison between multi-loop control and ideal decoupled control. This means that if $\beta_i > 1$, then the controller effort within a multi-loop environment is bigger than the one required for a SISO system and therefore a decoupler is recommended. A small value of the RDG means that the controller output does not have to move too far from its steady-state to compensate the effects of the load disturbance. Mathematically, multi-loop control is preferred when [4]:

$$|\beta_1| + |\beta_2| < 2 \qquad (5)$$

From the ten possible load disturbances, it was observed that just five play a significant role in the process. These load disturbances are DV14, DV16, DV18, DV19 and DV22. Each of the load disturbances DV18, DV19 and DV22 only affects one controlled variable CV25, CV26 and CV25 respectively. The load disturbance DV14 affects the controlled variables CV22 and CV26 while the load disturbance DV16 affects the controlled variables CV24 and CV26. Therefore, the RDG analysis only makes sense when studying the disturbance variables DV14 and DV16. Fig. 4 shows the effects of the disturbance variable DV16 on controlled variables CV24 and CV26 in terms of RDG. Fig. 5 gives the RDG terms related to disturbance DV14. As observed, the sum of RDG does not surpass considerably the upper limit in the whole frequency range. This means the control loop pairings obtained using RGA are appropriate for disturbance rejection is this process.

## IV. Control Performance

The primary variables were controlled by a combination of Kappa Factor control with conventional feedback control while the secondary outputs using only PI controllers. The Kappa Factor control is a particular controller used in Pulp Mills especially designed for disturbance rejection. Its operation together with PI controllers allows free-offset set-point tracking. Essentially, the Kappa factor can be understood as the ratio between lignin content in the pulp and the chemical agent entering the bleach tower. The general equations of the Kappa Factor are defined as [2]:

$$K_f = a_n K_{in}^n + a_{n-1} K_{in}^{n-1} + \ldots + a_1 K_{in} + a_0 \qquad (6)$$

$$F_x = K_f F_p C_p K_{in} / X \qquad (7)$$

where $K_f$, $F_x$, $X$, $F_p$, $C_p$ and $K_{in}$ are, respectively, the Kappa factor, chemical flow rate (manipulated variable), chemical composition, pulp flow rate, pulp consistency and pulp upstream Kappa no. The product of variables $K_f$, $C_p$ and $F_P$ determines the amount of lignin entering the bleach tower. Coefficients $a_0$ to $a_n$ define a polynomial of order $n$ which expresses the functionality between the Kappa Factor and the incoming Kappa no. Based on industrial experience it is possible to determine this functionality.

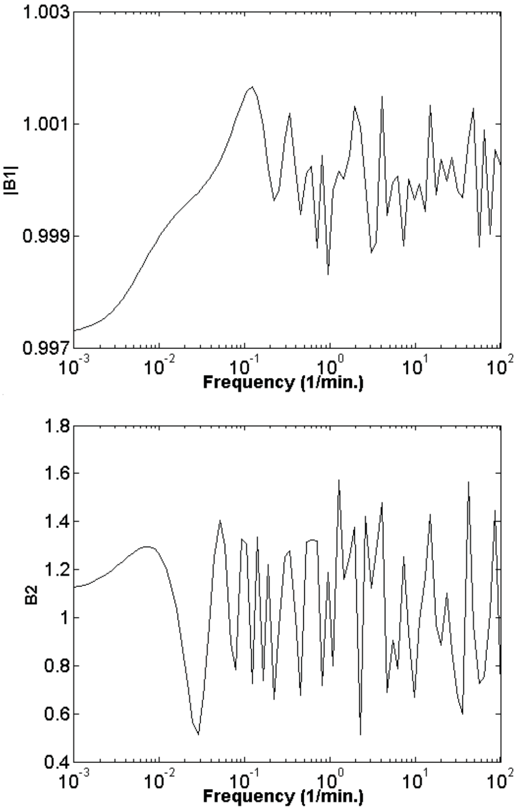


Fig. 4. Dynamic RDG for disturbance variable DV16 and controlled variables CV24 and CV26

Controller settings for each control loop were calculated using different internal model control tuning (IMC) methodologies. Rivera et al [14] have proposed tuning methods for PID controllers based on the IMC methodology. Table 4 gives the IMC tuning rules for the following first order plus delay model:

$$G(s) = \frac{Ke^{-\alpha s}}{1+\tau s} \qquad (8)$$

TABLE IV.
IMC TUNING FORMULAE

| Controller Type | $K_c$ | $\tau_I$ | Recommended λ |
|---|---|---|---|
| PI | $\tau/(\lambda K)$ | $\tau$ | λ/α>1.7 |
| Improved PI | $(2\tau+\alpha)/(2\lambda K)$ | $\tau+\alpha/2$ | λ/α>1.7 |

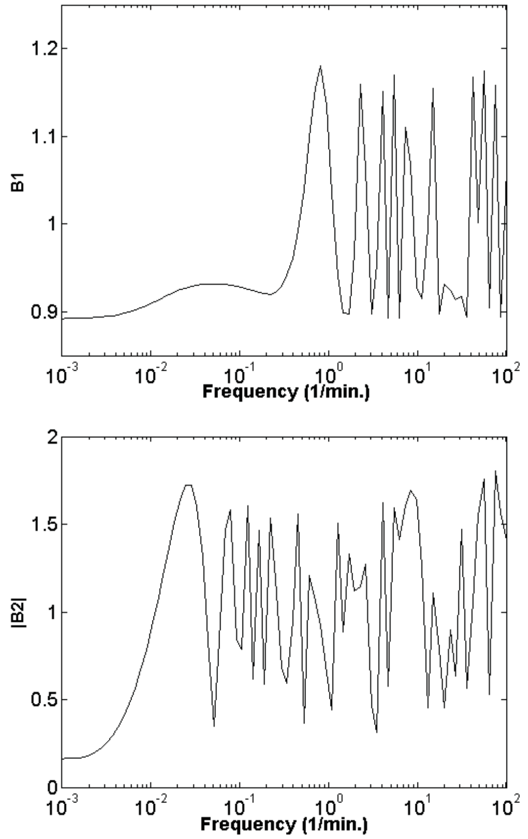
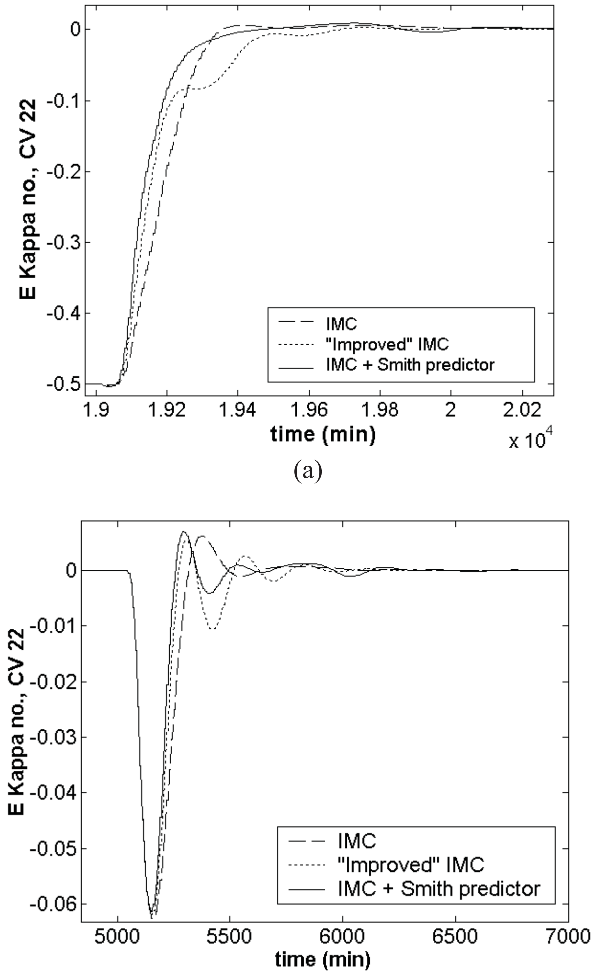


Fig. 5. Dynamic RDG for disturbance variable DV14 and controlled variables CV22 and CV26

A conservative value of λ=2.5α was used in order to avoid excessive control action due to model-plant mismatch. For variables with no time delay, the value of λ was adjusted by inspection of the controlled responses. Additionally, controller settings were calculated using Smith-Predictors in the quality variables. Using Smith-Predictors, instead of tuning each controller without considering the time delay of the process, a value of λ=2α was introduced in the classical IMC formulae. The best setting of controller parameters was chosen by the behaviour of the system for set-point tracking and disturbance rejection performance. Integral of absolute error (IAE), settling time, time constant and percentage of overshoot were the parameters used for evaluating these performances. PI controllers combined with Smith-predictors, and tuned with classical IMC, providing the set-points of the Kappa factor controllers related to the quality variables of the process and PI-only controllers tuned using IMC controlling the secondary variables give the best control performance. Smith predictors allow the controller designer to provide to the process controllers larger controller gains and smaller reset times, making the controlled response faster.

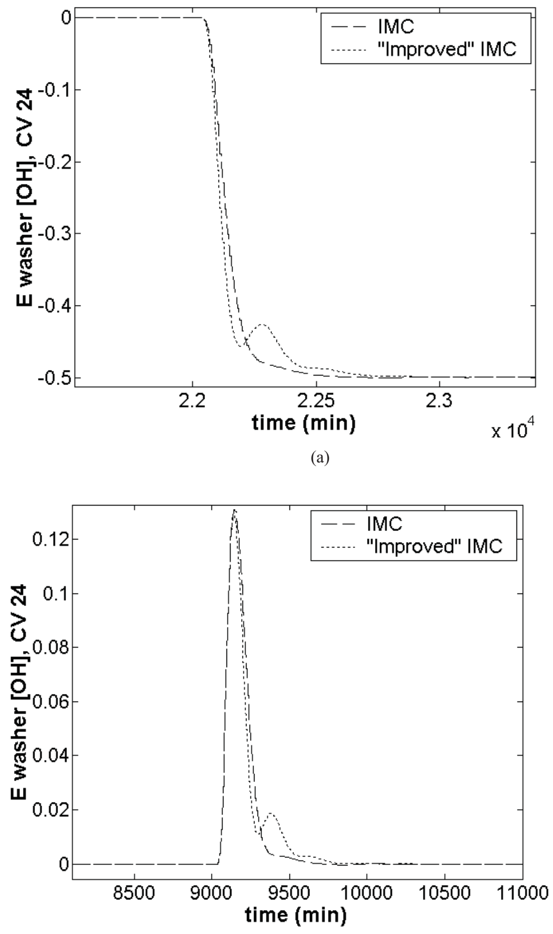


(a)

(b)

Fig. 6. Set-point tracking and disturbance rejection performances for CV22

Fig. 6 (a) and Fig. 7 (a) show the set-point tracking performances of the primary variables CV22 and CV24. Examples of disturbance rejection performances for the three primary variables are shown in Fig. 6 (b), Fig. 7 (b) and Fig. 8. Fig. 6 (b) represents the disturbance rejection performance of the primary variable CV22 under a positive step change of magnitude 0.25 in the disturbance variable DV14. By the same way, Fig. 7 (b) and Fig. 8 represent, respectively, the disturbance rejection performance of the primary variables CV24 and CV26 under a positive step change of magnitudes 0.25 in respectively the disturbance variables DV16 and DV19. As it can be observed, the use of Smith predictors significantly improves the control performance.



(a)

(b)

Fig. 7. Set-point tracking and disturbance rejection performances for CV24

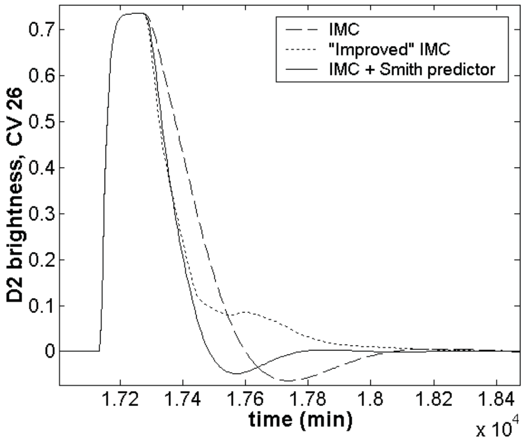


Fig. 8. Disturbance rejection performance for CV26

## V. Conclusions

Plant wide control of simulated benchmark bleach plant with emphasis on Kappa number control and brightness control is presented in this study. Dynamic RDG analysis confirms the input-output setting provided by the authors of the benchmark based on RGA analysis. Additionally, it is demonstrated that PI controllers combined with Smith-predictors, and tuned with classical IMC, providing the set-points of the Kappa factor controllers related to the quality variables of the process offers an appropriate methodology to control the main objectives of the plant in terms of set-point tracking and disturbance rejection performances.

## References

[1] J. J. Castro and F. J. Doyle III, "Plantwide control of the fiberline in a pulp mill", *Ind. Eng. Chem. Res.*, vol. 41, pp. 1310-1320, 2002.

[2] C. Vanbrugghe, M. Perrier, A. Desbiens, and P. Stuart, "Real time optimization of a bleach plant using an IMC-based optimization algorithm", *Pulp and Paper Canada*, 2004

[3] J. J. Castro and F. J. Doyle III, "A pulp mill benchmark problem for control: problem description", *Journal of Process Control*, vol. 14, pp. 17-29, 2004.

[4] G. Stanley, M. Marino-Galarraga, and T.J. McAvoy, "Shortcut operability analysis. 1. The relative disturbance gain", *Ind. Eng. Chem. Process Des. Dev.*, vol. 24, pp. 1181-1188, 1985.

[5] E. H. Bristol, "On a measure of interaction in multivariable process control", *IEEE Trans. on Auto. Control*, vol. AC-11, pp. 133-134, 1966.

[6] F. G. Shinskey, "Predict distillation column response using relative gains", *Hydrocarbon Processing*, May 1977.

[7] P. Grosdidier, M. Morari, and B. Holt, "Closed-loop properties for steady-state gain information", *Ind. Eng. Chem. Process Des. Dev.*, vol. 24, pp. 221-235, 1985.

[8] M. Hovd and S. Skogestad, "Simple frequency-dependent tools for control system analysis, structure selection and design", *Automatica*, vol. 25, no. 5, pp. 989-996, 1992.

[9] C. Thurston, Hydrocarbon Processing, vol. 60, pp. 125, 1981.

[10] T. J. McAvoy, AIChE J., "Connection between relative gain and control loop stability and design", vol. 27, pp. 613-619, 1981.

[11] N. Jensen, "Commons on some results on dynamic interaction analysis of complex control systems", *Ind. Eng. Chem. Process Des. Dev.*, vol. 24, pp. 228-229, 1985.

[12] E. Bristol, Paper presented at the AIChE 71st Annual Meeting, Miami, FL, Nov 1978.

[13] M. Witcher and T. J. McAvoy, *ISA Trans.* , vol. 16, pp. 35, 1977.

[14] D. E. Rivera, M. Morari, and S. Skogestad, "Internal model control – 4. PID controller design", *Ind. Eng. Chem. Process Des. Dev.*, vol.25, pp. 252-265, 1986.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Eco-efficiency of control configurations using Exergy

M. T. Munir is working as a Post doc in the Chemical and Materials Engineering Department, The University of Auckland New Zealand (email: mmun047@aucklanduni.ac.nz).

W. Yu is a lecturer at the Chemical and materials department, The University of Auckland New Zealand (email:wyu048@aucklanduni.ac.nz).

B. R. Young is a professor at the Chemical and materials engineering department, The University of Auckland New Zealand (Phone: +64 9 923 5606; Fax: +64 9 373 7463; email: b.young@auckland.ac.nz).

***Abstract*— For general multi-input multi-output (MIMO) processes, thermodynamic properties like exergy have been previously used for the development of eco-efficiency analysis tools e.g. Relative Exergy Array (REA) [1, 2], and Exergy Eco-efficiency Factor (EEF) [3]. The REA is easy to use, compares several control scheme candidates for single units and ranks them according to their exergy interactions between control loops. The EEF provides means to determine the effect of a control scheme on the eco-efficiency of the whole process. In this paper, our intention is to justify the use of EEF and REA, compare EEF recommendations to REA, and consider EEF for the whole process eco-efficiency and discuss the similarities and differences between REA and EEF. Furthermore these results of testing REA and EEF on a realistic case study will be analysed to provide some practical recommendations on their usage.**

**Keywords- Relative exergy array (REA); Control configurations; Exergy eco-efficiency factor (EEF); Eco-efficiency**

## I. Introduction

Most industrial chemical processes are multi-input multi-output (MIMO) in nature. These systems can be either controlled by a multivariable or centralized MIMO controller or by a set of single-input single-output (SISO) controllers. Since centralized multivariable controllers are complex and lack integrity, decentralized control systems have more attractive advantages: i) simple algorithms; ii) ease of understanding by plant operating personnel; and iii) standard control design developed for common unit operations [4]. Therefore, they are more often selected for control of MIMO systems.

For a decentralized type of control system, the control system structure decides the best control scheme. Control scheme selection pertains to the pairing of manipulated (MV) and controlled variables (CV). For a process or plant, control scheme selection is a straightforward task, provided that no interactions are present between the various control loops in multi-loop control schemes of that process or plant. However, this is rarely the case in process control design practice. A well performing control scheme selection is essential because the incorrect pairing of MV and CV will result in poor performance.

There are several techniques and methods for designing or selecting decentralized MIMO control schemes, such as the relative gain array (RGA), the Niederlinski index (NI), singular value decomposition (SVD), the condition number (CN) and Morari's resiliency index (MRI) [5-7].

In some cases, these steady state information based techniques (RGA, NI, SVD and CN) can lead to incorrect conclusions concerning the control scheme selection. Dynamic effects are also included in some techniques/methods to minimize the deficiency of these steady state techniques. For example, this happens in dynamic relative gain array (DRGA), internal model control (IMC) and effective relative gain array (ERGA). Using the steady state (RGA, NI, SVD and CN) and dynamic (DRGA, IMC and ERGA) techniques, well performing control schemes are further selected. The viability of a well performing/selected control scheme is further validated by the dynamic simulation of responses to the various process disturbances.

Steady state and dynamic techniques for selecting MIMO control schemes focus on controllability and control loop stability. Control system structure or control scheme selection is a large part of process control. In this age of ubiquitous industrialization and in the wake of decreasing energy resources, increasing energy costs and energy crises, special attention must be paid to the control system structure or control scheme selection, as a poorly structured control scheme can lose a great deal of energy from a process or plant. In the wake of energy crises, control scheme selection must also consider energy usage, cost and ecological impacts. This can be achieved by integrating control scheme selection and energy cost/energy usage/environmental impacts.

Thermodynamic properties such as exergy have the potential to amalgamate several elements into a single domain, these include; control system structure/control scheme selection and energy cost/energy usage/environmental impacts. The concept of exergy indicates what is wasted in terms of energy or the eco-efficiency of the process/plant. Exergy is the component of energy which is available for useful work. Kotas [8, pp 7] defines exergy as "The maximum amount of work/useful energy drawn from a process/material stream as it comes from its original state (process condition) to the ultimate dead state (reference state) during which it interacts only with the environment". At the ultimate dead state, the

Corresponding author: b.young@auckland.ac.nz process/material stream is in thermodynamic (thermal, mechanical and chemical) equilibrium with the environment [9].

The total energy of a material stream has two main parts (available energy and unavailable energy). Exergy of a material stream accounts for the quality of energy or available energy. According to the 1st law of thermodynamics energy in and out of a process is equal. According to the 2nd law of thermodynamics exergy in is greater than exergy out due to entropy generation or exergy destruction during the process. The fraction of non-available energy normally increases from input to output due to exergy destruction during a process. The increase in non-available energy is also called available energy loss as shown in Figure 2.

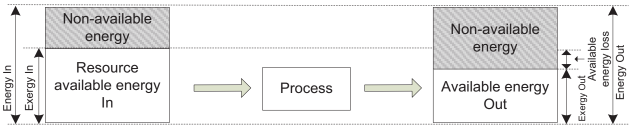


Figure 1. Energy and Exergy

Exergy accounts for the quality of energy/fraction of energy that can be fully converted into useful work and/or other types of energies. Every irreversible process causes exergy destruction leading to an increased exergy/energy requirement or entropy production. Since exergy can detect and evaluate the causes of the thermodynamic imperfections of the considered process/material stream, it can therefore be used as a measure to evaluate the eco-efficiency of the process.

Eco-efficiency is also a measure of progress in green chemical engineering growth. The concept of eco-efficiency can be traced back to the 1960s as the concept of environmental efficiency, or as a business when linked to sustainable development. Eco-efficiency has a role in expressing how efficient economic activity is with regard to nature's goods/services. The concept of eco-efficiency focuses on methods of resource use, obtaining economic and environmental progress through efficient use of resources and lower pollution/emissions.

The World Business Council for Sustainable Development (WBCSD) states that eco-efficiency is achieved through the practice of producing "valuable goods and services that satisfy human needs and bring quality of life with reduced environmental impacts and resource intensity in line with the Earth's estimated carrying capacity." In other words eco-efficiency means producing more with less. An eco-efficient process is ecologically friendly and economically viable. Ecologically friendly practices signify a process with reduced consumption of energy/destroyed exergy. This diminution in energy utilization reduces the operational expenses of that process. Efficient use of energy/exergy plays an important role for sustainability and minimizing environmental impacts.

Exergy has many valuable uses in process design and control to facilitate these complex tasks. There are many exergy based tools and methods, such as; efficiency concepts, exergy flow diagrams, relative exergy array (REA) and exergy eco-efficiency factor (EEF). The REA is used to measure control scheme interactions and exegetic efficiency for the various possible MIMO control schemes [1, 2, 10]. However, the REA only measures the eco-efficiency within the scope of single, decentralized control loops. To overcome this deficiency in REA, the EEF is defined and employed for the eco-efficiency analysis of the whole process [3]. More details of REA and EEF are presented in section II.

In this work, the intention is to justify the use of REA and EEF coupled with classical controllability tools (RGA, NI, SVD and CN), compare EEF recommendations to REA, and consider EEF for the whole process eco-efficiency and discuss the similarities and differences between REA and EEF. The results of REA and EEF are also analyzed to provide some practical recommendations on their usage.

This article is further organized as follows. After this general and brief introduction related to this research, more details of REA and EEF are discussed and explained. Following this, the results of REA and EEF are explained with a realistic case study. Finally, the results are discussed and conclusions drawn.

## II. RELATIVE EXERGY ARRAY (REA) AND EXERGY ECO-EFFICIENCY FACTOR (EEF)

Relative gain is defined as "The ratio of open loop process gain in an isolated loop to apparent process gain in the same loop when all other control loops are closed and are in perfect control", [11]. RGA is a matrix composed of elements defined as the ratio of open-loop to closed-loop gains. One of its elements, relative gain, $\lambda_{ij}$, which relates the $j^{th}$ input $u_j$ and the $i^{th}$ output $y_i$, can be expressed by the following Equation (1).

$$\lambda_{ij} = \frac{\left(\frac{\Delta y_i}{\Delta u_j}\right)_{all-loops-open}}{\left(\frac{\Delta y_i}{\Delta u_j}\right)_{all-loops-closed(in-perfect-control)-except-u_j-loop}} \qquad (1)$$

### A. Relative exergy array (REA)

Exergy helps to develop many tools for the facilitation of process design and control, such as exergetic efficiency, REA and EEF.

For the understanding of Relative Exergy Array (REA), the concept of exergetic efficiency is introduced. Exergetic efficiency is defined as the ratio of the exergy out to the exergy in to a process as shown in Equation (2). A general process for exergetic efficiency calculation is shown in Figure 2.

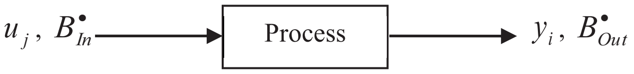


Figure 2. A general control loop portion

$$\eta = \frac{B^{\bullet}_{Out}}{B^{\bullet}_{In}} \tag{2}$$

where $\eta$ = Exergetic efficiency, $B^{\bullet}_{Out}$ = Total exergy going out of a process and $B^{\bullet}_{In}$ = Total exergy coming in to a process.

In a control loop a manipulated variable (MV) is paired with a controlled variable (CV) and both are linked with each other by some process as shown in Figure 2. In a control loop, the exergy of the manipulated variable ($u_j$) stream is the total exergy coming in to the process ($B^{\bullet}_{In}$) and the exergy of the control variable ($y_i$) stream is the total exergy going out of a process ($B^{\bullet}_{Out}$) as shown in Figure 2. The ratio of the exergy of the control variable stream to the exergy of the manipulated variable stream gives the exergetic efficiency.

The Initial exergetic efficiency is based on the exergy of the manipulated variable stream and the exergy of the control variable stream before a step input in the exergy of the manipulated variable stream. The exergy gain ratio is calculated after a step change in the exergy of the manipulated variable stream. The exergy gain ratio defined by Equation (3) gives the exergetic efficiency of the process.

$$\tau_{ij} = \frac{\Delta B(y_i)}{\Delta B(u_j)} = \frac{B^{\bullet}_{Out,Final} - B^{\bullet}_{Out,Initial}}{B^{\bullet}_{In,Final} - B^{\bullet}_{In,Initial}} \tag{3}$$

where $\tau_{ij}$ = Exergy gain ratio

Equation 3 is also called the generic exergy gain ratio which allows an alternate measurement of exergetic efficiency, defined by Equation 2. The open loop exergy gain of a control loop can be calculated via Equation 3 when all other loops are open. The open loop exergy gain is analogous to the open loop gain in the Relative Gain Array (RGA) as shown in Equation 4.

$$k_{ij} = \frac{\Delta y_i}{\Delta u_j} = \frac{y_{i,Final} - y_{i,Initial}}{u_{j,Final} - u_{j,Initial}} \tag{4}$$

where $k_{ij}$ = simple gain ratio

In the open loop gain ratio calculation, all loops are open and no interactions from other loops affect the considered loop ($u_j$ - $y_i$). When all other loops except ($u_j$ - $y_i$) are closed then loop interactions from other loops affect the considered loop ($u_j$ - $y_i$). These loop interactions change the effect of manipulated variable stream step input to the controlled variable stream of control loop ($u_j$ - $y_i$) in terms of control variable $y_i$ and its exergy. So a great variation in open loop gain or open loop exergy gain to closed loop gain or closed loop exergy gain, accounts for large interactions within the control loops. This translates to an alternate measure of loop interactions by using exergy values. With loop interaction measurement it also accounts for exergetic efficiency of a control loop ($u_j$ - $y_i$) effected by interaction from other loops.

So from the above discussion relative exergy gain is defined by Equation 5, (analogous to relative gain in RGA as shown in Equation 1).

$$\gamma_{ij} = \frac{\left(\dfrac{\Delta B_{total}(y_i)}{\Delta B_{total}(u_j)}\right)_{\text{all loops open}}}{\left(\dfrac{\Delta B_{total}(y_i)}{\Delta B_{total}(u_j)}\right)_{\text{all loops closed (in perfect control) except } u_j \text{ loop}}} \tag{5}$$

where $\gamma_{ij}$ = Relative exergy gain ratio

For all pairing combinations in multi-loop SISO control, relative gains ($\lambda_{ij}$) are calculated. These relative gains are arranged into an array called RGA ($\Lambda$) as shown by Equation 6. Similarly for all pairing combinations in multi-loop SISO control, relative exergy gains ($\gamma_{ij}$) are calculated. These relative exergy gains are arranged into an array called REA ($\Gamma$) as shown by Equation 7. The interpretations of RGA and REA values are given by [6, 12].

$$\Lambda = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1n} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_{n1} & \lambda_{n2} & \cdots & \lambda_{nn} \end{bmatrix} \tag{6}$$

$$\Gamma = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1n} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \gamma_{n1} & \gamma_{n2} & \cdots & \gamma_{nn} \end{bmatrix} \tag{7}$$

where $\Lambda$ = Relative gain array (RGA) and $\Gamma$ = Relative exergy gain array (REA)

The definition of the REA implies that the system or process under consideration is linear so that the gains calculated are valid for the full range of operation of the control structure. In the case of non-linear systems or processes the results for the REA have to be recalculated for each range of the process where the linear assumption holds.

The REA can be used mainly as a screening tool for candidate control structures, but for a complete detailed analysis of the control structure performance, a detailed dynamic simulation is recommended.

### B. *Exergy eco-efficiency factor (EEF)*

Exergy eco-efficiency factor (EEF) is based on an understanding of the total exergy of each material stream in and out of the thermodynamic process as shown in Figure 2.

The ratio, Equation (2) is the exergetic efficiency of this process (Figure 2) which is a measure of eco-efficiency. This general process is a portion of the control loop between the manipulated and the control variables. However it does not provide any information about how the control loop configuration affects this exergetic efficiency. EEF connects the control loop configuration to the eco-efficiency. The exergy eco-efficiency factor for a control pair ($u_j$, $y_i$), is defined as,

$$\tau_{ij} = (\Delta B_{out} - \Delta B_{in}) \frac{\Delta u_j}{\Delta y_i} \qquad (8)$$

where $\Delta u_j$ denotes a step change of the *MV* $u_j$, $\Delta y_i$ denotes a response in the *CV*, $y_i$, caused by a step change of $u_j$, and $\Delta B_{out}$ and $\Delta B_{in}$ represent the exergy differences caused by the *MV* step change for exergy out and exergy in, respectively. For example, if $\tau_{21}$is less than $\tau_{22}$, it means that for the same amount of *CV* change $\Delta y_2$, using $u_1$, will cause less exergy destruction than using $u_2$. The final interpretation is that the control pairing ($u_1$, $y_2$) is more eco-efficient than the pairing ($u_2$, $y_2$).

The EEF shows how much exergy will be destroyed by using different *MV* to control the same amount of *CV* change. It can provide a quantitative measurement of the exergy consumption. By comparing the sum of EEF of one control configuration with another, the approximate amount of exergy to be saved can be obtained. This will be very useful in some situations, for example, when control configuration A makes an exergy saving of 10%, compared to control configuration B, the implementation of control configuration B is more expensive than it is for control configuration A, Thus, on this basis, control configuration A may be selected.

For a 2 x 2 example, if $\tau_{21}$ is less than $\tau_{22}$, it means that for the same amount of *CV* change, $\Delta y_2$ using *MV* $u_1$, will cause less exergy change/loss than will using *MV* $u_2$. The final interpretation is that control pair ($u_1$, $y_2$) is more eco-efficient than pair ($u_2$, $y_2$).

EEF helps engineers to build an eco-efficient process which is ecological friendly and economically viable. Since exergy accounts for the quality of energy, thus it can be used as a measure to evaluate the eco-efficiency for a process design. A process is called eco-efficient if it uses a relatively small amount of energy or the destruction of exergy is low.

Control loop configurations can be determined by techniques such as RGA, NI and SVD, it is usual that several candidate control loop configurations can be implemented. In regard to eco-efficiency, the EEF can be used to select the best control loop configuration from among the candidates in the sense of eco-efficiency.

EEF calculation for different cases with all possible control schemes is difficult. It increases the computational load on process design engineers. The development of an algorithm/software package for EEF calculation has been done, which combines together a commercial simulator, VMGSim and Excel to calculate the EEF. A potential help is obtainable in EEF calculation by using a commercial simulator VMGSim [13].

### C. *Validation of EEF Results*

Dynamic simulation is the best way to validate the proposed eco-efficiency factor. By recording the exergy consumptions of several control configurations, the most eco-efficient control configuration can be identified and dynamic results compared to the results from the eco-efficiency factor.

Dynamic exergy versus time can be approximated by several exergy calculations at different conditions during the dynamic response of a process. The exergy values of the process dynamic response at different time intervals are calculated. As chemical simulators still do not have the ability to directly calculate and display the total exergy of a material stream, these simulators cannot automatically calculate exergy versus time at every point. Simulators such as the HYSYS and the VMGSim can only calculate steady state exergy values at given process conditions. For dynamic exergy versus time, different process condition points are selected during the process dynamic response due to step input disturbances. The selection of calculation points depends on the dynamic response of the process. In order to get the maximum information in regard to dynamic response, if the variation in the process conditions with time is large, then the time interval between the selected points is decreased. With less variation in the process conditions, the time interval between the selected points is increased. Then the exergy values are calculated on these selected points during the dynamic process response. Exergy values at different points are calculated with the procedure developed in [14]. Then those exergy points are used to approximate the dynamic exergy response versus time.

## III. CASE STUDY

### A. *Process description*

A monochlorobenzene (MCB) separation process is selected for this case study, and is used to show how the REA and EEF provides information regarding the best control loop configuration among the candidates in the sense of eco-efficiency. MCB plant consists of three main units: a flash vessel (F1), an absorption column (T1) and a distillation column (T2), as shown in Figure **3**. VMGSim (process simulator) with the NRTL activity thermodynamic model was used for the simulation of the MCB separation process. The detailed information regarding feed conditions and column specification can be found in [15].

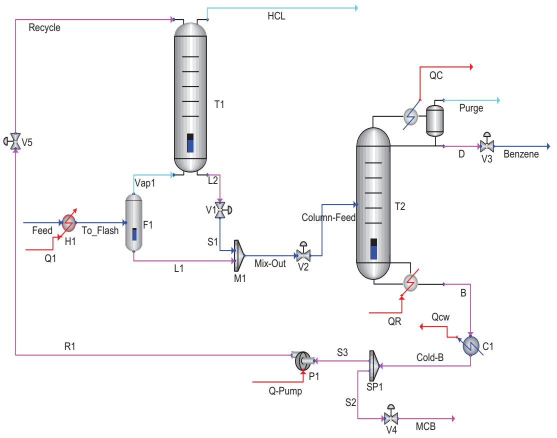


Figure 3. MCB separation process schematic

For comparison, three basic control configurations can be defined for the dual composition control of the distillation column (T2) and composition of HCl ($x_{HCl}$) leaving in the vapour stream of the absorber, namely $LVQ_{cw}$, $LBQ_{cw}$ and $DVQ_{cw}$. Each configuration is comprised of three main composition control loops (name of each configuration). For example, in the $LVQ_{cw}$ control configuration, *L (Reflux rate)* is used to control the composition of the top product ($x_D$), boil-up rate ($V$) is used to control the composition of the bottom product ($x_B$), and cooler duty ($Q_{cw}$) is used to control the composition of vapour stream leaving the absorber ($x_{HCl}$). It behaves as a pseudo 3 x 3 system because; i) three inventory control loops of distillation column and other control loops of MCB plant are assumed to be under perfect control; and ii) Control loops other than the composition control are not interacting with the composition loops, therefore are not included in the analysis.

## B. *Results and Discussion*

The simulation model shown in Figure 3 was used to do the required step tests. Step tests are performed to obtain the necessary information to calculate the RGA, DRGA, REA, NI and CN as shown in Table 1.

The $LVQ_{cw}$ and $DVQ_{cw}$ control configurations are further selected based on classical controllability tools (RGA, NI, SVD and CN) results, to ascertain the most eco-efficient control configuration ($LVQ_{cw}$ or $DVQ_{cw}$) by using the REA and EEF.

The REA results, as shown in Table 1, depict that all three control configurations ($LVQ_{cw}$, $LBQ_{cw}$ and $DVQ_{cw}$) have exergy interactions and the thermodynamic (exergetic) efficiencies of the control loops are affected by these exergy interactions. For the $LVQ_{cw}$ configuration, the REA results show that the exergy changes for open loop are smaller and opposite in direction to the exergy changes caused by loop interactions. For the $LBQ_{cw}$ control configuration, the REA results show that exergy changes due to loop interactions are almost equal and larger than the exergy changes due to the open loop operation. For the $DVQ_{cw}$ control configuration, the REA results show that exergy changes for open loop operation are much smaller and are in the same direction as the exergy changes caused in closed loop operation, except for one element.

Table 1. RGA, DRGA, REA, NI and CN results for the whole MCB plant

| Configurations | RGA | DRGA | REA | NI | CN |
|---|---|---|---|---|---|
| $LVQ_{cw}$ | $\begin{bmatrix} 6.30 & -4.76 & -0.54 \\ -5.80 & 6.50 & 0.35 \\ 0.52 & -0.70 & 1.20 \end{bmatrix}$ | $\begin{bmatrix} 3.80 & -2.80 & 0.02 \\ -1.40 & 2.60 & -0.23 \\ -1.40 & 1.20 & 1.20 \end{bmatrix}$ | $\begin{bmatrix} 0.006 & -0.25 & 1.25 \\ 1.21 & -0.23 & 0.02 \\ -0.21 & 1.48 & -0.27 \end{bmatrix}$ | 1.15 | 31 |
| $LBQ_{cw}$ | $\begin{bmatrix} 0.88 & 0.10 & 0.02 \\ 0.12 & 0.92 & -0.05 \\ 0.00 & -0.03 & 1.02 \end{bmatrix}$ | $\begin{bmatrix} 0.84 & -1.60 & 1.76 \\ 0.18 & 0.38 & 0.43 \\ -0.03 & 2.20 & -1.19 \end{bmatrix}$ | $\begin{bmatrix} 0.88 & 0.13 & -0.007 \\ 0.12 & 0.61 & 0.26 \\ -0.002 & 0.26 & 0.74 \end{bmatrix}$ | -1.0 | 16 |
| $DVQ_{cw}$ | $\begin{bmatrix} 0.52 & 0.49 & -0.01 \\ 0.39 & 0.51 & 0.10 \\ 0.10 & 0.00 & 0.91 \end{bmatrix}$ | $\begin{bmatrix} 0.91 & 0.03 & 0.06 \\ 0.03 & 1.02 & -0.05 \\ 0.06 & -0.05 & 0.99 \end{bmatrix}$ | $\begin{bmatrix} 0.11 & 0.85 & 0.03 \\ 0.83 & 0.14 & 0.03 \\ 0.05 & 0.006 & 0.94 \end{bmatrix}$ | 0.97 | 23 |

According to the guidelines for interpreting the REA results, an element in REA close to the value of 1.0 indicates that thermodynamic (exergetic) efficiency and exergy changes of the control loop are not affected by other loops. The REA results, as shown in Table 1, show that leading diagonal elements of only the $LBQ_{cw}$ control configuration are close to the value of 1.0 and that this would be a good candidate for selection because its thermodynamic (exergetic) efficiency and exergy changes are not affected by loop interactions.

The interpretation of the RGA, DRGA, NI, CN and REA results depict that a trade-off exists in this process between controllability, thermodynamic (exergetic) efficiency and exergy changes. The $LBQ_{cw}$ control configuration is the best choice for selection, from the thermodynamic (exergetic) efficiency and exergy changes points of view, but it is the worst from the controllability point of view. When the results of the classical controllability techniques (RGA, DRGA, NI and CN) and REA contradict each other (as in this case), then dynamic simulation for validation is required

As it is affected by control loop interactions, the REA provides a means for the analysis of the thermodynamic (exergetic) efficiency and exergy changes of a control loop. The REA does not provide any information concerning the eco-efficiency analysis of the whole process or plant.

Table 2. EEFs for MCB plant

| Control pairs | ($L$, $x_D$) | ($D$, $x_D$) | ($V$, $x_B$) | ($B$, $x_B$) | ($Q_{cw}$, $x_{HCl}$) |
|---|---|---|---|---|---|
| EEF (kW. kgmole/h) or (kW) | 94.3 | 3.41 | 78.72 | 1.08 E3 | 1.28 E4 |

The EEF was proposed for the eco-efficiency analysis of the whole process or plant. The EEF was developed to minimize the limitations of the REA e.g. the REA measures eco-efficiency solely within the scope of control loops. The EEF determines the eco-efficiency of the whole plant because its calculation is based on the total exergy destroyed in a process, total exergy coming in and going out of a process. A higher EEF value indicates that selection of that control configuration will result in higher exergy destruction, and vice versa. As the EEF provides the means to determine the true eco-efficiency of the whole plant, EEF analysis is preferred over REA analysis.

There are some similarities and differences between the EEF and REA. For example: the EEF is affected by the recycling of materials and energy streams like REA [10, 16]; and unlike the REA, the EEF considers a single possible control scheme at a time for analysis [3].

EEF results for whole MCB plant, as shown in Table 2, indicate that the control pair ($Q_{cw}$, $x_{HCl}$) will use the most exergy and be the least eco-efficient control pair, and the control pair ($D$, $x_D$) is the most eco-efficient control pair. Since both control configurations ($LVQ_{cw}$ and $DVQ_{cw}$) include the same control pairs ($V - x_B$, $Q_{cw} - x_{HCl}$), we only need to compare the EEF for control pairs ($L$, $x_D$) and ($D$, $x_D$). For controlling $x_D$ if we use $D$, it will make an exergy saving ≈ 2.0 % compared to use of $L$. Dynamic simulation procedure explained in section II.C is used to validate the steady state EEF results.

## IV. Conclusions

The REA integrates concepts of process control and thermodynamic efficiency (exergetic efficiency). The REA provides a means to analyze the exergetic efficiency of a control structure as it is affected by control structure loop interactions. The EEF is employed for eco-efficiency analysis of the whole process, to overcome the deficiencies in REA. The EEF facilitates in the deciding of controllable and eco-efficient control schemes for the whole process. As the EEF provides the means to determine the true eco-efficiency of the whole plant, EEF analysis is preferred over REA analysis.

The $LVQ_{cw}$ and $DVQ_{cw}$ control configurations were selected based on classical controllability techniques. Both the $LVQ_{cw}$ and the $DVQ_{cw}$ control configurations are equally controllable, but the $DVQ_{cw}$ control configuration is preferred over the $LVQ_{cw}$ because it causes less exergy destruction (eco-efficient) than the $LVQ_{cw}$ control configuration.

## Acknowledgment

The authors would like to acknowledge the support of the Industrial Information and Control Centre ($I^2C^2$), Chemical and Materials Engineering department, Faculty of Engineering, The University of Auckland, New Zealand.

## References

[1] Montelongo-Luna, J.M., W.Y. Svrcek, and B.R. Young, *The relative exergy array—a new measure for interactions in process design and control.* The Canadian Journal of Chemical Engineering, 2010. **89**(3): p. 545-549.

[2] Munir, M.T., W. Yu, and B.R. Young, *Determination of Plant-wide Control Loop Configuration and Eco-Efficiency, G.P. Rangaiah and V. Kariwala (Eds.)*, in *Plantwide Control: Recent Developments and Applications*. 2012, John Wiley & Sons, ISBN:9780470980149.

[3] Munir, M.T., W. Yu, and B.R. Young, *Control loop configuration and eco-efficiency*, in *FOCAPO/CPC-VIII*2012: Savannah, Georgia, USA.

[4] Marlin, T.E., *Process control: design process and control system for dynamic performance*. 2 ed. 2000, New York: McGraw Hill.

[5] Seborg, D.E., T. F. Edgar, and D.A. Mellichamp, *Process Dynamics and Control*. 1989, New York: John Wiley & Sons.

[6] Svrcek, W.Y., D.P. Mahoney, and B.R. Young, *A Real-Time Approach to Process Control*. 2006, Chichester: John Wiley & Sons Ltd.

[7] Seborg, D.E., T. F. Edgar, D.A. Mellichamp, and F.J. Doyle, *Process Dynamics and Control*. 3rd ed. 2010, New York: John Wiley & Sons.

[8] Kotas, T.J., *The exergy concept and exergy losses*, in *The exergy method of thermal plant analysis*. 1985, Butterworth-Heinemann Ltd: London.

[9] Kotas, T.J., *The exergy method of thermal plant analysis*. 1985, London: Butterworths. Medium: X; Size: Pages: 344.

[10] Munir, M.T., W. Yu, and B.R. Young, *Recycle effect on the relative exergy array.* Chemical Engineering Research and Design, 2012. **90**(1): p. 110-118.

[11] Bristol, E.H., *On a new measure of interactions for multivariable process control.* IEEE Trans. Automat. Control, 1966. **11**: p. 133-134.

[12] Montelongo-Luna, J.M., W.Y. Svrcek, and B.R. Young, *The relative exergy array - a new measure for interactions in process design and control.* The Canadian Journal of Chemical Engineering, 2010: p. Accepted.

[13] Munir, M.T., W. Yu, and B.R. Young, *A software algorithm/package for control loop configuration and eco-efficiency* ISA Transactions: (Accepted), 2012.

[14] Munir, M.T., J.J. Chen, and B.R. Young, *A computer program to calculate the stream exergy using the visual basic graphical interface*, in *Chemeca 2010*2010: Adelaide, Australia.

[15] Seider, W.D., J.D. Seader., and D.R. Lewin., *Product and Process Design Principles: Synthesis, Analysis, and Evaluation*. 2nd ed. 2004, New York: John Wiley.

[16] Munir, M.T., W. Yu, and B.R. Young, *Eco-efficiency and control loop configuration for recycle systems*, in *American control conference*2012: Montreal, Canada.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Low Eigenvalue Sensitivity Eigenstructure Assignment to Linear Parameter Varying Systems

Fengming Shi and Ron J. Patton
Department of Engineering
University of Hull
Hull, HU6 7RX, UK
shi_fengming@163.com, R.J.Patton@hull.ac.uk

*Abstract*— **This paper is concerned with the assignment of a desired eigenstructure to linear parameter-varying (LPV) systems as an extended version of the corresponding eigenstructure assignment problem for linear time-invariant systems. Based on a complete parametric solution of parametric generalized Sylvester matrix equation, a controller design method is proposed to guarantee a low sensitivity of the closed-loop eigenvalues. The observer state feedback structure is considered for output feedback control design. An example of control for a satellite attitude system is used to demonstrate the usefulness of the proposed approach.**

***Keywords-*** *Eigenstructure Assignment, Linear Parameter Varying systems, Low Eigenvalue Sensitivity, Sylvester Matrix Equation, Observer State Estimate Feedback*

## I. Introduction

Gain scheduling control has been widely used in practical applications [1, 2] to handle the nonlinearity of real systems. The classic gain scheduling approach consists in designing linear controllers for several operating points and then applying an interpolation strategy to obtain a global controller. Consequently, powerful tools for linear systems can be applied to nonlinear plants. In spite of the numerous applications, there has been no formal framework for gain scheduling until the early 1990s [3]. This framework gives heuristic rules to ensure global stability, but it does not provide a systematic design procedure. The linear parameter varying (LPV) approach to system modeling, estimation and control followed on from gain-scheduling as a strategy for attempting to model non-linear parametric variations using a time-varying linear systems approach [4, 5]

Eigenstructure assignment has been used in many applications and has been proven to be a useful tool both for analysis and design of linear time invariant (LTI) systems [6-8]. This method allows the designer to satisfy directly damping, settling time, and mode decoupling specifications by choosing the eigenvalues and eigenvectors. That is because the transient response of an LTI system is completely specified by the system eigenstructure. Generally, the eigenvalues determine the decay (or growth) rate of the response and the left and right eigenvectors fix the shape of the response [8]. Also, minimum eigenvalue sensitivity to model parameter variation and other performance requirements such as minimum gain control can be accommodated using explicit choices of the free controller parameters [8]. For years, many researchers have attempted to generalize the conventional notions of eigenvalues and eigenvectors for linear time-invariant systems to linear time-varying (LTV) systems [9-12]. Although existing eigenstructure assignment techniques and algorithms for LTV systems confirm the value of using this approach, eigenstructure assignment of LTV systems remains a difficult problem that is still in a state of development.

However, there are very few examples in the literature where eigenstructure assignment is applied within an LPV framework [13, 14]. In [13, 15], polynomial eigenstructure assignment of LTI systems was extended to solve the corresponding eigenstructure assignment problem for LPV systems using output feedback. In [14], a parametric approach for eigenstructure assignment, appropriate for LTI systems, was extended to LPV plants using state feedback. However, with the conditions given in [13], the choice of controller structure, the matching conditions and the solution of the controller are not unique and  require much more additional criteria to constrain the order of controller gains. As discussed in [16], the method used in [14] to calculate the controller parameters is complex.

In this paper, a low eigenvalue sensitivity eigenstructure assignment method for LPV systems is proposed. A general complete parametric solution of the corresponding parametric generalized Sylvester matrix equation [14] is introduced. A parametric eigenstructure assignment is presented based on the proposed solution approach. Using the eigenstructure assignment design freedom, low eigenvalue sensitivity is achieved by projecting the desired eigenstructure into an allowable subspace. An observer-based state estimate feedback controller structure is chosen within an output feedback framework. An algorithm is proposed to calculate a state feedback controller with state observer. The remainder of this paper is organized as follows. Section II briefly introduces a parametric solution of the parametric generalized Sylvester matrix equation, and the eigenstructure assignment to LPV system is presented. In Section III, the definition of the overall eigenvalue sensitivity of matrix to LTI systems is reviewed and extended to LPV systems. The main results and an algorithm are also presented in Section III. Section IV gives an example of a satellite system to illustrate the application of the theory. Finally, conclusions are drawn up in Section V.

Fengming Shi acknowledges joint scholarship funding support from the CSC (China) and the Hull China Scholarship.

## II. Parametric Eigenstructure Assignment for LPV systems

### A. LPV systems

Reference [5] considers that LPV Systems are linear time-varying plants whose state-space matrices are fixed functions of some vector of varying parameters $\theta(t)$. LPV systems can be described by state-space equations of the form:

$$\left.\begin{aligned}\dot{\boldsymbol{x}}(\boldsymbol{t}) = \boldsymbol{A}\big(\boldsymbol{\theta}(\boldsymbol{t})\big)\boldsymbol{x}(\boldsymbol{t}) + \boldsymbol{B}\big(\boldsymbol{\theta}(\boldsymbol{t})\big)\boldsymbol{u}(\boldsymbol{t})\\ \boldsymbol{y}(\boldsymbol{t}) = \boldsymbol{C}\big(\boldsymbol{\theta}(\boldsymbol{t})\big)\boldsymbol{x}(\boldsymbol{t}) + \boldsymbol{D}\big(\boldsymbol{\theta}(\boldsymbol{t})\big)\boldsymbol{u}(\boldsymbol{t})\end{aligned}\right\} \tag{1}$$

where $x(t) \in R^n, u(t) \in R^r and\ y(t) \in R^m$ are the state vectors, the input vectors and measured output vector, respectively. $A(.), B(.), C(.), D(.)$, with corresponding dimensions, are known continuous function of a time-varying parameters vector $\theta(t)$ which satisfies:

$$\theta(t) = \left[\theta_1(t), \cdots \theta_{n_\theta}(t)\right]^T \in \Theta, \forall\, t \geq 0$$

where $\Theta$ is a compact set. The subscript $t$ is omitted through the remainder of the paper without causing confusion.

From a practical point of view, LPV systems have at least two interesting interpretations. They can be viewed as LTI plants subject to time-varying parametric uncertainty $\theta(t)$. On the other hand, they can be models of linear time-varying plants or result from the linearization of nonlinear plants along the trajectories of the parameter θ. From the second view, the parameter $\theta(t)$ can be measured in real time during system operation. Consequently, the control strategy can exploit the available measurements of θ to increase performance. The LPV controller design approach proposed in this paper is based on the second view point.

### B. Parametric Eigenstructure Assignment using state feedback

Consider an LPV system given in form of (1). A linear parameterised state feedback law:

$$u = K(\theta)x, K(\theta) \in R^{r\times n}$$

is applied, such that the closed-loop system is in the following form:

$$\dot{x} = (A(\theta) + B(\theta)K(\theta))x$$

Following [9, 14], the closed-loop self-conjugate eigenvalue set can be described as $\Lambda = \{\lambda_i(\theta): \lambda_i(\theta) \in C, i = 1,2,..\tilde{n}, 1 \leq \tilde{n} \leq n\}$, for which the algebraic and geometric multiplicities of the eigenvalue $\lambda_i$ are denoted by $q_i$ and $r_i$, respectively. Then in the Jordan form of the matrix $A_{cl}(\theta) = A(\theta) + B(\theta)K(\theta)$, there are $r_i$ Jordan blocks, associated with the i-th eigenvalue $\lambda_i$, of orders $p_{ij}$, $j = 1,2,\ldots, r_i.$ , $p_{ij}$, $q_i\ and\ r_i$ satisfy the relations:

$$\sum_{j=1}^{r_i} p_{ij} = q_i, \sum_{i=1}^{\tilde{n}} q_i = n$$

Denoting the left and right eigenvectors and generalized eigenvectors of matrix $A_{cl}(\theta)$ associated with $\lambda_i$ by $L_i(\theta)$ and $R_i(\theta)$, respectively, it follows that:

$$(\lambda_i(\theta)I - A_{cl}(\theta))R_{ij,k}(\theta) = -R_{ij,k-1}(\theta),$$

$$R_{ij,0}(\theta) = 0$$

$$(\lambda_i(\theta)I - A_{cl}(\theta))^T L_{ij,k}(\theta) = -L_{ij,k-1}(\theta),$$

$$L_{ij,0}(\theta) = 0$$

for $k = 1,2, \ldots p_{ij}, j = 1,2, \ldots r_i, and\ i = 1,2, \ldots, \tilde{n}$
and

$$L^T(\theta)R(\theta) = I$$

Hence, the problem to assign a desired closed-loop eigenstructure to a system using a state feedback controller is to find a solution of the parametric equation:

$$A(\theta)R(\theta) + A(\theta)W(\theta) = R(\theta)F(\theta)$$

where $A(\theta), B(\theta)$ are the state space matrices, $R(\theta)$ is the desired right eigenvector matrix, $F(\theta)$ is the desired eigenvalues *diagnosis matrix* and $W(\theta)$ is an auxiliary matrix.

Theorem 1 is introduced to show how the eigenvectors and generalized eigenvectors can be parameterised.

**Theorem 1 [14]**

Let $[A(\theta)\ B(\theta)]$ be controllable, and matrix $B(\theta)$ be of full-column rank, then all the solutions of matrix equation:

$$A(\theta)R(\theta) + B(\theta)W(\theta) = R(\theta)F(\theta)$$

are given by:

$$\begin{bmatrix} R_{ij}^k \\ w_{ij}^k \end{bmatrix} =$$

$$\begin{bmatrix} N\big(\theta, \lambda_i(\theta)\big) & \cdots & \frac{1}{(k-1)!}\frac{d^{k-1}}{d\lambda^{k-1}} N\big(\theta, \lambda_i(\theta)\big) \\ D\big(\theta, \lambda_i(\theta)\big) & \cdots & \frac{1}{(k-1)!}\frac{d^{k-1}}{d\lambda^{k-1}} D\big(\theta, \lambda_i(\theta)\big) \end{bmatrix} \begin{bmatrix} f_{ij}^k(\theta) \\ \vdots \\ f_{ij}^1(\theta) \end{bmatrix} \tag{2}$$

$$\mathrm{k} = 1, 2, \ldots., \mathrm{p}_{ij}, \mathrm{j} = 1,2, \ldots.\mathrm{q}_i, \mathrm{i} = 1,2, \ldots.\mathrm{n}'$$

where $f_{ij}^k \in C^r$ are arbitrarily chosen parameter vectors; N(θ, λ) and D(θ, λ) are right co-prime matrix polynomials satisfying:

$$[\lambda(\theta)I - A(\theta)]^{-1}B(\theta) = N\big(\theta, \lambda(\theta)\big)D^{-1}\big(\theta, \lambda(\theta)\big) \tag{3}$$

From Theorem 1, it can be seen that the desired eigenvectors and generalized eigenvectors can be parameterised by (2) [14].

**Remark 1:**

- Theorem 1 concerns an extension of the eigenstructure assignment of LTI systems to an LPV modeling framework by introducing a solution of the parametric Sylvester matrix equation.
- Theorem 1 gives a clear analytical, complete, and explicit parametric solution expressed by the eigenvalues of the matrix $A_{cl}(\theta)$ and a group of free parameters, namely $f_{ij}$ . By specially choosing the free parameters given in (2), solutions with desired properties can be obtained.

## III. Main Result

It is known in [8, 17] that if a system has low sensitivity to perturbations and parameter variations in the system matrices then there may be a low chance of the closed-loop system becoming unstable compared with the case when controllers are used that are not based on sensitivity minimization. Hence, the eigenvalue sensitivity of a closed-loop system to modeling errors should be given suitable consideration. For the sake of simplicity, only the overall eigenvalue (i.e. Wilkinson) sensitivity is considered here [8, 17].

### *A. Overall Eigenvalue Sensitivity*

The overall eigenvalue sensitivity of the matrix $X$ is defined [8, 17] as:

$$\eta(R) = \|R\|_2 \|R^{-1}\|_2$$

where $R$ is the right eigenvector matrix of the matrix $X$.

Similarly, in this study the overall eigenvalue sensitivity of the parameter varying matrix $X(\theta)$ is defined as:

$$\eta(R(\theta)) = \sup_{\theta\in\Theta} \|R(\theta)\|_2 \|R(\theta)^{-1}\|_2$$

where $R(\theta)$ is the right eigenvector matrix of the matrix $X(\theta)$.

Suppose that the right eigenvector matrix $R$ is unitary, i.e., $R^T R = I$ then $\eta(R) = 1$ . This indicates that if $R$ is a unitary matrix the corresponding eigenvalues are perfectly conditioned and hence minimally sensitive to perturbations or parameter variations.

These observations provide the basis for the algorithms to be described in this paper. The objective of the paper is to show how to assign a set of closed-loop eigenvectors which match the columns of a unitary matrix as closely as possible. If this process is successful, a perfectly conditioned set of closed-loop eigenvalues results [8, 17].

### *B. Performance function*

The desired eigenvectors must be projected into the allowable subspace which is optimal according to some performance function while the desired eigenvectors are not in the allowable subspace. As argued previously, to achieve overall low eigenvalue sensitivity, an LPV system performance function can be defined as:

$$J_p = \left(R_{d,i}(\theta) - R_i(\theta)\right)^T \widetilde{W}_i \left(R_{d,i}(\theta) - R_i(\theta)\right) \qquad (4)$$

where θ is the varying parameter and other symbols have the same meaning as in LTI system case.

If the right eigenvectors are parameterized as:

$$R_i(\theta) = P_{R,i}(\theta) W_i(\theta)$$

The solution that minimises the LPV performance function $J_p$ is obtained by setting:

$$W_{o,i}(\theta) = \left(P_{R,i}^T(\theta)\widetilde{W}_i P_{R,i}(\theta)\right)^{-1} P_{R,i}^T(\theta)\widetilde{W}_i R_{d,i}(\theta) \qquad (5)$$

The least-squares best-fit LPV right eigenvectors can be computed by:

$$R_{o,i}(\theta) = P_{R,i}(\theta) W_{o,i}(\theta)$$

Now set $R_{d,}(\theta) = I$ , the low eigenvalue sensitivity performance function of LPV system is:

$$J_{pl} = \left(R_{d,i}(\theta) - I_i(\theta)\right)^T \widetilde{W}_i \left(R_{d,i}(\theta) - I_i(\theta)\right)$$

### *C. Results and Algorithms*

Consider an LPV system described by (1), and the performance function is described as (4). The solution that minimises the performance function $J_{pl}$ is obtained by setting:

$$\begin{bmatrix} f_{oij}^k(\theta) \\ \vdots \\ f_{oij}^1(\theta) \end{bmatrix} = \left(P_{R,ijk}^T(\theta)\widetilde{W}_i P_{R,ijk}(\theta)\right)^{-1} P_{R,ijk}^T(\theta)\widetilde{W}_i I_{,i}(\theta) \; (6)$$

where:

$$P_{R,ijk}(\theta) = \left[N\left(\theta,\lambda_i(\theta)\right) \quad \cdots \quad \frac{1}{(k-1)!}\frac{d^{k-1}}{d\lambda^{k-1}} N\left(\theta,\lambda_i(\theta)\right)\right], \; (7)$$

$$\mathrm{k} = 1, 2, \ldots., \mathrm{p}_{ij}, \mathrm{j} = 1,2, \ldots.\mathrm{q}_i, \mathrm{i} = 1,2, \ldots.\mathrm{n'}$$

The least-squares best-fit LPV right eigenvector can be computed by

$$R_{oij}^k(\theta) = P_{R,ijk} \begin{bmatrix} f_{oij}^k(\theta) \\ \vdots \\ f_{oij}^1(\theta) \end{bmatrix} =$$

$$\left[N\left(\theta,\lambda_i(\theta)\right) \quad \cdots \quad \frac{1}{(k-1)!}\frac{d^{k-1}}{d\lambda^{k-1}} N\left(\theta,\lambda_i(\theta)\right)\right] \begin{bmatrix} f_{oij}^k(\theta) \\ \vdots \\ f_{oij}^1(\theta) \end{bmatrix},$$

The corresponding auxiliary matrix can be computed as:

$$w_{oij}^k(\theta) = \left[D\left(\theta,\lambda_i(\theta)\right) \quad \cdots \quad \frac{1}{(k-1)!}\frac{d^{k-1}}{d\lambda^{k-1}} D\left(\theta,\lambda_i(\theta)\right)\right] \begin{bmatrix} f_{oij}^k(\theta) \\ \vdots \\ f_{oij}^1(\theta) \end{bmatrix}$$

The following introduces a non-iterative method which involves the direct projection of a unitary matrix into the allowable eigenvector subspace.

**Algorithm 1**

**Step** 1: Chose a set of desired closed-loop eigenvalues $\Lambda = \{\lambda_i(\theta): \lambda_i(\theta) \in \mathrm{C}, \mathrm{i} = 1,2,..\tilde{\mathrm{n}},\ 1 \le \tilde{\mathrm{n}} \le n\}$

**Step** 2: Get $N\ (\theta, \lambda\ (\theta))$ and $D\ (\theta, \lambda\ (\theta))$ satisfying (3) by applying right co-prime factorization

**Step** 3: Project each column of the unitary matrix $U = [U_1 U_2 \cdots U_3]$ into each of the allowable eigenvector subspaces corresponding to each closed-loop eigenvalue using (6) and (7). For each column $U_i$ of $U$, this produces a total of $n$ achievable right eigenvectors $R_{ij}(\theta), j = 1, ..n$;

**Step** 4: Calculate the $n^2$ misalignment angles given by:

$$\alpha_{ij}(\theta) = \sup_{\theta \in \Theta} \cos^{-1}\left(\frac{|U_j^T R_{ij}(\theta)|}{\|U_j\|_2 \|R_{ij}(\theta)\|_2}\right)$$

**Step** 5: Choose the assignment from the $n!$ possibilities which have the smallest sum of misalignment angles $\alpha_{ij}$.

**Step** 6: Calculate the controller $K(\theta) = W(\theta)R(\theta)^{-1}$

**Remark 2**

- The algorithm is an LPV-extended version of the existing right eigenstructure assignment scheme via the Sylvester matrix equation for LTI systems.
- The above algorithm will assign the closed-loop eigenvectors as close to a unitary matrix as possible to achieve optimum sensitivity.
- In the algorithm, if the rank of the control input matrix $B(\theta)$ is equal to the rank of the system matrix $A(\theta)$the desired right eigenvectors (a unitary matrix) as well as the desired eigenvalues can be achieved exactly.
- For a special case, if the desired eigenvalue is constant $\Lambda = \{\lambda_i: \lambda_i \in \mathrm{C}, \mathrm{i} = 1,2,..\tilde{\mathrm{n}},\ 1 \le \tilde{\mathrm{n}} \le n\}$, and if the achieved eigenvectors are parameter-independent, then the closed-loop system could be time invariant.
- As suggested in [8] $n!$ could be very large when $n$ is large. Hence, $n < 7$ is suggested to use in the proposed approach.
- For the Step 4, grid method [18, 19] is proposed to tackle the high dimensionality and nonlinear nature of the optimization problem.

## IV. An Example

Now, an example is given of a satellite attitude control problem to show how to calculate the controller for a given LPV system to achieve low eigenvalue sensitivity. The example is given in [14]. Consider an LPV system:

$$\left.\begin{aligned} \dot{x} &= A(\theta)x + B(\theta)u \\ y &= Cx + Du \end{aligned}\right\} \tag{8}$$

where

$$A(\theta) = \begin{bmatrix} 0 & a12 & 0 & 0 & a15 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & a34 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ a51 & 0 & 0 & 0 & 0 & a56 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$B(\theta) = \begin{bmatrix} b11 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & b32 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & b53 \\ 0 & 0 & 0 \end{bmatrix}, C = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$D = 0$$

The coefficients of $A(\theta)\ and\ B(\theta)$ are defined as:

$$a12 = 4(\mathrm{Iz} - \mathrm{Iy})\frac{w_0^2}{\mathrm{Ix}}, \mathrm{a15} = 4(\mathrm{Ix} + \mathrm{Iz} - \mathrm{Iy})\frac{\mathrm{w_0}}{\mathrm{Ix}},$$

$$\mathrm{a51} = (\mathrm{Iy} - \mathrm{Iz} - \mathrm{Ix})\frac{\mathrm{w_0}}{\mathrm{Iz}}, \mathrm{a56} = (\mathrm{Ix} - \mathrm{Iy})\frac{w_0^2}{\mathrm{Iz}},$$

$$\mathrm{a34} = 3(\mathrm{Iz} - \mathrm{Ix})\frac{w_0^2}{\mathrm{Iy}}, \mathrm{b11} = \frac{1}{\mathrm{Ix}}, \mathrm{b32} = \frac{1}{\mathrm{Iy}}, \mathrm{b53} = \frac{1}{\mathrm{Iz}},$$

$$\mathrm{w_0} = 1.1 * 10^{-3}, \mathrm{Ix} = 661.4723\ , \mathrm{Iy} \in [2620 \quad 3700],$$

$$\mathrm{Iz} \in [2780 \quad 3850]$$

$\theta$ is defined as $\theta = \{\theta_1, \theta_2\} = \{\mathrm{Iy}, \mathrm{Iz}\}$, where Iy and Iz are the moment of inertia parameters.

### A. Controller design

The desired eigenvalues are set to be the same as in [18]:

$$\Lambda = \{-1.00 \pm 2.063i \quad -1.023 \pm 2.354i, \quad -1.592 \pm 1.676i\}$$

Using elementary transformations and the rational matrix factorization method [20, 21], it follows that:

$$N(\theta, s) = \begin{bmatrix} s & 0 & 0 \\ 1 & 0 & 0 \\ 0 & s & 0 \\ 0 & 1 & 0 \\ 0 & 0 & s \\ 0 & 0 & 1 \end{bmatrix},$$

$$D(\theta, s) = \begin{bmatrix} \frac{s^2 - a12}{b11} & 0 & \frac{s * a15}{b11} \\ 0 & \frac{s^2 - a34}{b32} & 0 \\ \frac{s * a51}{b53} & 0 & \frac{s^2 - a56}{b53} \end{bmatrix}$$

The desired eigenvector matrix is set to be the identity matrix $I_6$ to force the closed-loop system to have low eigenvalue sensitivity to parameter uncertainty.

Using the above algorithm, the desired eigenvector Identity Matrix ($I_6$) is projected to the allowable subspace. The calculated controller is:

$$K = \begin{bmatrix} \frac{3.185}{b11} & k_{12} & 0 & 0 & \frac{a15}{b11} & 0 \\ 0 & 0 & \frac{2.0466}{b32} & k_{24} & 0 & 0 \\ \frac{a51}{b53} & 0 & 0 & 0 & \frac{2}{b53} & k_{36} \end{bmatrix} \quad (9)$$

$$k_{12} = \frac{5.3454-\text{a12}}{b11}, k_{24} = \frac{6.588+\text{a34}}{b32}, k_{36} = \frac{5.2539+\text{a56}}{b53}$$

When the designed state feedback controller (9) is applied to the original model (8); the closed-loop matrix is obtained as in (10).

$$A - BK = \begin{bmatrix} -3.185 & a_{cl} & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -2.0466 & -6.588 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -2 & -5.2539 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \quad (10)$$

$$a_{cl} = 2 * a12 - 5.3454$$

### B. Observer design

Based on the Separation Principle, an observer state estimate feedback is used to achieve a form of output feedback control. The observer design can be achieved by recognizing the duality between the state feedback control and state estimation problems. A full order observer for the LPV system is considered with the following structure:

$$\left.\begin{aligned} \dot{\hat{x}} &= A(\theta)\hat{x} + B(\theta)u + L(\theta)(y - \hat{y}) \\ \hat{y} &= C\hat{x} + Du \end{aligned}\right\} \quad (11)$$

where $\hat{x} \in R^n$ is the estimated state and $L(\theta)$ is the designed observer gain.

To make the estimated states converge to real system states fast enough, the real part of the observer eigenvalues should be large enough. So, the desired eigenvalues for the observer system are chosen as:

$$\Lambda = \{-3 \pm 3\text{i}, \quad -6 \pm 3\text{i}, \quad -9 \pm 6\text{i}\}$$

Using the proposed procedure, the obtained observer gain is:

$$L = \begin{bmatrix} 72.04 & 18.03 & 0 & 0 & \frac{a15}{b11} & 0 \\ 0 & 0 & \frac{2.046}{b32} & \frac{6.588 + \text{a34}}{b32} & 0 & 0 \\ \frac{a51}{b53} & 0 & 0 & 0 & \frac{2}{b53} & \frac{5.254 + \text{a56}}{b53} \end{bmatrix}$$

### C. Simulation Result

In the simulation model, the initial conditions are set to be $x(0) = [0, 0.0175, 0, 0.0175, 0, 0.0175]^T$. The simulation results are shown in Fig. 1. From the time responses it can be seen that their steady-state errors converge to zero asymptotically. And from the result of the closed-loop system matrix, it can be seen that the modes are decoupled from each other. This is because the system eigenstructure is also considered which would make the system more insensitive to parameter perturbations.

## V. Conclusion

In this paper, a low eigenvalue sensitivity eigenstructure assignment approach is presented for LPV systems via observer/state feedback, based on the complete parametric solution of a parametric generalized Sylvester matrix equation. Furthermore, for the sake of practical applications, an example is used to demonstrate the usefulness and advantage of the proposed LPV control scheme. The results show that the closed-loop system transient response performance requirements are satisfied and low eigenvalue sensitivity to perturbation and parameter variation is achieved. The dynamic output feedback controller should be considered in the case where the Separation Principle breaks down.

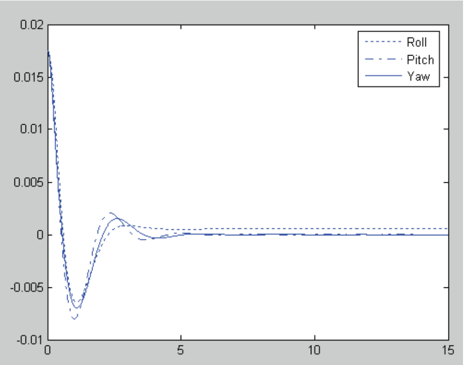


Figure 1. Observer/state feedback closed-loop time response

## VI. References

[1] W. J. Rugh and J. S. Shamma, "Research on gain scheduling," Automatica, vol. 36, pp. 1401-1425, 2000.

[2] D.J.Leith and W. E.Leithead, "Survey of Gain-Scheduling Analysis and Design," Journal of Control, vol. 73, pp. 1001--1025, 1999.

[3] J. S. Shamma and M. Athans, "Analysis of gain scheduled control for nonlinear plants," IEEE Transactions on Automatic Control, vol. 35, pp. 898-907, 1990.

[4] J. S. Shamma and M. Athans, "Gain Scheduling: Potential Hazards and Possible Remedies," in American Control Conference, 1991, pp. 516-521.

[5] P. Apkarian, P. Gahinet, and G. Becker, "Self-scheduled $H_\infty$ control of linear parameter-varying systems: a design example," Automatica, vol. 31, pp. 1251-1261, 1995.

[6] K. M. Sobel, E. Y. Shapiro, and A. N. Andry, "Eigenstructure assignment," International Journal of Control, vol. 59, pp. 13-37, 1994.

[7] B. A. White, "Eigenstructure assignment: a survey," Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering, vol. 209, pp. 1-11, 1995.

[8] G. P. Liu and R. Patton, Eigenstructure Assignment for Control System Design: John Wiley& Sons, Inc., 1998.

[9] G. Duan, G. Wu, and W. Huang, "Eigenstructure Assignment for Linear Time-Varying Systems," Science in China (Series A, English Eition), vol. 34, pp. 246-256, 1991.

[10] H. C. Lee and J. W. Choi, "Linear time-varying eigenstructure assignment with flight control application," IEEE Transactions on Aerospace and Electronic Systems, vol. 40, pp. 145-157, 2004.

[11] J. J. Zhu, "A necessary and sufficient stability criterion for linear time-varying systems," in Proceedings of the Twenty-Eighth Southeastern Symposium on System Theory, 1996, pp. 115-119.

[12] J. W. Choi, H. C. Lee, and J. J. Zhu, "Decoupling and tracking control using eigenstructure assignment for linear time-varying systems," International Journal of Control, vol. 74, pp. 453-464, 2001.

[13] B. A. White, L. Bruyere, and A. Tsourdos, "Missile autopilot design using quasi-LPV polynomial eigenstructure assignment," IEEE Transactions on Aerospace and Electronic Systems, vol. 43, pp. 1470-1483, 2007.

[14] G. Cai, C. Hu, and G. Duan, "Eigenstructure assignment for linear parameter-varying systems with applications," Mathematical and Computer Modelling, vol. 53, pp. 861-870, 2011.

[15] F. Wang, A. Tsourdos, R. Zbikowski, and B. A. White, "Quasi-LTV Polynomial Eigenstructure Assignment Control for Formation Flying Around Sun-Earth L2 Point," Automatic Control in Aerospace, vol. 18, 2007.

[16] L. Zhang, "The attitude dynamics and control of on-orbit refueling spacecraft," Master, Harbin Institute of Technology, Harbin, 2009.

[17] J. H. Wilkinson, The algebraic eigenvalue problem: Oxford University Press, 1965.

[18] F. Wu, "Control of Linear Parameter Varying Systems," Ph.D, University of California at Berkeley, 1995.

[19] A. Packard and M. Kantner, "Gain scheduling the LPV way," in Decision and Control, 1996., Proceedings of the 35th IEEE, 1996, pp. 3938-3941 vol.3934.

[20] G.-R. Duan, "On the solution to the Sylvester matrix equation AV+BW=EVF," Automatic Control, IEEE Transactions on, vol. 41, pp. 612-614, 1996.

[21] G. R. Duan, "Solutions of the equation AV+BW=VF and their application to eigenstructure assignment in linear systems," IEEE Trans. on Automatic Control, vol. 38, pp. 276-280, 1993.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# *Phase modulation of robust variable structure control for nonlinear aircraft*

Yimeng Tang, Ron J Patton
Department of Engineering
University of Hull
Hull, UK
Yimeng.Tang@2010.hull.ac.uk; r.j.patton@hull.ac.uk

*Abstract*—**This paper concerns phase plane description and modulation of the performance for a nonlinear Unmanned Aerial Vehicle (UAV) system based on Variable Structure Control (VSC) by reaching sliding mode. The novelty lies in the application of a phase analysis approach to achieve a robust controller for a complex nonlinear system. The aircraft dynamics are introduced and approximately linearized and decoupled on-line using feedback linearization theory. Then the sliding mode control (SMC) scheme is accomplished for the decoupled sub-channels. The phase modulation method is applied for theoretically ensuring further robustness. The simulation results demonstrate the efficiency and effectiveness of the proposed strategy.**

***Keywords-nonlinear aircraft; feedback linearization; sliding mode; phase modulation; robust controller***

## I. INTRODUCTION

Flight control systems have strict real application requirements to achieve high reliability against model uncertainties. The model-based approach to sustainable control for dynamic systems (based on analytical redundancy instead of hardware redundancy) has long been emphasized for achieving robustness and minimizing the effects of modelling uncertainty to the system [1-4]. Some approaches are based on robust control or passive fault tolerant control (FTC) as an alternative to achieve system reconfiguration [5-7]. Here we consider the robust approach for passive flight FTC system.

Advanced high-performance aircraft, not only have the characteristics of high nonlinearity and are Multiple Input and Multiple Output (MIMO) from a control standpoint, but also require high manoeuvrability with static instability [8]. For the purpose of efficiency and simplification, the feedback linearization technique is well-proven and has been developed to be one of the feasible control strategies in the study of nonlinear system, especially for aircraft [9], [10]. Feedback linearization can remove nonlinear features from the system and provide a linearized and decoupled closed-loop form. In addition to these features, dynamic linearization has advantages such as insensitivity to parameter changes and disturbances, and simplicity in physical realization [11-13].

For the linearized and decoupled aircraft system, a further robust controller is required to achieve tracking accuracy and passive FTC performance. As a main mode of VSC, the SMC technique turns out to be characterized by high simplicity and robustness [14], [15]. The main idea at the basis of the SMC strategy is the design of a particular control surface to coerce the controlled system trajectories into the sliding manifold to achieve expected performance via rapid switching between positive and negative control gains, resulting in variable structure of the control law. An advantage of sliding behavior is its insensitivity similar as the on-line feedback linearization strategy. The undesired phenomenon of so-called "chattering" of real-time SMC system, which is due to the finite switching frequency, could be avoided by using an approximated saturation function instead of the sign function during SMC system design [16-18]. In this study, the simultaneously worked feedback linearization controller and sliding mode controller optimize the system response against most of disturbances and even chattering.

As another aspect to this work, the literature of the development of SMC within phase modulation is well summarized in the work of [19] and [20]. By just using a real-time system response signal, a simple and effective phase modulation strategy can be used to purposely rectify the controller structure and parameters. This method also facilitates an approach to theoretical analysis for ensuring robustness of the SMC system, before adding external disturbances for certification. The approach, also presents a unique and efficient criterion for studying system characterization, especially when using linearised systems approaches for which the most suitable mathematical description is difficult to conclude because of unpredictable error during linearization and decoupling.

The contribution of this paper lies in the application of the combined SMC strategy based on feedback linearization with phase modulation theory to a nonlinear aircraft system. The aircraft example includes full force and moment longitudinal and lateral dynamics, which are deliberately linearised and decoupled into three second order sub-systems. The phase diagrams are then feasibly obtained and successfully used as criteria for achieving system robustness. The designed SMC systems illustrate the efficiency for real-time application.

Section II introduces the theoretical foundation of the control strategies. The mathematical model and linearization processing for a nonlinear UAV, the Machan, are introduced in Section III. The modulation analysis and simulation results are

Yimeng Tang's study at the University of Hull [2010-2013] is supported by the China Scholarship Council (CSC) and the Hull-China scholarship.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

given in Section IV to illustrate the proposed approach. The concluding discussion is given in Section V.

## II. Control System Scheme

### A. Nonlinear Feedback Linearization

The concept of feedback linearization makes use of the principle of transforming a smooth non-linear dynamical system into linear input-output form [21].

For the MIMO affine nonlinear system:

$$\left.\begin{aligned} x &= f(x) + G(x)u \\ y &= H(x) \end{aligned}\right\} \tag{1}$$

where $x = (x_1, \ldots, x_n)^T$ is an n-D (dimensional) states vector of the system, $u \epsilon R^m$ and $y \epsilon R^m$ are the input and output vectors of the system, $f \epsilon R^n$ is a sufficiently smooth vector field. $G(x) = [g_1\,(x)\; g_2\,(x) \ldots g_m\,(x)]^T$, $g_i\;(i = 1,2, \ldots, m)$ is an m-D sufficiently smooth vector field , $H(x) = [h_1\,(x)\; h_2\,(x) \ldots h_m\,(x)]^T$, $h_i\;(i = 1,2, \ldots, m)$ is a sufficiently smooth scalar function.

For the affine nonlinear system (1), define $\gamma_i$ to be the smallest integer such that at least one of the inputs appears in $y_i^{(\gamma_i)}$ using Lie derivatives as:

$$y_i^{(\gamma_i)} = L_f^{\gamma_i} h_i + \sum_{j=1}^{m} (L_{g_j} L_f^{\gamma_i - 1} h_i u_j) \tag{2}$$

with at least one of the $L_{g_j} L_f^{\gamma_i - 1} h_i \neq 0 \forall x$, and $u_j$ is the $jth$ row of $u$. The input-output relation can then be defined as:

$$\begin{bmatrix} y_1^{(\gamma_1)} & y_2^{(\gamma_2)} & \ldots & y_m^{(\gamma_m)} \end{bmatrix}^T = A(x) + B(x)u \tag{3}$$

$$A(x) = \begin{bmatrix} L_f^{\gamma_1} h_1(x) & L_f^{\gamma_2} h_2(x) & \ldots & L_f^{\gamma_m} h_m(x) \end{bmatrix}^T \tag{4}$$

$$B(x) = \begin{bmatrix} L_{g_1} L_f^{\gamma_1 - 1} h_1 & \cdots & L_{g_m} L_f^{\gamma_1 - 1} h_1 \\ \vdots & \ddots & \vdots \\ L_{g_1} L_f^{\gamma_m - 1} h_m & \cdots & L_{g_m} L_f^{\gamma_m - 1} h_m \end{bmatrix} \tag{5}$$

If the matrix $A(x) \epsilon R^{m \times m}$ is invertible, then the system can be linearized by decoupling the non-linear terms in (3) by choosing $u$ as follows:

$$u = B^{-1}(x)[-A(x) + v] \tag{6}$$

which leads to the closed-loop decoupled, linear system:

$$\begin{bmatrix} y_1^{(\gamma_1)} & y_2^{(\gamma_2)} & \ldots & y_m^{(\gamma_m)} \end{bmatrix}^T = [v_1 \quad v_2 \quad \cdots \quad v_m]^T \tag{7}$$

Once linearization has been achieved, any further control objectives may be easily met [22], [23].

Furthermore, if the system has the total relative degree $\gamma = \gamma_1 + \gamma_2 + \cdots + \gamma_m < n$, the standard MIMO form for system states can be reformulated as:

$$\left.\begin{aligned} z_1^1 = h_1(x), z_2^1 = L_f h_1(x), \ldots, z_{\gamma_1}^1 = L_f^{\gamma_1 - 1} h_1(x) \\ z_1^2 = h_2(x), z_2^2 = L_f h_2(x), \ldots, z_{\gamma_2}^2 = L_f^{\gamma_2 - 1} h_2(x) \\ \vdots \\ z_1^m = h_m(x), z_2^m = L_f h_m(x), \ldots, z_{\gamma_m}^m = L_f^{\gamma_m - 1} h_m(x) \end{aligned}\right\} \tag{8}$$

where $z$ are the chosen transformed states for linearization and decoupling. The system in (8) can be converted into a pseudo-linearized system as:

$$\left.\begin{aligned} \dot{z}_1^1 &= z_2^1 \\ &\vdots \\ \dot{z}_{\gamma_1}^1 &= a_1(z, v) + \textstyle\sum_{j=1}^{m} b_j^1(z, v) u_j \\ &\vdots \\ \dot{z}_1^m &= z_2^m \\ &\vdots \\ \dot{z}_{\gamma_m}^m &= a_m(z, v) + \textstyle\sum_{j=1}^{m} b_j^m(z, v) u_j \\ v &= P(z, v) + Q(z, v)u \\ y_1 &= z_1^1 \\ &\vdots \\ y_m &= z_1^m \end{aligned}\right\} \tag{9}$$

where

$$\left.\begin{aligned} a_i(z, v) &= L_f^{\gamma_i} h_i \cdot \Phi^{-1}(z, v) \\ b_j^i(z, v) &= L_{g_i} L_f^{\gamma_i - 1} h_i \cdot \Phi^{-1}(z, v) \\ p_i(z, v) &= L_f \eta_i \cdot \Phi^{-1}(z, v) \\ q_{ij}(z, v) &= L_{g_i} \eta_i \cdot \Phi^{-1}(z, v) \end{aligned}\right\} \tag{10}$$

Note that $\in R^{n-\gamma}$ , $Q \in R^{(n-\gamma) \times m}$ and the aforesaid state transform $\Phi: x \rightarrow (z, v)$ is a diffeomorphism that maps $x$ onto standard coordinates. The transformed input vector $v$ has the expression from (9) as:

$$v = \Phi(z, v, u) \tag{11}$$

Once the linearization has been achieved, any further control objectives may be easily met.

### B. Variable Structure Control and Robustness Analysis

As mentioned above, the nonlinear terms of the system can be eliminated by selecting an appropriate set of input transformations. However, the input-output linearization only fits in with systems with accurate models. In order to ensure control system robustness in the presence of system uncertainties, such as parameter uncertainties or unmodelled dynamics, the sliding mode control based on variable structure theory is chosen and applied to the linearized system.

Taking into account the presence of uncertainties in the nonlinear system (1), and (7) becomes:

$$\begin{bmatrix} y_1^{(\gamma_1)} & y_2^{(\gamma_2)} & \ldots & y_m^{(\gamma_m)} \end{bmatrix}^T = (A + \Delta A) + (B + \Delta B)u \tag{12}$$

where the uncertainties represented by $||\Delta B||$ and $||\Delta A||$ are bounded. The switching surface is chosen as [24]:

$$s_i = \left(\frac{d}{dt} + \lambda_i\right)^{\gamma_i} \left(y_i - y_{iref}\right) \tag{13}$$

where $\lambda_i$ is a positive constant and $y_{iref}$ the reference command signal. Differentiating (13) with respect to time $t$ leads to:

$$\dot{s}_i = y_i{}^{\gamma_i} + \sum_{j=0}^{\gamma_i - 1} k_{ij} \left(y_i^j - y_{iref}^j\right) - y_{iref}^{\gamma_i} \tag{14}$$

In order to design a robust controller, the exponential approaching law is selected for SMC. The control input can then be expressed in the following form [25]:

$$u = B^{-1}(x)\{Y_{ref} - A - KY - \varepsilon \operatorname{sgn}(s)\} \tag{15}$$

$$KY = \begin{bmatrix} \sum_{j=0}^{\gamma_1 -1}\left[k_{1j}\left(y_1^j - y_{1ref}^j\right)\right] \\ \sum_{j=0}^{\gamma_2 -1}\left[k_{2j}\left(y_2^j - y_{2ref}^j\right)\right] \\ \vdots \\ \sum_{j=0}^{\gamma_m -1}\left[k_{mj}\left(y_m^j - y_{mref}^j\right)\right] \end{bmatrix} \tag{16}$$

Substituting $u$ and (12) into (14) yields to:

$$\dot{s} = \Delta A - \varepsilon sgn(s) -$$
$$\Delta B(x)B^{-1}(x)\left[KY + \varepsilon sgn(s) + A - Y_{ref}\right] \tag{17}$$

where $Y_{ref} = \left[y_{1ref}^{\gamma_1} \quad y_{2ref}^{\gamma_2} \quad \cdots \quad y_{mref}^{\gamma_m}\right]^T$.

In order to guarantee asymptotic stability of the control system, $s^T\dot{s} < 0$ must be ensured yields to:

$$s^T\dot{s} \leq \|s\|[\|\varDelta A\| - \varepsilon + \|\varDelta BB^{-1}KY\| + \varepsilon\|\varDelta BB^{-1}\| + \|\varDelta BB^{-1}A\| + \|\varDelta BB^{-1}Y_d\|] \tag{18}$$

If $\|\varDelta B\| < \|B\|$, and satisfy $\varepsilon > (\|\varDelta A\| + \|\varDelta BB^{-1}KY\| + \|\varDelta BB^{-1}A\| + \left\|\varDelta BB^{-1}Y_{ref}\right\|)/(1 - \|\varDelta BB^{-1}\|)$, then $s^T\dot{s} < 0$ can be guaranteed. This means that the reaching condition of the sliding mode is tenable and the desired sliding motion is reachable by means of a suitable control law $u$ [26], [27].

### *C. Phase Modulation Analysis*

A proper selection of parameters that satisfy the above-cited conditions can ensure stability and robustness of the control system, although this is a sufficient but not necessary condition for reachability of the sliding mode. Moreover, the system uncertainties come from the unmodelled information and linearization error make the transformed system, even a simple one, more complicated to obtain desirable results.

Phase plane portraits are traditionally used to graphically show the SMC working performance [28], [29]. Recent research considers it as a modulation strategy to instructively adjust system structure and parameters for ensuring the system robustness [19], [20]. The basic principle of phase modulation is introduced in Fig. 1. The system real-time response error $e$ and its derivative form $\dot{e}$ are used to establishing the coordinates. The switching line $s = 0$ defined by VSC theory divides the phase plane into four regions dominated by negative and positive feedback, respectively. Once the control system reach the designed sliding surface, the phase trajectory will frequently cross the switching line to repeatedly enter the different regions for compensating system uncertainties through switching between negative and positive feedback with high-frequency [30].

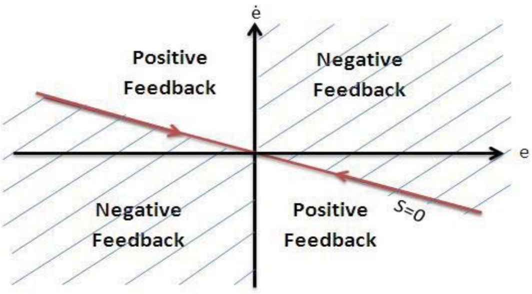


Figure 1. Phase plane with switching line

## III. NONLINEAR MACHAN AIRCRAFT

The aircraft chosen is a UAV (or remotely piloted vehicle), the Machan, used as a development vehicle by Marconi Avionics, RAE Farnbrough and NASA Dryden for research on high incidence flight and non-linear control laws [31]. The methodology has wider application for FTC for systems with known nonlinear dynamic structure.

The Machan Euler equations relate the forces $X$, $Y$, $Z$ and moments $L$, $M$, $N$ in the aircraft body axes to the angular and linear velocities in the inertial axes are shown as:

$$\left.\begin{aligned} m(\dot{u} + qw - rv) &= X \\ m(\dot{v} + ur - pw) &= Y \\ m(\dot{w} + vp - qu) &= Z \\ I_x\dot{p} + (I_z - I_y)rq &= L \\ I_y\dot{q} + (I_x - I_z)pr &= M \\ I_z\dot{r} + (I_y - I_z)pq &= N \end{aligned}\right\} \tag{19}$$

where, $m$ is the mass of the aircraft; $I_x$, $I_y$, $I_z$ are the moments of inertia about the axes through the centre of gravity parallel to the aircraft body axes; $u$, $v$ and $w$ are the forward, side and vertical velocity of the aircraft respectively; $p$, $q$ and $r$ are the roll, pitch and yaw rates, respectively.

The aerodynamic force and moment equations are:

$$\left.\begin{gathered} X = X_E - Dcos\alpha + (L_w + L_T)sin\alpha - mgsin\theta \\ Y = Y_a + mgcos\theta sin\phi \\ Z = -(L_W + L_T)cos\alpha - Dsin\alpha + mgcos\theta cos\phi \\ L = L_a + L_E \\ M = M_a + L_w(cg - 0.25)\bar{c} - L_q(l_t + 0.25 - cg) \\ N = N_a \end{gathered}\right\} \tag{20}$$

where $\alpha$ (degrees) is the angle of attack; $\theta$ and $\phi$ (degree) are the pitch and roll angles, respectively; $Y_a$ (N) is the side force; $cg$ (m) is the position of the aircraft centre of gravity; $X_E$ (N) is the thrust force due to the engine; $l_t$ (N·m$^{-1}$) is the tail moment; $D$ (N) is the force acting on the airframe; $L_w$, $L_T$ and $L_q$ (N) represent the wing, total tail and tail lift due to the pitch rate respectively; $M_a$, $N_a$ and $L_a$ (N·m$^{-1}$) are the pitching, yawing and rolling moment components respectively; $\bar{c}$ (m) is the mean aerodynamic chord and $L_E$ (N·m$^{-1}$) is the rolling moment due to the engine.

The first order non-linear engine dynamic is given as:

$$\dot{X}_E = (P_{max}T_H\delta_p - X_EU_2)/K_e \tag{21}$$

where, $P_{max}$, $T_H$, $\delta_p$, $K_e$ and $U_2$ represent the maximum engine power, the throttle demand, the propeller efficiency, the engine rise rate and the air flow rate, respectively. The parameter details are given in Aslin 1985. The open-loop Machan UAV is unstable, thus a closed-loop "base-line" control system must be configured for stability before the further robust or FTC system can be developed.

To simplify the system, this paper only considers the angle states and their rates. Thus the system state vector $x$, and the output state vector $y$ for the nonlinear aircraft model are chosen as:

$$x = [\phi \quad \theta \quad \psi \quad p \quad q \quad r]^T$$
$$y = [\phi \quad \theta \quad \psi]^T \tag{22}$$

where $\psi$ is yaw angle.

## *A. State-space Description*

The states $p$, $q$ and $r$ in (19) can be expressed as:

$$\left.\begin{aligned} \dot{p} &= \frac{(I_y - I_z)}{I_x} rq + \frac{L}{I_x} \\ \dot{q} &= \frac{(I_z - I_x)}{I_y} pr + \frac{M}{I_y} \\ \dot{r} &= \frac{(I_x - I_y)}{I_z} pq + \frac{N}{I_z} \end{aligned}\right\} \tag{23}$$

Additionally, the roll, pitch and yaw angles $\phi$, $\theta$and $\psi$ can be expressed in the terms of $p$, $q$ and $r$ as:

$$\left.\begin{aligned} \dot{\phi} &= p + qsin\phi tan\theta + r\, cos\phi tan\theta \\ \dot{\theta} &= qcos\phi - r\, sin\phi \\ \dot{\psi} &= qsin\phi sec\theta + rcos\phi sec\theta \end{aligned}\right\} \tag{24}$$

The real input vectors for the aircraft system are as:

$$u_{real} = [\delta_{\mathrm{a}} \quad \delta_{\mathrm{e}} \quad \delta_{\mathrm{t}}]^{\mathrm{T}} \tag{25}$$

where $\delta_{\mathrm{a}}$, $\delta_{\mathrm{e}}$, $\delta_{\mathrm{t}}$ are the aileron, elevator and rudder input, respectively. By only considering the nonlinear part of system, the input vectors $u$ of feedback linearization and decoupling control issue in (1) is chosen as described in (23):

$$u = [u_1 \quad u_2 \quad u_3]^T = \begin{bmatrix} \frac{L}{I_x} & \frac{M}{I_y} & \frac{N}{I_z} \end{bmatrix}^T \tag{26}$$

Then $u_{real}$ can be easily calculated in the terms of $u$ since they are linearly related from the form of the forces and moments equations [31], which can be shown as:

$$u_{real} = [l_{\delta_{\mathrm{a}}} \quad l_{\delta_{\mathrm{e}}} \quad l_{\delta_{\mathrm{r}}}]^T u \tag{27}$$

where $l_{\delta_{\mathrm{a}}}$, $l_{\delta_{\mathrm{e}}}$, $l_{\delta_{\mathrm{r}}}$are all the linear parameters.

Thus the affine system in (1) for this nonlinear UAV is simplified as:

$$f(x) = \begin{bmatrix} p + qsin\phi tan\theta + r\, cos\phi tan\theta \\ qcos\phi - r\, sin\phi \\ qsin\phi sec\theta + rcos\phi sec\theta \\ \frac{(I_y - I_z)}{I_x} rq \\ \frac{(I_z - I_x)}{I_y} pr \\ \frac{(I_x - I_y)}{I_z} pq \end{bmatrix}$$
$$G(x) = \begin{bmatrix} 0_{3\times3} \\ I_{3\times3} \end{bmatrix}$$
$$H(x) = [I_{3\times3} \quad 0_{3\times3}] \tag{28}$$

## *B. System Feedback Linearization and Decoupling*

For achieving linearization and decoupling, the pseudo-linearized system states in (9) for transforming the nonlinear system in (28) are chosen as [32]:

$$\left.\begin{aligned} z_1^1 &= \phi \\ z_2^1 = \dot{z}_1^1 &= p + qsin\phi tan\theta + r\, cos\phi tan\theta \\ z_1^2 &= \theta \\ z_2^2 = \dot{z}_1^2 &= qcos\phi - r\, sin\phi \\ z_1^3 &= \psi \\ z_2^3 = \dot{z}_1^3 &= qsin\phi sec\theta + rcos\phi sec\theta \end{aligned}\right\} \tag{29}$$

To check whether this coordinate transformation is invertible, the Jacobi Matrix for $z = z(x)$ is organized as:

$$\nabla z = \frac{\partial z(x)}{dx} =$$
$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ (qcos\phi - rsin\phi)tan\theta & (qsin\phi + rcos\phi)sec^2\theta & 0 & 1 & sin\phi tan\theta & cos\phi tan\theta \\ 0 & 1 & 0 & 0 & 0 & 0 \\ -qsin\phi - rcos\phi & 0 & 0 & 0 & cos\phi & -sin\phi \\ 0 & 0 & 1 & 0 & 0 & 0 \\ (qcos\phi - rsin\phi)sec\theta & (qsin\phi + rcos\phi)sec\theta tan\theta & 0 & 0 & sin\phi sec\theta & cos\phi sec\theta \end{bmatrix} \tag{30}$$

where $z = [z_1^1 \quad z_2^1 \quad z_1^2 \quad z_2^2 \quad z_1^3 \quad z_2^3]^T$.

The rank of $\nabla z$ is $det(\nabla z) = -sec\theta \neq 0$and $z = z(x)$ is a sufficiently smooth vector field with inversion, thus $z = z(x)$ is a global diffeomorphism of the system in (28) [9].

The input transformation $\nu$ in (6) takes the form:

$$\nu = P + Qu \tag{31}$$

where $\nu = [\nu_1 \quad \nu_2 \quad \nu_3]^T$ , $P = [P_1 \quad P_2 \quad P_3]^T$ , $Q = [Q_1 \quad Q_2 \quad Q_3]^T$ follow the description in (9), (10) as:

$$P(\bar{z}) = \frac{\partial z(x)}{dx} f(x)_{|x=f^{-1}(z)} \tag{32}$$

$$Q(\bar{z}) = \frac{\partial z(x)}{dx} G(x)_{|x=f^{-1}(z)} \tag{33}$$

Now, by choosing $\bar{z} = [z_1^1 \quad z_1^2 \quad z_1^3]^T$ to be the three output channels for the transformed system, each of these channels can be expressed as:

$$\begin{aligned} P_1 = (qcos\phi - rsin\phi)[(p &+ qsin\phi tan\theta \\ &+ r\, cos\phi tan\theta)tan\theta \\ &+ (qsin\phi + rcos\phi)sec^2\theta] + (I_y - I_z)rq/I_x \\ &+ (I_z - I_x)prsin\phi tan\theta/I_y \\ &+ (I_x - I_y)pqcos\phi tan\theta/I_z \end{aligned}$$

$$\begin{aligned} P_2 = (-qsin\phi - rcos\phi)&(p + qsin\phi tan\theta + r\, cos\phi tan\theta) \\ &+ (I_z - I_x)prcos\phi/I_y \\ &- (I_x - I_y)pq\, sin\phi/I_z \end{aligned}$$

$$\begin{aligned} P_3 = (qcos\phi - rsin\phi)[sec\theta&(p + 2qsin\phi tan\theta \\ &+ 2r\, cos\phi tan\theta)] \\ &+ (I_z - I_x)prsin\phi sec\theta/I_y \\ &+ (I_x - I_y)pqcos\phi sec\theta/I_z \end{aligned}$$

$$Q_1 = [1 \quad sin\phi tan\theta \quad cos\phi tan\theta]^T$$
$$Q_2 = [0 \quad cos\phi \quad -sin\phi]^T$$
$$Q_3 = [0 \quad sin\phi sec\theta \quad cos\phi sec\theta]^T \tag{34}$$

After input-output feedback linearization for the system, the nonlinear aircraft dynamics have been theoretically decoupled into three 2nd order linear sub-systems with the transformed states and outputs described in (3) as:

$$\left.\begin{aligned} \dot{z} &= A_0 z + B_0 v \\ y &= C_0 z \end{aligned}\right\} \tag{35}$$

where

$$A_0 = \begin{bmatrix} A_0^1 & & \\ & A_0^2 & \\ & & A_0^3 \end{bmatrix}, A_0^i = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} (i = 1,2,3)$$

$$B_0 = \begin{bmatrix} B_0^1 & & \\ & B_0^2 & \\ & & B_0^3 \end{bmatrix}, B_0^i = \begin{bmatrix} 0 \\ 1 \end{bmatrix} (i = 1,2,3)$$

$$C_0 = \begin{bmatrix} C_0^1 & & \\ & C_0^2 & \\ & & C_0^3 \end{bmatrix}, C_0^i = [1 \quad 0] (i = 1,2,3) \tag{36}$$

Each of the three sub-systems is a single-input single-output (SISO) 2nd order linear system in controllable (or phase variable) canonical form and is hence suitable for phase portrait analysis. For the 2nd order system in phase variable canonical form, the quality of the sliding mode invariance properties are satisfied, which means that once the states reach the sliding surface, the system dynamic performance is critically decided by the parameters of the designed SMC system [29].

### C. *Sliding Mode Controller Design*

For the Machan system, the thrust input $\delta_t$ related to all the states is set as a very small constant value $T_c$to limit it's effect on the nonlinear aircraft dynamics, which could be modeled as system uncertainty. The SMC theory is used to achieve sub-controllers for each decoupled channel. The complete control system scheme is shown as Fig. 2.

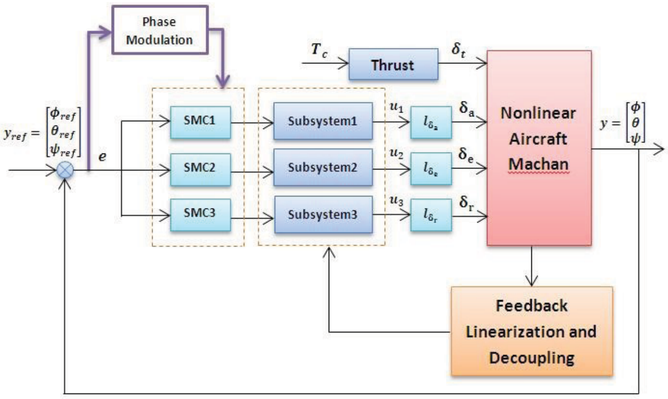

Phase Modulation
$T_c$
Thrust
$\delta_t$
$y_{ref} = \begin{bmatrix} \phi_{ref} \\ \theta_{ref} \\ \psi_{ref} \end{bmatrix}$
$e$
SMC1
SMC2
SMC3
Subsystem1
Subsystem2
Subsystem3
$u_1$
$u_2$
$u_3$
$l_{\delta_a}$
$l_{\delta_e}$
$l_{\delta_r}$
$\delta_a$
$\delta_e$
$\delta_r$
Nonlinear Aircraft Machan
$y = \begin{bmatrix} \phi \\ \theta \\ \psi \end{bmatrix}$
Feedback Linearization and Decoupling

Figure 2. Control system scheme

For the 2nd order subsystems, the switching surface can be derived from (13) as:

$$s_i = k_i e_i + \dot{e}_i (i = 1,2,3) \tag{37}$$

where $k_i$ is the adjustable parameter for SMC; $e_i = y_{ref} - y_i$ is system output error; $y_{ref}$ and $y_i$ are system reference input and real output, respectively.

The time derivative of (37) is then given as:

$$\dot{s}_i = k_i \dot{e}_i + \ddot{e}_i = k_i \left( y_{ref} - z_i^0 \right) + \left( \dot{y}_{ref} - z_i^1 \right) (i = 1,2,3) \tag{38}$$

By choosing a proper $k_i$ for each subsystem, the sliding surface would take on desired characteristics. The SMC approaching law used in this system has the proportional form:

$$\dot{s}_i = -\varepsilon_i \mathrm{sgn} s_i (i = 1,2,3) \tag{39}$$

where $\varepsilon_i$ is positive constant. From (37)-(39), the subsystem control inputs $v_i (i = 1,2,3)$ derived from SMC are as:

$$v_i = k_i \dot{e}_i + \varepsilon_i \mathrm{sgn} s_i \tag{40}$$

The above system with discontinuous control is termed a VSC since the effect of the switching surface is to alter the system feedback structure. The state trajectories on either sides of the surface $s_i = 0$ will remain in the vicinity of the sliding manifold since $s_i \dot{s}_i < 0$ on this surface. Once $v_i$ is obtained, the system chosen input vector $u$ and real input $u_{real}$ could be easily calculated from (31) and (27).

## IV. SIMULATION AND ANALYSIS

The SMC parameters are chosen as $K = [k_1, k_2, k_3]^T = [10,10,4]^T$, $\varepsilon = [\varepsilon_1, \varepsilon_2, \varepsilon_3]^T = [3,3,10.6]^T$.

The system responses and their phase trajectories shown in Fig. 3 appear to be stable but the phase trajectories for the states $\phi$ and $\psi$ are tangential to the switching line instead of crossing it, which indicates that the controllers are of dubious value without robustness. The reason for this phenomenon is that the linearized system may have conjugate poles too close to the imaginary axis, which makes the system readily unstable through the high controller gain action.

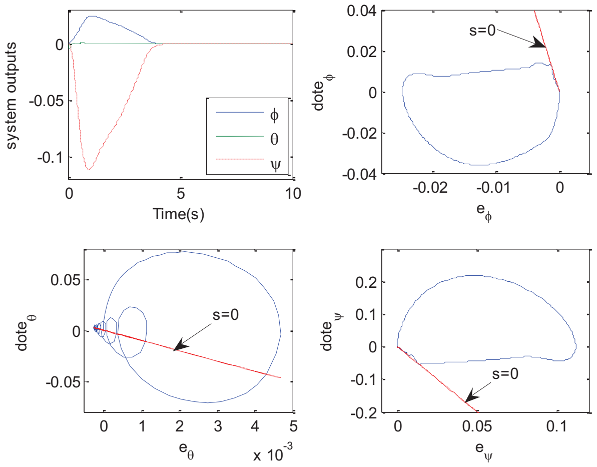


Figure 3. System responses and coordinating phase portraits

The solution is to add integrators for the corresponding states to replace system poles [30]. The new SMC parameters $K$ are chosen as: $[k_1, k_2, k_3]^T = [8,15,5]^T$, $\varepsilon$ is the same one as above. The developed system responses and their phase trajectories are shown in Fig. 4. The trajectories in Fig. 4 satisfy the principle of phase criterion for SMC described in Section II. For this case, the actuator faults $F$ are chosen as step signals $F = \left[ f_{\delta_a} \; f_{\delta_e} \; f_{\delta_r} \right]^T = [0.3,0.1,0.3]^T$ acting at time $t = 2s$. The system responses shown in Fig. 5 demonstrate the control scheme robustness and the efficiency of the modulation strategy.

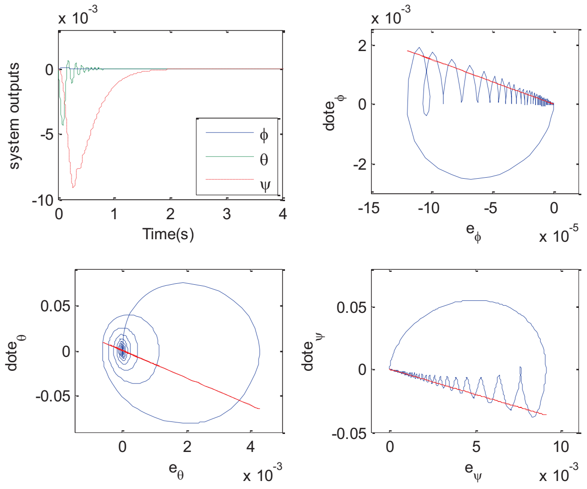


Figure 4. System responses and coordinating phase portraits

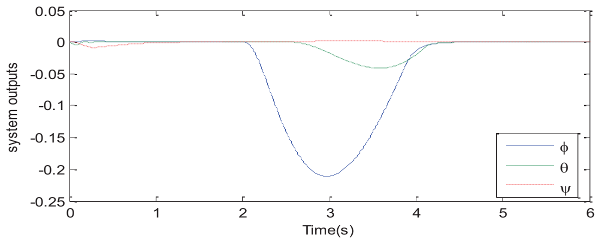


Figure 5. System responses with faults

## V. CONCLUSION

The design of a passive FTC scheme for the highly nonlinear dynamics of a UAV, the Machan is achieved via VSC theory based on feedback linearization. The inner stable control loop is designed for on-line linearizing and decoupling the nonlinear system. The principle of phase plane analysis is outlined for observing the system robustness. The further SMC sub-systems are developed using the phase modulation principle to guarantee robust performance and robust stability. The designed system responses illustrate that this strategy is feasible valid and a very promising passive approach to robust FTC for flight systems.

## REFERENCES

[1] R. J. Patton, “Fault-tolerant control systems: The 1997 situation”, IFAC Symp. Safeprocess’97, Hull, UK, pp.1029-1052, 1997.

[2] J. Boskovic, S. Li, and K. Raman, “Reconfigurable flight control design using multiple switching controllers and on-line estimation of damage-related parameters”, IEEE Int. Confer. on Contr. Appl., Sep 2000.

[3] Y. Wang, L. Xie, and C. E. de Souza, “Robust control of a class of uncertain nonlinear systems”, system&control letters, vol.19, August 1992, pp. 139-149.

[4] M. Bodson and J. E. Groszkiewicz, “Multivariable adaptive algotithms for reconfigureable flight control”, IEEE Trans. Contr. Syst. Technol., vol. 5, pp. 217-229, March 1997.

[5] M. Blanke, “Fault-tolerant control systems”, pp.171-196, Springer-Verlag, Denmark, 1999.

[6] J. Chen, and R. J. Patton, “Robust model-based fault diagnosis for dynamic systems”, Kluwer Academic, 1999.

[7] J. Doyle, K. Zhou, K. Glover and B. Bodenheimer, “Mixed $H_2$-H∞ performance objectives: robust performance analysis”, IEEE Trans. on Autom. Contr., vol. 39, pp. 1564-1575, 1994.

[8] M. Donald, “Automatic flight control systems”, Prentice Hall, 1990.

[9] S. H. Lane and R. F. Stengel, “Flight control design using non-linear inverse dynamics”, Automatica, vol. 24, pp. 471-483, 1988.

[10] Y. Ochi and K. Kanai, “Design of restructurable flight control systems using feedback linearization”, J. of Guid., Contr. & Dyn.,vol. 14, 1991.

[11] S. Snell, D. Enns and W. Garrard, “Nonlinear inversion flight control for a super-manoeuvrable aircraft”, J. of Guid. Contr. & Dyn, vol. 15, 1992.

[12] Y. Wu, and Q. Zou, “Robust Inversion-based 2-DOF control design for output tracking: piezoelectric-actuator example”, IEEE Trans. Contr. Syst. Technol., vol. 17, pp. 1069– 1082, 2009.

[13] N. Hovakimyan, E. Lavretsky and C. Cao, “Dynamic inversion of multi-input non-affine systems via time-scale separation”, American Control Conference, vol. 14, 2006.

[14] D. Young, U. Ozguner and V. Utkin, “A control engineering guide to sliding mode control”, IEEE Trans. on Contr. Syst. Technol., vol. 7, pp. 328-342, 1999.

[15] X. Yan, and C. Edwards, “Nonlinear robust fault reconsturctionh and eatimation using a sliding mode observer”, Automatica, vol. 43, Sep 2007.

[16] C. Edwards and S. K. Spurgeon, “Sliding mode control: theory and application”, Taylor & Francis, London, 1998.

[17] V. I. Utkin, J. Guldner and J. Shi, “Sliding mode control in electromechanical systems”, Taylor & Francis, London, 1999.

[18] G. Bartolini, A. Ferrara and E. Usani, “Chattering avoidance by second-order sliding mode control”, IEEE Trans. on Autom. Contr., vol.43, pp. 241-246, 1998.

[19] S. Ma, A. J. Wilkinson, and K. S. Paulson, “A phase modulation-based ultrasonic communication system using variable structure control”, IEEE ICCT, pp. 857-860, Nov 2010.

[20] S. Ma, A. J. Wilkinson and K. S. Paulson, “Performance analysis of a phase modulation-based ultrasonic receiver using variable structure control”, IEEE ICCT, pp.776-779, 2011.

[21] J.-J. E. Slotine, and W. Li, “Applied nonlinear control”, Prentice Hall, USA, 1991.

[22] Y. Tang, R. J. Patton, “Active FTC for nonlinear aircraft based on feedback linearization and robust estimation”, accepted for presentation at the IFAC Symposium Safeprocess 2012, Mexico City, August 2012.

[23] Y. Tang, R. J. Patton, “Fault tolerant flight control for UAV”, accepted for presentation at the Mediterranean conference, Barcelona, July 2012.

[24] Y. B. Shtessel and J. Buffington, “Multiple time scale flight control using reconfigurable sliding mode”, AIAA J. of Guid., Contr. and Dyn., vol. 22, pp.873-883, 1999.

[25] K. D. Young, “A control engineering’s guide to sliding mode control”, IEEE International Workshop on Variable Structure Systems.pp.1-14, Tokyo, Dec 1996.

[26] M. Allen, F. Barnelli-Zazzera and R. Scattolini, “Sliding mode control of large flexible space structure”, Control Engineering Practice, vol.8, pp.861-871, 2000.

[27] B. A. White and P. M. Silson, “Reachability in variable structure control systems”, IEE Pro. D Control Theory & Appl., vol. 131, pp.85-91, 1984.

[28] B. A. White, “Range-space dynamics of scalar-variable-structure control systems”, IEE Pro. D Control Theory & Appl., vol. 133, pp.35-42, 1986.

[29] A. S. I. Zinober, “Deterministic control of uncertain systems”, Peter Peregrinus, London, 1990.

[30] Ma, "Increasing the Capacity of an Ultrasonic Communication System using Variable Structure Control," PhD thesis, Department of Engineering, The University of Hull, UK, 2011.

[31] P. Aslin, “Aircraft simulation and robust flight control system design”, DPhil thesis, Department of Electronics, University of York, UK, 1985.

[32] G. Meyer, L. R. Hunt, “Application of nonliear transfomation to automatic flight control”, Automatica, Vol. 20, pp.103-107, 1984.

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# Frequency-Domain Tuning of Fixed-Structure Control Systems

Pascal Gahinet and Pierre Apkarian

*Abstract*— **This paper presents two new MATLAB-based tools for tuning fixed-structure linear control systems in the frequency domain: `hinfstruct` and `looptune`. These tools can directly tune control architectures with multiple feedback loops and multiple fixed-order, fixed-structure control elements. The controller parameters are tuned using non-smooth $H_\infty$ optimization but little a-priori knowledge of the $H_\infty$ theory is required. This makes such tools ideally suited for real-world applications where the control system structure and complexity are constrained. An application to helicopter control is discussed and the results of an extensive benchmark of `hinfstruct` vs. `hifoo` are reported.**

## I. Introduction

$H_\infty$ theory [1], [2], [3], [4] provides powerful techniques for synthesizing controllers in the frequency-domain. Typical design requirements such as speed of response, control bandwidth, disturbance rejection, and robust stability are naturally expressed as constraints on the gain ($H_\infty$ norm) of well-chosen closed-loop transfer functions. In turn, efficient algorithms and software tools are available to synthesize MIMO controllers that satisfy such gain constraints [1], [5], [6].

Yet existing $H_\infty$ synthesis tools have practical limitations that have slowed their adoption in industry. $H_\infty$ controllers are monolithic whereas most embedded control architectures are decentralized collections of simple control elements such as gains and PID controllers. $H_\infty$ controllers tend to be opaque and complex (high number of states) whereas embedded controllers tend to be intuitive and have low complexity. And recasting the design requirements as a single aggregate $H_\infty$ constraint can be challenging for engineers. As a result, hand tuning and optimization-based tuning tend to remain the norm for decentralized control systems.

This paper presents new MATLAB tools for *Structured $H_\infty$ Synthesis* [6] that overcome the limitations listed above. These tools leverage state-of-the-art nonsmooth optimizers [7], [8] to directly and efficiently tune arbitrary control architectures. By "arbitrary," we mean any single- or multiple-loop block diagram arrangement containing any number and type of linear control elements, from simple gains and PIDs to more complex notch filters and state-space controllers. Some of these tools also automate the $H_\infty$ formulation, allowing users to tune the controller elements directly from high-level specifications. Finally, despite the lack of convexity, these tools perform well in practice, both in terms of speed of execution and quality of the solutions.

The paper is organized as follows. Section 2 discusses the standard formulation of structured $H_\infty$ synthesis and the representation of tunable control elements. Section 3 presents `hinfstruct`, a general-purpose tool for structured $H_\infty$ synthesis. Section 4 presents `looptune`, a more specialized tool that automates mainstream tuning tasks. Finally, Section 5 gives an application example and Section 6 reports the results of a comparison with `hifoo`.

## II. Framework for Tuning Fixed-Structure Control Systems

As mentioned earlier, our starting point is any linear control architecture with one or more fixed-structure blocks to tune, for example the feedback structure shown in Figure 4 where the shaded blocks are tunable. Since there are infinitely many possible architectures, we use a formal representation called *Standard Form* that is both general and convenient to work with. The Standard Form is depicted in Figure 1 and consists of two main components:

- An LTI model $P(s)$ that combines all fixed (non tunable) blocks in the control system
- A structured controller $C(s) = \text{Diag}(C_1(s), \ldots, C_N(s))$ that combines all tunable control elements. Each control element $C_j(s)$ is assumed to be linear time invariant and to have some prescribed structure.

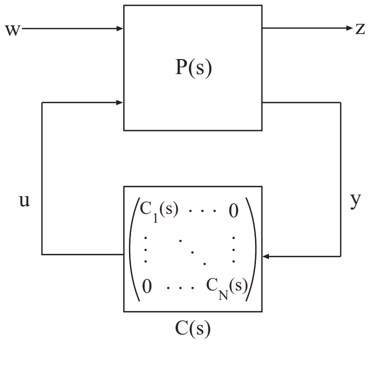

w
z
P(s)
u
y
$\begin{pmatrix} C_1(s) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & C_N(s) \end{pmatrix}$
C(s)

Fig. 1. Standard Form for Structured $H_\infty$ Synthesis

External inputs to the system such as reference signals and disturbances are gathered in $w$ and performance-related outputs such as error signals are gathered in $z$. The closed-loop transfer function from $w$ to $z$ is given by the linear-fractional transformation (LFT) [9]:

$$T_{zw}(s) = F_l(P, C) := P_{zw} + P_{zu}C(I - P_{yu}C)^{-1}P_{yw}. \quad (1)$$

P. Gahinet is with MathWorks, 3 Apple Hill, Natick, MA 01760-2098, USA `Pascal.Gahinet@mathworks.com`

Pierre Apkarian is with ONERA and Institut de Mathématiques, Université Paul Sabatier, 2, av. Ed. Belin, 31055, Toulouse, France `Pierre.Apkarian@onera.fr`

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

It is well known from Robust Control theory that any block diagram can be rearranged into this Standard Form by isolating the tunable blocks and collapsing the rest of the diagram into $P(s)$. The resulting model has the same structure as the uncertain models used in $\mu$ analysis (with the uncertainty blocks $\Delta_j(s)$ replaced by the control elements $C_j(s)$) [10], [11], [12]. Figure 1 is also reminiscent of standard $H_\infty$ synthesis [1] but differs in one key aspect, namely, the special structure of the controller $C(s)$. Since systematic procedures and automated tools are available to transform any architecture to the Standard Form of Figure 1, we henceforth assume that the control architecture is specified in this form.

The next challenge is to describe the tunable control elements. Again we are faced with a wide range of possible structures, from simple gains and PIDs to more complex lead-lag and observer-based controllers. Since our approach is based on optimization, it is natural to use parameterizations of such components. For example, a PID can be parameterized by four scalars $K_p, K_i, K_d, T_f$ as

$$C_j(s) = K_p + \frac{K_i}{s} + \frac{K_d s}{T_f s + 1}. \qquad (2)$$

Similarly, a state-space controller with fixed order $n$ can be parameterized by four matrices $A, B, C, D$ of suitable sizes. To parameterize more general elements like the lowpass and notch filters

$$\frac{a}{s+a}, \quad \frac{s^2 + 2\zeta_1\omega_0 s + \omega_0^2}{s^2 + 2\zeta_2\omega_0 s + \omega_0^2}, \qquad (3)$$

we introduce a basic building block called "real parameter" (`realp`). If $a$ is a real parameter, then any well-posed rational function $R(a)$ can be written as the LFT:

$$R(a) = F_l(M, a \otimes I) \qquad (4)$$

where $M$ is a fixed matrix and $a \otimes I := \mathrm{Diag}(a, \ldots, a)$ [13]. Since any interconnection of LFT models is an LFT model, it is easily seen that if the control element $C_j(s)$ is a rational function of $a$, the Standard Form of Figure 1 can be rearranged so that $C_j(s)$ is replaced by a block-diagonal matrix with copies of $a$ on the diagonal. This remains true when $C_j(s)$ depends on multiple parameters $a_1, \ldots, a_M$. In other words, we can absorb the specific structure of $C_j(s)$ into $P(s)$ and keep only the low-level tunable parameters $a_1, \ldots, a_M$ in the $C(s)$ block of Figure 1. Note that unlike $\mu$-analysis, "repeated" blocks do not affect the optimization outcome and only incur some small overhead.

For example, consider the lowpass filter $F(s) = \frac{a}{s+a}$ where $a$ is tunable. This tunable element is specified by:

```
a = realp('a',1);  % a initialized to 1
F = tf(a,[1 a]);   % creates a/(s+a)
```

This automatically builds the following LFT representation of $F(s)$:

$$F(s) = F_l\left(\left(\begin{array}{ccc} 0 & 0 & 1 \\ 1/s & -1/s & 0 \\ 1/s & -1/s & 0 \end{array}\right), \left(\begin{array}{cc} a & 0 \\ 0 & a \end{array}\right)\right). \qquad (5)$$

Note that while it is possible to parameterize $F(s)$ using a single copy of `a`, the convenience of the syntax shown above more than makes up for the small overhead incurred by the extra copies of $a$.

## III. Structured $H_\infty$ Synthesis

Now that we have a framework for describing arbitrary control architectures and linear control elements, we turn to the question of using $H_\infty$ synthesis to tune the controller parameters in the Standard Form of Figure 1. $H_\infty$ synthesis is a frequency-domain method for enforcing typical control design requirements. At the heart of the method is the $H_\infty$ norm, which measured the peak input/output gain of a given transfer function:

$$\|H(s)\|_\infty := \max_\omega \overline{\sigma}(H(j\omega)). \qquad (6)$$

In the SISO case, this norm is just the peak gain over frequency. In the MIMO case, it measures the peak 2-norm of the frequency response $H(j\omega)$ over frequency.

It is well-known from robust control theory [14] that classical design requirements (bandwidth, roll-off, disturbance attenuation, stability margins) can be recast as normalized $H_\infty$ constraints of the form

$$\|W_j(s)T_j(s)\|_\infty < 1, \quad j = 1, \ldots, M \qquad (7)$$

where the $T_j$'s are suitable closed-loop transfer functions and the $W_j$'s are weighting functions that reflect the nature and parameters of each requirement. So a typical controller tuning task consists of adjusting the controller parameters to satisfy the constraints (7). Introducing

$$H(s) := \mathrm{Diag}(W_1(s)T_1(s), \ldots, W_M(s)T_M(s)). \qquad (8)$$

(7) is equivalent to $\|H(s)\|_\infty < 1$. Since each transfer function $T_j$ can be expressed as an LFT model depending on the structured controller $C(s) := \mathrm{Diag}(C_1(s), \ldots, C_N(s))$, the Standard Form of $H(s)$ looks like

$$H(s) = F_l\left(P(s), \mathrm{Diag}(C(s), \ldots, C(s))\right) \qquad (9)$$

where $P(s)$ is a fixed LTI model. So *independently* constraining two or more closed-loop transfer functions $T_j(s)$ as in (7) leads to repeating the controller $C(s)$ multiple times in the Standard Form. The resulting block-diagonal controller structure is beyond the scope of standard $H_\infty$ algorithms but poses no problem in our framework since this merely amounts to repeating the tunable blocks along the diagonal (see Section II for a discussion of repeated blocks). This is an important advantage over traditional $H_\infty$ synthesis where all requirements must be expressed in terms of a *single* MIMO closed-loop transfer function $T(s)$.

Summing up, decentralized controller tuning can be recast as a structured $H_\infty$ synthesis problem where the controller has a block-diagonal structure, each block being parameterized and possibly repeated. In turn, we can use the nonsmooth algorithms described in [7], [15], [16] to optimize the controller parameters and enforce the constraint $\|H(s)\|_\infty < 1$. The MATLAB sofware described here is based on [7], [17] and consists of three main components:

- Simple objects to specify tunable parameters (realp) and elementary control elements such as gains, low-order transfer functions, and PIDs
- Overloaded arithmetic and interconnection algebra to automatically build the Standard Form of $H(s)$ by combining/connecting together ordinary LTI models, tunable elements, and weighting functions
- The hinfstruct function for minimizing the $H_\infty$ norm of $H(s)$ with respect to the tunable controller parameters. This function can be seen as the counterpart of hinfsyn for structured $H_\infty$ synthesis

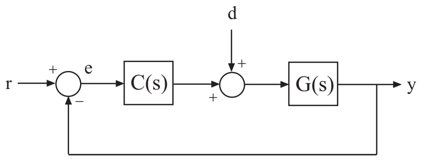


Fig. 2. Elementary feedback loop.

To get a feel for these tools, consider the simple scenario where the requirements for the feedback loop of Figure 2 can be expressed as

$$\|w_S S\|_\infty < 1, \quad \|w_T T\|_\infty < 1 \tag{10}$$

where $S = 1/(1 + L)$, $T = L/(1 + L)$, $L = GC$, and $w_S, w_T$ are suitable frequency-weighting functions. Also assume that $C(s)$ is constrained to be a PID controller. Using the sofware, you can construct a parametric model of $H(s) = \text{Diag}(w_S S, w_T T)$ as follows:

```
G = tf([1 2],[1 5 10]);       % plant model
C = ltiblock.pid('C','pid'); % tunable PID
S = feedback(1,G*C);
T = feedback(G*C,1);
H0 = blkdiag(wS * S, wT * T);
```

The result H0 is a MATLAB representation of the (untuned) Standard Form for $H(s)$ and depends on the tunable PID block C. Next invoke hinfstruct to tune the PID controller gains so as to enforce $\|H(s)\|_\infty < 1$:

```
H = hinfstruct(H0);
```

The output H contains the tuned Standard Form of $H(s)$ and you can access the tuned value of the PID controller C with

```
C = getBlockValue(H,'C')
```

Note that hinfstruct actually minimizes the $H_\infty$ norm of $H(s)$ but can be configured to terminate as soon as the target value of 1 is achieved. Also, hinfstruct can be configured to automatically run multiple optimizations from randomly generated starting points. This helps mitigate the local nature of the optimizer and increases the likelihood of finding parameter values that meet the design requirements. See [18], [6], [19] for more details and examples.

## IV. Automated Tuning of Feedback Loops

While hinfstruct addresses the first two practical limitations of traditional $H_\infty$ synthesis tools, familiarity with the $H_\infty$ methodology is still required to turn typical design specifications into a well-posed $H_\infty$ optimization problem. This difficulty is exacerbated in multi-loop control systems because of scaling and coupling issues. For example, a poor choice of units in one feedback channel may skew the sensitivity function and lead to an ill-posed $H_\infty$ problem [14, p. 5-8]. Also, classical one-loop-at-a-time stability margins may be misleading when cross-coupling exists between feedback loops [20]. Such challenges led us to seek ways to automate the $H_\infty$ formulation of high-level requirements. This turns out to be possible and to often lead to satisfactory results. We now discuss this "push-button" approach and the looptune function that embodies it.

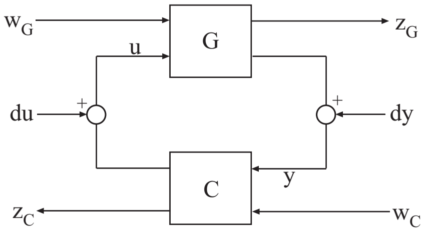


Fig. 3. Generic MIMO feedback loop.

Our starting point is the generic MIMO feedback loop of Figure 3 where $G$ represents the "plant" and $C$ represents the overall controller. Any control structure can be rearranged in this fashion by using the measurement signals $y$ and control signals $u$ to separate the controller from the plant. Both $G$ and $C$ may contain tunable elements, which allows for co-tuning of plant and controller parameters. To formulate an $H_\infty$ synthesis problem for this feedback system, observe that most controller tuning tasks involve some combination of the following requirements:

1) **Performance**: The feedback loops should have high gain at low frequency to reject disturbances and follow setpoint changes
2) **Roll off**: The feedback loops should have low gain at high frequency to guard against unmodeled dynamics and measurement noise
3) **Stability**: The feedback loops should be stable with enough margin to sustain typical amounts of gain and phase variations at the plant inputs and outputs.

Typically, the first two requirements amount to shaping the open-loop response to have integral action at low frequency and roll off in excess of -20 dB/decade at high frequency. The transition from high to low open-loop gain occurs in the *gain crossover band* (an interval in the MIMO case since in general it is neither possible nor desirable to make all loops cross at the same frequency). The gain crossover band determines the response time and bandwidth of the control system. Using the standard "mixed-sensitivity" formulation, we can express these loop-shaping requirements as

$$\left\| \begin{pmatrix} W_{LF} S_i \\ W_{HF}(I - S_i) \end{pmatrix} \right\|_\infty < 1 \tag{11}$$

where $S_i$ is the sensitivity function at the plant inputs $u$ and the weighting functions $W_{LF}, W_{HF}$ reflect the desired loop shape. Note that $S_i$ should be replaced by the sensitivity

$S_o$ at the plant outputs if there are more controls $u$ than measurements $y$.

For the stability requirement, we use the notion of multivariable disk margins discussed in [20]. This measure guarantees robustness against simultaneous gain and phase variations at all plant inputs and outputs, which is much stronger than one-loop-at-a-time stability margins. With the notation

$$L(s) = \begin{pmatrix} 0 & G(s) \\ C(s) & 0 \end{pmatrix}, \quad X(s) = (I+L)(I-L)^{-1}, \tag{12}$$

the robust stability condition is

$$\mu(X(s)) < 1/\alpha \tag{13}$$

where $\mu(.)$ denotes the structured singular value for a diagonal block structure [4] and the parameter $\alpha$ is a function of the desired gain and phase margins [20]. For tractability reasons, we replace this condition by:

$$\min_{D} \max_{\omega\in[\omega_1,\omega_2]} \|D^{-1}X(jw)D\| < 1/\alpha \tag{14}$$

where $D$ is a constant and diagonal scaling matrix and $[\omega_1, \omega_2]$ is some interval containing the gain crossover band. The rationale for this simplification is that (a) stability margins are worst near the gain crossovers, and (b) the gain crossover band is typically narrow enough that we can get away with a constant rather than frequency-dependent D-scaling in the $\mu(.)$ upper bound.

Note that the scaling $D = \begin{pmatrix} D_o & 0 \\ 0 & D_i \end{pmatrix}$ is equivalent to the plant I/O scaling $G \to D_o^{-1}GD_i$. In other words, $D$ automatically corrects scaling issues in the vector signals $u$ and $y$, e.g., $u$ having both small and large entries due to a poor choice of units. Because the $H_\infty$ norm is not invariant under I/O scaling, this turns out to be essential to formulate a meaningful $H_\infty$ synthesis problem [14, Remark 1]. Finally, (14) is tractable in our framework if we treat the diagonal entries of $D$ as tunable parameters ($D^{-1}X(jw)D$ is an LFT in the controller and scaling matrix $D$).

Summing up, for a given crossover frequency/band, tuning the controller amounts to finding a scaling $D$ and controller parameter values that satisfy

$$\left\| \begin{pmatrix} W_{LF}D_i^{-1}S_iD_i \\ W_{HF}(I - D_i^{-1}S_iD_i) \end{pmatrix} \right\|_\infty < 1 \tag{15}$$

$$\max_{\omega\in[\omega_1,\omega_2]} \|\alpha D^{-1}X(jw)D\| < 1. \tag{16}$$

Note that we use the scaled input sensitivity $D_i^{-1}S_iD_i$ instead of $S_i$ to take advantage of the $u$ scaling provided by $D$.

The `looptune` function [6] formulates and solves this $H_\infty$ optimization problem. The basic interface is

```
[G,C] = looptune(G0,C0,wc,Req1,Req2,...)
```

where `G0,C0` are (untuned) parametric models of $G$ and $C$, `wc` is the target crossover frequency/band, and the optional arguments `Req1,Req2,...` specify additional requirements such as maximum gain or setpoint tracking. Note that `wc` can be omitted and replaced by more sophisticated loop shaping requirements, thus providing a fair amount of flexibility.

## V. Helicopter Example

This section presents an application to a challenging helicopter control problem. We use an 8-state model of the Westland Lynx helicopter at the hovering trim condition. The controller generates commands $d_s, d_c, d_T$ in degrees for the longitudinal cyclic, lateral cyclic, and tail rotor collective using measurements of $\theta, \phi, p, q, r$ (pitch and roll angles and roll/pitch/yaw rates). For details and data, see [21] and the demo in [6]. The controller structure is shown in Figure 4 and consists of two feedback loops:

- The inner loop (static output feedback SOF) provides stability augmentation and decoupling
- The outer loop (PI controllers PI1-PI3) provides the desired setpoint tracking performance.

The main control objective is to track setpoint changes in $\theta, \phi, r$ with zero steady-state error, settling times of about 2 seconds, minimal overshoot, and minimal cross-coupling.

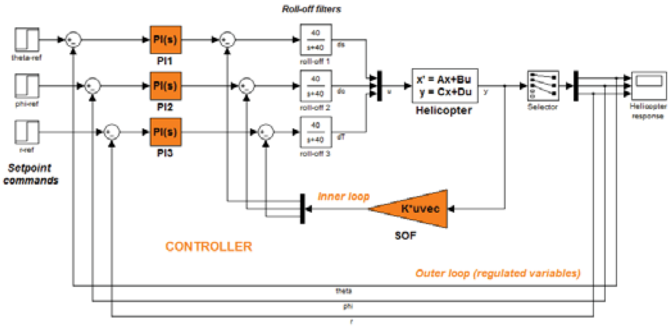


Fig. 4. Control architecture for Westland Lynx helicopter.

Since the helicopter is modeled in Simulink we use the `slTunable` interface to quickly set up the `looptune` optimization. With this interface you just specify the Simulink blocks to tune, the measurement and control signals (controller I/Os), and the I/O signals of interest for closed-loop analysis:

```
ST0 = slTunable('helico',...
                {'PI1','PI2','PI3','SOF'});
ST0.addControl('u')
ST0.addMeasurement('y')
ST0.addIO({'theta_ref','phi_ref','r_ref'},'in')
ST0.addIO({'theta','phi','r'},'out')
```

This information is used to automatically parameterize the tuned blocks and linearize the Simulink model to extract the plant model $G$ and a parametric model of the controller $C$. Note that the static-output-feedback gain is initialized to zero and the PI controllers to $1 + 1/s$, values for which the closed-loop response is unstable.

We want the outer loop to settle in about 2 seconds so the open-loop bandwith should be at least 2 rad/s (based on first-order characteristics). The inner loop must typically be faster so we seek a gain crossover band between 10 and 30 rad/s:

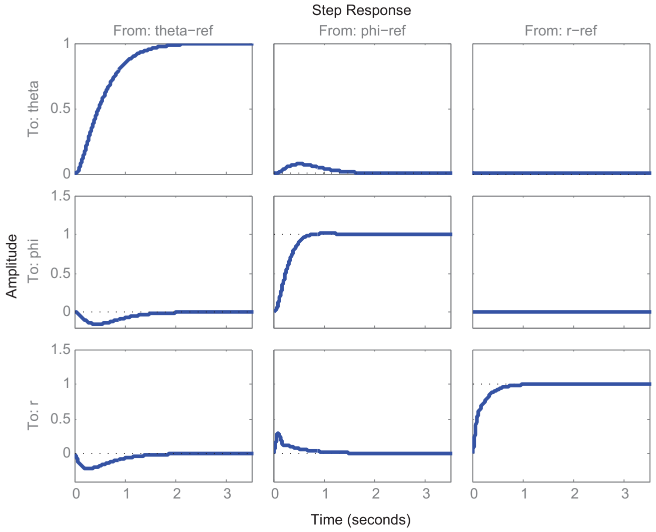


Fig. 5. Closed-loop responses to $\theta, \phi, r$ commands.

```
wc = [10,30];
```

Because there are fewer actuators (3) than measurements (5), integral action in the open-loop response is not enough to guarantee that $\theta, \phi, r$ will track the setpoint commands $\theta_{\text{ref}}, \phi_{\text{ref}}, r_{\text{ref}}$. We therefore add an explicit tracking requirement with a 2-second response time:

```
TR = TuningGoal.Tracking(...
      {'theta_ref','phi_ref','r_ref'},...
      {'theta','phi','r'},2);
```

Finally, we specify the desired minimum gain and phase margins:

```
Opt = looptuneOptions('GainMargin',5,...
                      'PhaseMargin',40);
```

We can now tune the controller parameters with `looptune`:

```
ST = ST0.looptune(wc,TR,Opt);
```

The final $H_\infty$ norm is 1.28, again close to 1, and the closed-loop step responses are shown in Figure 5. These responses settle in less than two seconds with no overshoot and small cross-coupling. The tuned values are:

$$\begin{aligned}
PI_1(s) &= 0.45 + 12.2/s \\
PI_2(s) &= -0.17 - 9.15/s \\
PI_3(s) &= -0.66 - 8.22/s \\
SOF &= \begin{pmatrix} 7.24 & -0.84 & -0.0031 & 1.02 & -0.045 \\ -1.30 & -2.45 & 0.01 & -0.14 & -0.21 \\ -1.43 & 0.81 & -1.47 & -0.23 & 0.17 \end{pmatrix}
\end{aligned}$$

Tuning these 21 parameters starting from an unstable initial guess took 11 seconds on a 64-bit PC with a 3 GHz dual-core processor and 6 GB of RAM.

## VI. BENCHMARK

This section reports the result of a comprehensive benchmark of `hinfstruct` vs. `hifoo`. Both `hifoo` and `hinfstruct` implement state-of-the-art nonsmooth programming techniques. `hifoo` is a two-stage technique where a smooth linesearch BFGS approach is followed by nonsmooth gradient sampling. This means gradients are randomized around the current iterate to refine or establish optimality in the second phase. `hinfstruct` exploits extension sets of the Clarke sub-differential at each iteration and derives a tangent subproblem in the form of a nonsmooth convex QP approximation of the original problem. A search direction is then computed and a linesearch is carried out. `hinfstruct` is fully deterministic and does not use randomization except for (optionally) choosing the starting point. Both techniques have local optimality certificates.

Our assessment is based on 234 test cases from the *COMP*l$_e$*ib* benchmark library [22] and compares the following versions:

- `hinfstruct` from [6]
- `hifoo 3.5` with `hanso 2.1` [15].

Details on the test cases can be found at [19]. Both codes are run in default mode with 3 starting points in each case. For `hifoo` we run the gradient sampling phase to enhance accuracy and provide an optimality certificate.

A graphical comparison of the achieved objective values and execution times for both techniques appears in Figure 6. Recall that the objective value is the closed-loop $H_\infty$ norm. The top plot shows a bar chart of the $H_\infty$-norm log-ratios:

$$\log_2(H_\infty\text{-norm } \texttt{hinfstruct}/H_\infty\text{-norm } \texttt{hifoo})\,.$$

Bars to the left of the vertical line $x = 0$ indicate that `hinfstruct` terminated with a lower $H_\infty$-norm than `hifoo`, and bars to the right indicate the opposite. A bar of unit length materializes improvement by a factor 2, a bar of length 2 improvement by a factor of 4, etc. Similarly, the bottom plot shows a bar chart of the CPU time log-ratios:

$$\log_{10}(\text{cpu time } \texttt{hinfstruct}/\text{cpu time } \texttt{hifoo})\,.$$

Here a bar of unit length means 10 times faster, a bar of length 2 means 100 times faster, etc.

Further comparison of CPU times for problems where `hinfstruct` and `hifoo` terminate with essentially the same objective value is shown in Figure 7. These results show that `hinfstruct` is significantly faster and often more accurate than `hifoo`.

## CONCLUSION

We have presented a new methodology and tool set for tuning fixed-structure SISO or MIMO control systems. While our approach is rooted in $H_\infty$ theory, it is clear that these tools and techniques are not restricted to Robust Control applications and can be seen as a systematic framework for tuning decentralized control architectures and exploring the trade-offs between performance and complexity.

## REFERENCES

[1] J. Doyle, K. Glover, P. P. Khargonekar, and B. A. Francis, "State-space solutions to standard $H_2$ and $H_\infty$ control problems," *IEEE Trans. Aut. Control*, vol. AC-34, no. 8, pp. 831–847, Aug. 1989.

[2] G. Stein and J. C. Doyle, "Beyond singular values and loop shapes," *J. Guidance and Control*, vol. 14, pp. 5–16, 1991.

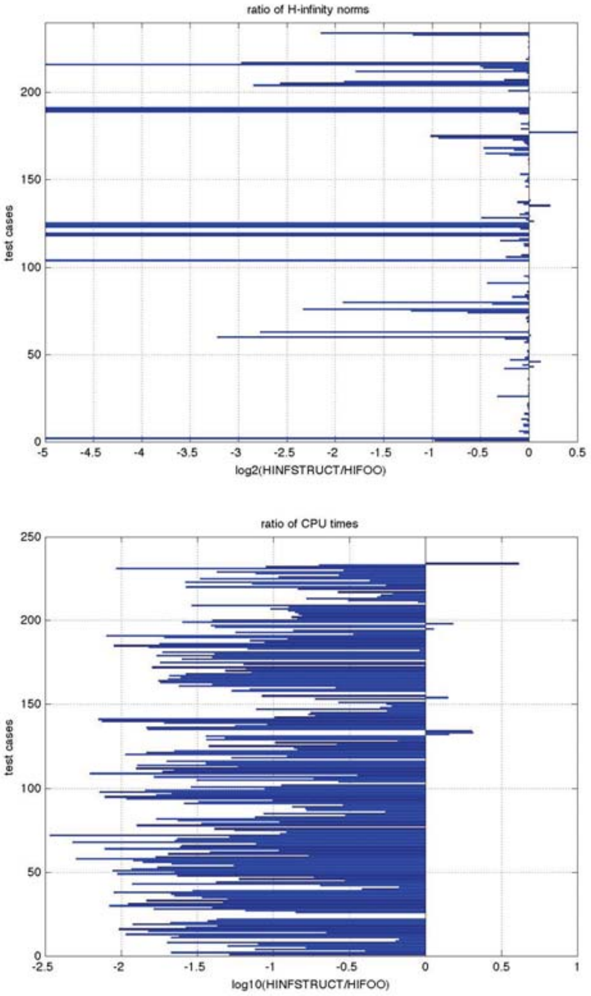


Fig. 6. Comparison of achieved objectives (top) and CPU times (bottom)

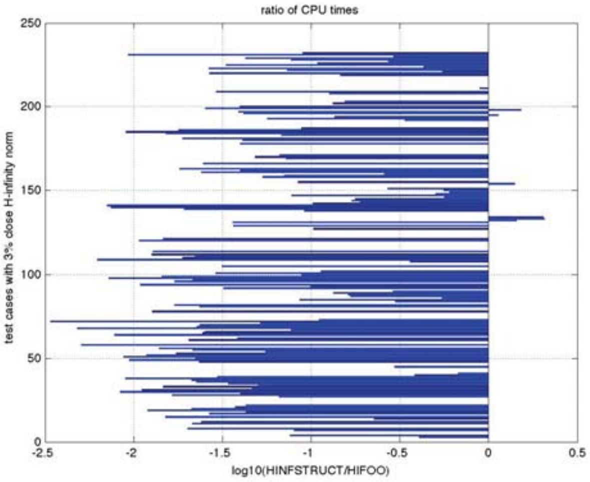


Fig. 7. Comparison of CPU times when final values differ by $3\%$ or less

[3] D. McFarlane and K. Glover, "A loop shaping design procedure using $H_\infty$ synthesis," *IEEE Trans. Aut. Control*, vol. 37, no. 6, pp. 759–769, 1992.

[4] K. Zhou, J. C. Doyle, and K. Glover, *Robust and Optimal Control*. Prentice Hall, 1996.

[5] P. Gahinet and P. Apkarian, "A linear matrix inequality approach to $\mathrm{H}_\infty$ control," *Int. J. Robust and Nonlinear Control*, vol. 4, pp. 421–448, 1994.

[6] *Robust Control Toolbox 4.1*. The MathWorks, Inc., Natick, MA, USA, March 2012.

[7] P. Apkarian and D. Noll, "Nonsmooth $H_\infty$ synthesis," *IEEE Trans. Aut. Control*, vol. 51, no. 1, pp. 71–86, 2006.

[8] P. Apkarian, V. Bompart, and D. Noll, "Nonsmooth structured control design with application to PID loop-shaping of a process," *Int. J. Robust and Nonlinear Control*, vol. 17, no. 14, pp. 1320–1342, 2007.

[9] R. M. Redheffer, "On a certain linear fractional transformation," *J. Math. and Phys.*, vol. 39, pp. 269–286, 1960.

[10] J. Doyle, A. Packard, and K. Zhou, "Review of LFT's, LMI's and $\mu$," in *Proc. IEEE Conf. on Decision and Control*, vol. 2, Brighton, Dec. 1991, pp. 1227–1232.

[11] A. Varga and G. Looye, "Symbolic and numerical software tools for LFT-based low order uncertainty modeling," in *Proc. CACSD'99 Symposium, Cohala*, 1999, pp. 1–6.

[12] A. Varga and J. Magni, "Enhanced LFR-toolbox for MATLAB," *Aerospace Science and Technology*, vol. 9, no. 2, pp. 173–180, 2005.

[13] C. J. Bett and M. Lemmon, "On linear fractional representations of multidimensional rational matrix functions," Eindhoven, University of Technology, Tech. Rep., 1997.

[14] S. Skogestad and I. Postlethwaite, *Multivariable Feedback Control - Analysis and Design*. Wiley, 1996.

[15] J. V. Burke, D. Henrion, A. S. Lewis, and M. L. Overton, "HIFOO - a MATLAB package for fixed-order controller design and $H_\infty$ optimization," in *5th IFAC Symposium on Robust Control Design*, Toulouse, France, July 2006.

[16] S. Gumussoy, M. Millstone, and M. L. Overton, "$H_\infty$ strong stabilization via HIFOO, a package for fixed-order controller design," in *Proc. IEEE Conf. on Decision and Control*, Cancun, Mexico, 2008, pp. 4135–4140.

[17] P. Apkarian and D. Noll, "Nonsmooth optimization for multiband frequency domain control design," *Automatica*, vol. 43, no. 4, pp. 724–731, April 2007.

[18] P. Gahinet and P. Apkarian, "Structured $H_\infty$ synthesis in MATLAB," in *Proc. IFAC*, Milan, Italy, Aug. 2011.

[19] P. Apkarian, "Internet pages," http://pierre.apkarian.free.fr, 2010.

[20] J. Blight, R. Dailey, and D. Gangsassi, "Practical control law design for aircraft using multivariable techniques," *International Journal of Control*, vol. 59, no. 1, pp. 93–137, 1994.

[21] C. Luo, R. Liu, C. Yiang, and Y. Chang, "$H_\infty$ control design with robust flying quality," *Aerospace Science & Technology*, vol. 7, pp. 159–169, 2003.

[22] F. Leibfritz, "COMP$L_e$IB, COnstraint Matrix-optimization Problem LIbrary - a collection of test examples for nonlinear semidefinite programs, control system design and related problems," Universität Trier, Tech. Rep., 2003.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Robust $H_\infty$ Control for a Class of Uncertain Nonlinear Switched Systems

Ben Niu*, Rui Wang †, Georgi M. Dimirovski‡ and Jun Zhao §
* College of Information Science and Engineering, Northeastern University, Shenyang, 110819, PR China
Email: niubengj@163.com; niubenzg@gmail.com
† School of Aeronautics and Astronautics, Dalian University of Technology, Dalian, 116024, PR China
Email: ruiwang@dlut.edu.cn
‡Department of Computer Engineering, Dogus University, Kadikoy, Istanbul, TR-34722, Turkey
Email: gdimirovski@dogus.edu.tr
§College of Information Science and Engineering, Northeastern University, Shenyang, 110819, PR China
Email: zhaojun@mail.neu.edu.cn

***Abstract*—This paper focuses on the robust $H_\infty$ control problem for a class of nonlinear switched systems containing neutral uncertainties with average dwell time (ADT). Uncertainties are assumed to be nonlinearly dependent on state and state derivative and allowed to appear in channels of state, control input and disturbance input. The robust $H_\infty$ control problem of the switched system with stabilizable and unstabilizable subsystems is solvable if the stabilizable and unstabilizable subsystems satisfy certain conditions and admissible switching strategy among them. ADT and piecewise Lyapunov function approaches are applied to achieve the control design. A numerical example is provided to illustrate the effectiveness of the proposed results.**

## I. INTRODUCTION

The last decades have witnessed a rapidly growing interest from the control field in the study of switched systems [1-7]. More specifically, switched systems belonging to a class of hybrid dynamical systems contain a finite number of subsystems and a switching signal that must be designed in order to orchestrate the switching among the subsystems. Recently, there is increasing growth of interest in applying ADT switching to handle the switched systems [8-10]. As a class of typical controlled switching signals, ADT switching means that the number of switches in a finite interval is bounded and the average time between the consecutive switching is not less than a specified value. It is widely recognized that ADT switching is of practical and theoretical significance to deal with the related stability analyses and control syntheses problems.

As is well known, uncertainties are unavoidable in engineering control and are frequently the source of instability and performance deterioration. Thus during the past decades, the problems of stability analysis and controller synthesis with uncertainties have received much attention [11-13]. [14] studied the robust stabilization problem for a class of uncertain nonlinear cascaded systems, in which the uncertain parameters are from a known compact set. In [15], the problem of robust $l_2$-$l_\infty$ filtering for switched linear discrete-time systems with polytopic uncertainties and time-varying delays is investigated. Furthermore, neutral uncertainties describing many practical parameter perturbations are often nonlinearly state and nonlinearly state derivative dependent. [16] discussed the robust $L_2$-gain performance synthesis problem for a class of nonlinear systems with neutral uncertainties. However, few results have focused on switched systems with neutral uncertainties so far.

On the other hand, $H_\infty$ control theory for switched systems has attracted considerable attention by researchers and has been a hot topic in the control area [17-21]. Especially, results about nonlinear $H_\infty$ control of switched systems have progressively appeared to solve robust stabilization and disturbance attenuation issues [22-25]. The nonlinear $H_\infty$ control problem for switched systems can be stated as follows: Find a compensator, either state feedback or more general output feedback and a switching rule (if necessary) such that (1) the internal state of the closed-loop system is stable and (2) the $L_2$ gain of the mapping from the exogenous input disturbance to the controlled output is minimized or guaranteed to be less than or equal to a prescribed value. In [26], the $H_\infty$ control problem of switched systems has been addressed with ADT in both linear and nonlinear contexts. [27] investigated the $H_\infty$ control problem for a class of switched nonlinear cascade systems using the multiple Lyapunov function method.

In this paper, we discuss the problem of robust $H_\infty$ control for a class of nonlinear switched systems with neutral uncertainties. For the case where states are measurable, sufficient conditions for the switched system to be asymptotically stable with $H_\infty$-norm bound and design of both switching law and state feedback controller are proposed for all admissible uncertainties. ADT switching is used so that the results cover the case where stabilizable and unstabilizable subsystems both exist in the switched system. An numerical example is given to illustrate the applicability of the developed method. As compared to the existing results, this paper deals with neutral uncertainties. Additionally, uncertainties are also allowed to appear in channels of state, control input and disturbance input.

*Notation:* we use standard notations throughout this paper. $R^n$ denotes the n-dimensional real Euclidean space, and given a matrix $P, P > 0$ denotes that $P$ is positive definite, $P^T$ stands for the transpose of $P$, $I$ is the identity matrix, $\| \cdot \|$

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

represents either the Euclidean vector norm or the induced matrix 2-norm, and $\overline{\sigma}(\cdot)$ denotes the largest singular value of a matrix.

## II. Problem Statement And Preliminaries

In this paper, we consider a class of nonlinear switched systems described by equations of the form:

$$\begin{aligned} \dot{x} + \Delta j_{\sigma(t)}(\dot{x}, t) =& f_{\sigma(t)}(x) + \Delta f_{\sigma(t)}(x, t) + (c_{\sigma(t)}(x) \\ &+ \Delta c_{\sigma(t)}(x, t))\omega_{\sigma(t)}, \\ y =& h_{\sigma(t)}(x), \end{aligned} \tag{1}$$

where $\sigma(t) : [0, +\infty) \to I_m = \{1, \cdots, m\}$ is the switching signal, which is assumed to be a piecewise constant function depending on time, $x \in R^n$ is the state, $\omega_i \in R^{c_i}$ is the disturbance input belongs to $L_2[0, \infty)$ , $u_i \in R^{m_i}$ and $y \in R^{p_i}$ stand for the control input and the measurement output of the $i$th subsystem respectively. $f_i(x), c_i(x)$ and $h_i(x)$ are known smooth nonlinear function matrices of appropriate dimensions satisfying $f_i(0) = 0$ and $h_i(0) = 0$, $\Delta j_i(\dot{x}, t), \Delta f_i(x, t)$ and $\Delta c_i(x, t)$ represent unknown smooth nonlinear function matrices, $i \in I_m$.

The switching sequence $\sigma(t)$ associated with the switched system (1) is given by

$$\begin{aligned} \sum = \{x_0; (i_0, t_0), (i_1, t_1), \cdots, (i_k, t_k), \cdots, \\ |i_k \in I_m, k \in N\}, \end{aligned} \tag{2}$$

in which $t_0$ is the initial time, $x_0$ is the initial state. When $t \in [t_k, t_{k+1}), \sigma(t) = i_k$, the $i_k$th subsystem is active, and the trajectory $x(t)$ of the switched system (1) is the trajectory $x_{i_k}$ of the $i_k$th subsystem. As commonly assumed in the literature, we exclude Zeno behavior for all types of switching signal in this paper. In addition, we assume that the state of the switched system (1) does not jump at the switching instants, i.e., the trajectory $x(t)$ is everywhere continuous.

In this paper, we assume all uncertainties in the switched system (1) having the following properties.

**Assumption 1.** The uncertain functions $\Delta j_i(\dot{x}, t), \Delta f_i\ (x, t)$ and $\Delta c_i(x, t)$ are gain bounded smooth functions described as follows:

$$\begin{aligned} \Delta j_i(\dot{x}, t) &= e_{j_i}\delta_{j_i}(\dot{x}, t), \|\delta_{j_i}\| \le \|W_{j_i}\dot{x}\|, \\ \Delta f_i(x, t) &= e_{f_i}\delta_{f_i}(x, t), \|\delta_{f_i}\| \le \|W_{f_i}(x)\|, \\ \Delta c_i(x, t) &= e_{c_i}\delta_{c_i}(x, t), \|\delta_{c_i}\| \le \|W_{c_i}(x)\|, \end{aligned} \tag{3}$$

where $e_{j_i}, e_{f_i}, e_{c_i}$ are known constant matrices and $\delta_{j_i}, \delta_{f_i}, \delta_{c_i}$ are unknown function vectors with $\delta_{j_i}(0, t) = 0$ and $\delta_{f_i}(0, t)$ $= 0$. $W_{j_i}, W_{f_i}$ are known smooth function matrices, $W_{c_i}$ are given weighting matrices, $i \in I_m$.

Now, the robust $H_\infty$ control problem to be addressed in this paper can be represented as: given a constant $\gamma > 0$, design a switching law $i = \sigma(t)$ for the switched system (1) such that

(i) The autonomous system (1) is globally asymptotically stable when $\omega_i \equiv 0$.

(ii) System (1) has weighted $L_2$-gain from $\omega_i$ to $y$ for all admissible uncertainties, ie., there holds

$$\int_0^\infty e^{-\lambda\tau} y^T(\tau) y(\tau) d\tau \le \gamma^2 \int_0^\infty \omega_i^T(\tau)\omega_i(\tau) d\tau + \beta(x_0)$$

for some real-valued function $\beta(\cdot)$ with $\beta(0) = 0$.

**Assumption 2.** For robust $H_\infty$ control problem, suppose that not all the subsystems of the switched system (1) are stabilizable.

**Definition 1.** For any $T_2 > T_1 \ge 0$, let $N_\sigma(T_1, T_2)$ denote the number of switching of $\sigma(t)$ over $(T_1, T_2)$. If $N_\sigma(T_1, T_2) \le N_0 + \frac{T_2 - T_1}{\tau_a}$ holds for $\tau_a > 0, N_0 \ge 0$, then $\tau_a$ is called average dwell time.

**Definition 2.** For the switched system (1), suppose that $V_i(t)$ is the corresponding Lyapunov function for the $i$th subsystem, then $V(t)$ is called a piecewise Lyapunov function candidate if it can be written as $V(t) = V_{\sigma(t)}(x)$, where $V_{\sigma(t)}(x)$ is switched among $V_i(t)$ in accordance with the piecewise constant switching signal $\sigma(t)$.

## III. Main Results

For the switched system (1) with stabilizable and unstabilizable subsystems, the robust $H_\infty$ control problem is solvable if the stabilizable and unstabilizable subsystems satisfy certain conditions and admissible switching law among them, respectively. In what folllows, we give the design method for the robust $H_\infty$ control problem of the switched system (1).

Consider the switched system (1). Under Assumption 2, for the robust $H_\infty$ control problem, not all the subsystems are stabilizable, without loss of generality, we assume that the $i$th subsystem ($1 \le i \le s$) is stabilizable (where the positive integer s satisfies $1 \le s < m$), whereas the other subsystems of (1) are unstabilizable.

Then, for any piecewise constant switching signal $\sigma(t)$ and any $0 \le t_0 < t$, we let $\Pi^-(t_0, t)$ (resp.,$\Pi^+(t_0, t)$) denote the total activation time of stabilizable (resp., unstabilizable) subsystems during $(t_0, t)$. Then, we present the following switching law:

(F): Let $t_0 < t_1 < t_2 < \cdots < t_i$ ($\lim_{i\to\infty} t_i = \infty$) be a specified sequence of time instants satisfying $\max_i(t_{i+1} - t_i) = T < \infty$. Determined the switching signal $\sigma(t)$ so that the inquality

$$\frac{\Pi^-(t_i, t_{i+1})}{\Pi^+(t_i, t_{i+1})} \ge \frac{\beta + \lambda^*}{\alpha - \lambda^*} \tag{4}$$

holds on time every interval $[t_i, t_{i+1})(i = 0, 1, \cdots)$ with $\alpha > 0, \beta > 0$ and $\lambda^* \in (0, \alpha)$. Meanwhile, we choose $\lambda^* \le \alpha$ as the average dwell time scheme: for any $t > t_0$,

$$N_\sigma(t_0, t) \le N_0 + \frac{t - t_0}{\tau}, \tau > \tau^* = \frac{\ln u}{\lambda^*}. \tag{5}$$

Under the switching law (F) for any $t_0$, $t$ satisfying $t_{i-1} <$

$t_0 < t_i < t_{i+1} < \cdots < t_k < t$, we can infer

$$\begin{aligned}
&\beta\Pi^+(t_0,t) - \alpha\Pi^-(t_0,t)\\
\le&\beta(t_i - t_0) + \sum_{l=i}^{k-1}[\beta\Pi^+(t_l,t_{l+1}) - \alpha\Pi^-(t_l,t_{l+1})]\\
&+\beta(t-t_k)\\
\le&\beta(t_i-t_0) - \lambda^*(t_k-t_i) + \beta(t-t_k)\\
\le&(\beta+\lambda^*)(t_i-t_0) - \lambda^*(t-t_0) + (\beta+\lambda^*)(t-t_k). \qquad (6)
\end{aligned}$$

Since on any interval $[t_i, t_{i+1})$, the total activation time period of unstable subsystems satisfies $\Pi^+(t_i,t_{i+1}) \le \frac{\alpha-\lambda^*}{\alpha+\beta}(t_{i+1}-t_i)$ according to the requirement in (F), we get from (6) that

$$\beta\Pi^+(t_0,t) - \alpha\Pi^-(t_0,t) \le c - \lambda^*(t-t_0), \qquad (7)$$

Where $c = \frac{2(\beta+\lambda^*)(\alpha-\lambda^*)}{(\alpha+\beta)}T$.

The following theorem provides theoretical basis for the robust $H_\infty$ control problem of the switched system (1).

**Theorem 1.** Given any constant $\gamma > 0$, suppose that there exist radially unbounded positive definite differentiable functions $V_i(x)$, $i = 1, \cdots, m$, constants $\mu \ge 1$, such that the following inequalities

$$\begin{aligned}
&\frac{\partial V_i}{\partial x}f_i + \gamma_i^2 C_i^T C_i + \gamma_i^2(\frac{1}{2\gamma_i^2}\frac{\partial V_i}{\partial x}B_i + C_i^T D_i)R_i^{-1}(\frac{1}{2\gamma_i^2}\frac{\partial V_i}{\partial x}\\
&\cdot B_i + C_i^T D_i)^T + \alpha V_i < 0, i \le s, \qquad (8)
\end{aligned}$$

$$\begin{aligned}
&\frac{\partial V_i}{\partial x}f_i + \gamma_i^2 C_i^T C_i + \gamma_i^2(\frac{1}{2\gamma_i^2}\frac{\partial V_i}{\partial x}B_i + C_i^T D_i)R_i^{-1}(\frac{1}{2\gamma_i^2}\frac{\partial V_i}{\partial x}\\
&\cdot B_i + C_i^T D_i)^T - \beta V_i < 0, i > s, \qquad (9)
\end{aligned}$$

$$V_i \le \mu V_j, \qquad (10)$$

and

$$\alpha_1^*(\|x\|) \le V_i(x) \le \alpha_2^*(\|x\|), \quad i,j = 1,\cdots,m \qquad (11)$$

hold, where $\alpha_1^*(x)$ and $\alpha_2^*(x)$ are two class $K_\infty$ functions and

$$\begin{aligned}
&\gamma_i^2 = \frac{\gamma^2}{1+\bar\sigma(W_{c_i})/\lambda_{c_i}^2}, B_i = [c_i, \lambda_{j_i}e_{j_i}, \lambda_{f_i}e_{f_i}, \lambda_{c_i}e_{c_i}],\\
&C_i^T = [({}^1/_{\gamma_i})h_i^T, ({}^1/_{\lambda_{j_i}})f_i^T W_{j_i}^T, ({}^1/_{\lambda_{f_i}})W_{f_i}^T, 0],\\
&D_i^T = [0, ({}^1/_{\lambda_{j_i}})B_i^T W_{j_i}^T, 0, 0], R_i = I - D_i^T D_i \qquad (12)
\end{aligned}$$

with $\lambda_{j_i}, \lambda_{f_i}$, and $\lambda_{c_i}, i \in I_m$ are positive constants.

Then, the robust $H_\infty$ control problem of the switched system (1) is solvable under the switching condition (F) and the average dwell-time (5).

**Proof:** From Definition 2, we choose the following piecewise Lyapunov function candidate:

$$V(t) = V_{\sigma(t)}(x) \qquad (13)$$

for the switched system (1), where $V_{\sigma(t)}(x)$ is switched among the solution $V_i(x)$'s of (8)-(11) in accordance with the piecewise constant switching signal $\sigma$.

Regard neutral uncertainty $\Delta j_i(\dot x, t)$ as an exogenous disturbance and make a new extended disturbance input including it. In this case, define

$$d_i^T = [\omega_i^T, -({}^1/_{\lambda_{j_i}})\delta_{j_i}^T, ({}^1/_{\lambda_{f_i}})\delta_{f_i}^T, ({}^1/_{\lambda_{c_i}})\omega_i^T\delta_{c_i}^T]. \qquad (14)$$

Then, we can conclude that

$$\begin{aligned}
d_i^T d_i \le& \|\omega_i\|^2 + ({}^1/_{\lambda_{j_i}^2})\delta_{j_i}^T\delta_{j_i} + (\bar\sigma(W_{c_i})/\lambda_{c_i}^2)\|\omega_i\|^2\\
&+({}^1/_{\lambda_{f_i}^2})\delta_{f_i}^T\delta_{f_i}\\
\le&(1+\bar\sigma(W_{c_i})/\lambda_{c_i}^2)\|\omega_i\|^2 + ({}^1/_{\lambda_{j_i}^2})\delta_{j_i}^T\delta_{j_i}\\
&+({}^1/_{\lambda_{f_i}^2})\delta_{f_i}^T\delta_{f_i}, \qquad (15)
\end{aligned}$$

which means

$$-\gamma^2\|\omega_i\|^2 \le \gamma_i^2 d_i^T d_i + ({}^{\gamma_i^2}/_{\lambda_{j_i}^2})\delta_{j_i}^T\delta_{j_i} + ({}^{\gamma_i^2}/_{\lambda_{f_i}^2})\delta_{f_i}^T\delta_{f_i}.$$

Owing to Assumption 1, it holds that

$$\begin{aligned}
&\dot V + \|y\|^2 - \gamma^2\|\omega_i\|^2\\
=&\frac{\partial V_i}{\partial x}(f_i + \Delta f_i + c_i\omega_i + \Delta c_i\omega_i - \Delta j_i) + \|y\|^2 - \gamma^2\|\omega_i\|^2\\
=&\frac{\partial V_i}{\partial x}(f_i + e_{f_i}\delta_{f_i} + c_i\omega_i + e_{c_i}\delta_{c_i}\omega_i - e_{j_i}\delta_{j_i}) + \|y\|^2\\
&-\gamma^2\|\omega_i\|^2\\
=&\frac{\partial V_i}{\partial x}(f_i + B_i d_i) + h_i^T h_i - \gamma_i^2 d_i^T d_i + \frac{\gamma_i^2}{\lambda_{j_i}^2}\delta_{j_i}^T\delta_{j_i}\\
&+\frac{\gamma_i^2}{\lambda_{f_i}^2}\delta_{f_i}^T\delta_{f_i}. \qquad (16)
\end{aligned}$$

Furthermore

$$\begin{aligned}
\frac{\gamma_i^2}{\lambda_{j_i}^2}\delta_{j_i}^T\delta_{j_i} \le& \frac{\gamma_i^2}{\lambda_{j_i}^2}(f_i + \Delta f_i + c_i\omega_i + \Delta c_i\omega_i - \Delta j_i)^T W_{j_i}^T\\
&\cdot W_{j_i}(f_i + \Delta f_i + c_i\omega_i + \Delta c_i\omega_i - \Delta j_i)\\
=&\frac{\gamma_i^2}{\lambda_{j_i}^2}(f_i + B_i d_i)^T W_{j_i}^T W_{j_i}(f_i + B_i d_i)\\
=&\frac{\gamma_i^2}{\lambda_{j_i}^2}f_i^T W_{j_i}^T W_{j_i} f_i + \frac{2\gamma_i^2}{\lambda_{j_i}^2}f_i^T W_{j_i}^T W_{j_i} B_i d_i\\
&+\frac{\gamma_i^2}{\lambda_{j_i}^2}d_i^T B_i^T W_{j_i}^T W_{j_i} B_i d_i. \qquad (17)
\end{aligned}$$

Combining the previous two inequalities (16)-(17) and considering (8)-(9), then by completing the squares, there holds

$$\begin{aligned}
&\dot V(x(t)) + \|y\|^2 - \gamma^2\|\omega_i\|^2\\
=&\frac{\partial V_i}{\partial x}(f_i + B_i d_i) + h_i^T h_i - \gamma_i^2 d_i^T d_i + \frac{\gamma_i^2}{\lambda_{j_i}^2}f_i^T W_{j_i}^T W_{j_i} f_i\\
&+\frac{2\gamma_i^2}{\lambda_{j_i}^2}f_i^T W_{j_i}^T W_{j_i} B_i d_i + \frac{\gamma_i^2}{\lambda_{j_i}^2}d_i^T B_i^T W_{j_i}^T W_{j_i} B_i d_i\\
&+\frac{\gamma_i^2}{\lambda_{f_i}^2}\delta_{f_i}^T\delta_{f_i}\\
\le&\frac{\partial V_i}{\partial x}(f_i + B_i d_i) + \gamma_i^2 C_i^T C_i - \gamma_i^2 d_i^T R_i d_i + 2\gamma_i^2 C_i^T D_i d_i
\end{aligned}$$

$$
\begin{aligned}
=&\frac{\partial V_i}{\partial x} f_i - \gamma_i^2 \left\| R_i^{\frac{1}{2}} d_i - R_i^{-\frac{1}{2}} (\frac{1}{2\gamma_i^2} \frac{\partial V_i}{\partial x} B_i + C_i^T D_i)^T \right\|^2 \\
&+ \gamma_i^2 (\frac{1}{2\gamma_i^2} \frac{\partial V_i}{\partial x} B_i + C_i^T D_i) R_i^{-1} (\frac{1}{2\gamma_i^2} \frac{\partial V_i}{\partial x} B_i + C_i^T D_i)^T \\
&+ \gamma_i^2 C_i^T C_i \\
\le&\frac{\partial V_i}{\partial x} f_i + \gamma_i^2 C_i^T C_i + \gamma_i^2 (\frac{1}{2\gamma_i^2} \frac{\partial V_i}{\partial x} B_i + C_i^T D_i) R_i^{-1} \\
&\cdot (\frac{1}{2\gamma_i^2} \frac{\partial V_i}{\partial x} B_i + C_i^T D_i) \\
\le&\left\{ \begin{array}{l} -\alpha V_i, i \le s, \\ \ \beta V_i, \ \ i > s. \end{array} \right.
\end{aligned} \tag{18}
$$

Note that when $\omega(t) \equiv 0$, we know from (18) that for any $t \in [t_k, t_{k+1})\,(t_0 \le k \le N_\sigma(t_0, t))$, the piecewise Lyapunov function candidate (13) satisfies

$$
V(t) = V_{\sigma(t)}(t) \le \left\{ \begin{array}{ll} e^{-\alpha(t-t_0)} V_{\sigma(t_k)}(t_k), & if \ \ i \le s, \\ e^{\beta(t-t_0)} V_{\sigma(t_k)}(t_k), & if \ \ i > s. \end{array} \right. \tag{19}
$$

From (10), $V_{\sigma(t_k)}(t_k) \le \mu V_{\sigma(t_k^-)}(t_k^-)$ is true at the switching point $t_k$. Therefore, we obtain by induction that

$$
\begin{aligned}
V(t) &\le e^{\beta\Pi^+(t_k,t) - \alpha\Pi^-(t_k,t)} V_{\sigma(t_k)}(t_k) \\
&\le \mu e^{\beta\Pi^+(t_k,t) - \alpha\Pi^-(t_k,t)} V_{\sigma(t_k^-)}(t_k^-) \\
&\le \mu e^{\beta\Pi^+(t_k,t) - \alpha\Pi^-(t_k,t)} V_{\sigma(t_{k-1})}(t_{k-1}) \\
&\le \cdots \le \mu^k e^{\beta\Pi^+(t_k,t) - \alpha\Pi^-(t_k,t)} V_{\sigma(t_0)}(t_0) \\
&\le \mu^{N(t_0,t)} e^{\beta\Pi^+(t_k,t) - \alpha\Pi^-(t_k,t)} V_{\sigma(t_0)}(t_0),
\end{aligned} \tag{20}
$$

where $N(t_0, t)$ is the switching numbers in the time interval $(t_0, t)$.

Taking (5) and (7) into account, we get

$$
\begin{aligned}
V(t) &\le \mu^{N(t_0,t)} e^{\beta\Pi^+(t_k,t) - \alpha\Pi^-(t_k,t)} V_{\sigma(t_0)}(t_0) \\
&\le e^{N_0 \ln\mu + c} e^{-(\lambda^* - \frac{\ln\mu}{\tau})(t-t_0)} V_{\sigma(t_0)}(t_0) \\
&\le c_0 e^{-\lambda(t-t_0)} V_{\sigma(t_0)}(t_0),
\end{aligned} \tag{21}
$$

where $c_0 = e^{N_0 \ln\mu + c}$, $\lambda = (\lambda^* - \frac{\ln\mu}{\tau})$.

According to (11), we have

$$
\alpha_1^* \left(\|x\|\right) \le V_i\left(x\right) \le \alpha_2^* \left(\|x\|\right). \tag{22}
$$

Combining (20)-(22) gives

$$
\|x(t)\| \le \alpha_1^{*-1} (c_0 e^{-\lambda(t-t_0)} \alpha_2^* \left(\|x(t_0)\|\right), \tag{23}
$$

which means global asymptotic stability of the switched system (1) with $\omega(t) \equiv 0$. The proof of internal stability is completed.

It can be easily seen from (18) that for any $t \in [t_k, t_{k+1})\,(t_0 \le k \le N_\sigma(t_0, t))$, the piecewise Lyapunov function candidate (13) satisfies

$$
V(t) \le \left\{ \begin{array}{l} e^{-\alpha(t-t_k)} V_{\sigma(t_k)}(t_k) - \int_{t_k}^t e^{-\alpha(t-\tau)} \Gamma(\tau) d\tau, \\ if \ \ \sigma(t_k) = i \le s, \\ e^{\beta(t-t_k)} V_{\sigma(t_k)}(t_k) - \int_{t_k}^t e^{\beta(t-\tau)} \Gamma(\tau) d\tau, \\ if \ \ \sigma(t_k) = i > s. \end{array} \right.
$$

From (10), $V_{\sigma(t_k)}(t_k) \le \mu V_{\sigma(t_k^-)}(t_k^-)$ is true at the switching point $t_k$. Therefore, we obtain by induction that

$$
\begin{aligned}
V(t) \le& e^{\beta\Pi^+(t_k,t) - \alpha\Pi^-(t_k,t)} V_{\sigma(t_k)}(t_k) \\
&- \int_{t_k}^t e^{\beta\Pi^+(\tau,t) - \alpha\Pi^-(\tau,t)} \Gamma(\tau) d\tau \\
\le& \mu e^{\beta\Pi^+(t_k,t) - \alpha\Pi^-(t_k,t)} V_{\sigma(t_k^-)}(t_k^-) \\
&- \int_{t_k}^t e^{\beta\Pi^+(\tau,t) - \alpha\Pi^-(\tau,t)} \Gamma(\tau) d\tau \le \cdots \\
\le& \mu^k e^{\beta\Pi^+(t_0,t) - \alpha\Pi^-(t_0,t)} V_{\sigma(t_0)}(t_0) \\
&- \mu^k \int_{t_0}^{t_1} e^{\beta\Pi^+(\tau,t) - \alpha\Pi^-(\tau,t)} \Gamma(\tau) d\tau \\
&- \mu^{k-1} \int_{t_1}^{t_2} e^{\beta\Pi^+(\tau,t) - \alpha\Pi^-(\tau,t)} \Gamma(\tau) d\tau - \cdots \\
&- \mu^0 \int_{t_k}^{t} e^{\beta\Pi^+(\tau,t) - \alpha\Pi^-(\tau,t)} \Gamma(\tau) d\tau \\
=& \mu^k e^{\beta\Pi^+(t_0,t) - \alpha\Pi^-(t_0,t)} V_{\sigma(t_0)}(t_0) \\
&- \int_{t_0}^t \mu^{N_\sigma(\tau,t)} e^{\beta\Pi^+(\tau,t) - \alpha\Pi^-(\tau,t)} \Gamma(\tau) d\tau \\
=& e^{\beta\Pi^+(t_0,t) - \alpha\Pi^-(t_0,t) + N_\sigma(t_0,t)\ln\mu} V_{\sigma(t_0)}(t_0) \\
&- \int_{t_0}^t e^{\beta\Pi^+(\tau,t) - \alpha\Pi^-(\tau,t) + N_\sigma(\tau,t)\ln\mu} \Gamma(\tau) d\tau.
\end{aligned} \tag{24}
$$

Multiplying both sides of the above inequality by $e^{-N_\sigma(t_0,t)\ln\mu}$ leads to

$$
\begin{aligned}
&e^{-N_\sigma(t_0,t)\ln u} V(t) \\
&+ \int_{t_0}^t e^{\beta\Pi^+(\tau,t) - \alpha\Pi^-(\tau,t) - N_\sigma(t_0,\tau)\ln u} y^T(\tau) y(\tau) d\tau \\
\le& e^{\beta\Pi^+(t_0,t) - \alpha\Pi^-(t_0,t)} V_{\sigma(t_0)}(t_0) \\
&+ \gamma^2 \int_{t_0}^t e^{\beta\Pi^+(\tau,t) - \alpha\Pi^-(\tau,t) - N_\sigma(t_0,\tau)\ln\mu} \omega^T(\tau) \omega(\tau) d\tau.
\end{aligned}
$$

Under the switching law (F) and the average dwell time scheme (5) with $\sigma < \lambda^*$, we can obtain

$$
\begin{aligned}
&\int_{t_0}^t e^{-\alpha(t-\tau) - \sigma\tau} y^T(\tau) y(\tau) d\tau \\
\le& e^{c-\lambda^*} V_{\sigma(t_0)}(t_0) + \gamma^2 \int_{t_0}^t e^{c-\lambda^*(t-\tau)} \omega^T(\tau) \omega(\tau) d\tau.
\end{aligned} \tag{25}
$$

Integrating both sides of the foregoing inequality from $t_0$ to $\infty$ and rearranging the double-integral area, we obtain

$$
\begin{aligned}
&\int_{t_0}^\infty e^{-\sigma\tau} y^T(\tau) y(\tau) d\tau \\
\le& \frac{\alpha e^c}{\lambda^*} V_{\sigma(t_0)}(t_0) + \frac{\alpha e^c}{\lambda^*} \gamma^2 \int_{t_0}^\infty \omega^T(\tau) \omega(\tau) d\tau,
\end{aligned} \tag{26}
$$

which means that the switched system achieves the weighted disturbance attenuation level $\sqrt{\frac{\alpha e^c}{\lambda^*}}\gamma$ under the average dwell time scheme (5) and the switching law (F).

When the switched system (1) is in the following linear form:

$$\begin{aligned}[I+E_{j_i}\sum\nolimits_{j_i}(t)F_{j_i}]\dot{x}=&[A_i+E_{a_i}\sum\nolimits_{a_i}(t)F_{a_i}]\\&+[H_i+E_{h_i}\sum\nolimits_{h_i}(t)F_{h_i}]\omega_i,\\y=&C_ix,\end{aligned}\tag{27}$$

where the uncertain matrices satisfy $\sum_{\upsilon}(t)\sum_{\upsilon}(t)\le I, \upsilon\in\{j_i,a_i,h_i,i\in I_m\}$. Let $\delta_{j_i}=\sum_{j_i}(t)F_{j_i}\dot{x},\delta_{f_i}=\sum_{j_i}(t)F_{j_i}x,\delta_{g_i}=\sum_{h_i}(t)F_{h_i}x$, it is clear that $\upsilon\in\{j_i,a_i,h_i,i\in I_m\}$. satisfy Assumption 1 with $M_{j_i}=F_{j_i},W_{f_i}=F_{a_i},W_{c_i}=F_{h_i}$. Then, we have the following Theorem.

**Theorem 2.** Given any constant $\gamma>0$, suppose that there exist a set of positive definite matrices $P_i, i\in I_m$, constants $\alpha>0,\beta>0$ and $\mu\ge1$, such that the following inequalities

$$\begin{aligned}&P_iA_i+A_i^TP_i+\gamma_i^2C_i^TC_i+\gamma_i^2(\frac{1}{2\gamma_i^2}P_iB_i+C_i^TD_i)R_i^{-1}\\&\cdot(\frac{1}{2\gamma_i^2}P_iB_i+C_i^TD_i)^T+\alpha P_i<0, i\le s,\end{aligned}\tag{28}$$

$$\begin{aligned}&P_iA_i+A_i^TP_i+\gamma_i^2C_i^TC_i+\gamma_i^2(\frac{1}{2\gamma_i^2}P_iB_i+C_i^TD_i)R_i^{-1}\\&\cdot(\frac{1}{2\gamma_i^2}P_iB_i+C_i^TD_i)^T-\beta V_i<0, i>s,\end{aligned}\tag{29}$$

and

$$P_i\le\mu P_j,\quad i,j\in I_m\tag{30}$$

hold, where

$$\begin{aligned}&\gamma_i^2=\tfrac{\gamma^2}{1+\bar{\sigma}(F_{h_i})/\lambda_{p_i}^2},\hat{B}_i=[H_i,\lambda_{j_i}E_{j_i},\lambda_{f_i}E_{a_i},\lambda_{p_i}E_{h_i}],\\&C_i^T=\begin{bmatrix}(^1/_{\gamma_i})C_i\\(^1/_{\lambda_{j_i}})F_{j_i}A_i\\(^1/_{\lambda_{f_i}})F_{a_i}\\0\end{bmatrix},D_i=\begin{bmatrix}0\\(^1/_{\lambda_{j_i}})F_{j_i}B_i\\0\\0\end{bmatrix},\\&R_i=I-D_i^TD_i,\end{aligned}$$

with $\lambda_{j_i},\lambda_{f_i}$, and $\lambda_{c_i}, i\in I_m$ are positive constants.

Then, the switching strategy (5) satisfying (F) solve the robust $H_\infty$ control problem of the switched system (27).

**Proof:** The proof is similar to Theorem 1.

## IV. Example

In this section, we give a numerical example to illustrate the performance of the proposed approach.

**Example 1.** Consider the nonlinear switched system (1) with $\sigma=\{1,2\}$ and

$$\begin{aligned}&f_1(x)=\frac{1}{4}x,c_1=1,h_1=-\frac{1}{2}x,f_2(x)=-2x,c_2=-1,\\&h_2=x,\Delta j_1(\dot{x},t)=a_1\dot{x}\sin t,e_{j_1}=1,\delta_{j_1}=a_1\dot{x}\sin t,\\&W_{j_1}=1,\Delta j_2(\dot{x},t)=a_2\dot{x}\cos t,e_{j_2}=1,\delta_{j_2}=a_2\dot{x}\cos t,\\&W_{j_2}=1,\Delta f_1(x,t)=\frac{1}{2}b_1x\cos t,e_{f_1}=1,\delta_{f_1}=\frac{1}{2}b_1x\sin t,\\&W_{j_2}=\frac{1}{2}x,\Delta f_2(x,t)=b_2x\sin t,e_{f_2}=1,\delta_{f_2}=b_2x\sin t,\\&W_{f_2}=x,\Delta c_1(x,t)=c_1e^{-t},e_{c_1}=1,\delta_{c_1}=c_1e^{-t},W_{c_1}=1,\\&\Delta c_2(x,t)=c_2e^{-t},e_{c_2}=1,\delta_{c_2}=c_2e^{-t},W_{c_2}=1,\end{aligned}\tag{31}$$

and $a_i,b_i,c_i,i=1,2$ are unknown constants in the set $[0,1]$.

It is easy to check that the first subsystem is unstabilizable and the second one is stabilizable. Let $\gamma^2=2$ and $\lambda_{j_i}=\lambda_{f_i}=\lambda_{c_i}=1$, then according to Theorem 1, we obtain

$$\begin{aligned}&\gamma_1=\gamma_2=1,B_1=[-1,1,1,1],B_2=[1,1,1,1],\\&C_1=[x,-2x,x,0],C_2=[-\tfrac{1}{2}x,\tfrac{1}{4}x,\tfrac{1}{2}x,0],\\&D_1^T=\begin{bmatrix}0&-1&0&0\\0&1&0&0\\0&1&0&0\\0&1&0&0\end{bmatrix},D_2^T=\begin{bmatrix}0&1&0&0\\0&1&0&0\\0&1&0&0\\0&1&0&0\end{bmatrix},\end{aligned}$$

We choose $V_1(x)=2x^2,V_2(x)=4x^2$, and $\alpha=0.8,\beta=0.5$. Then following (8)-(9),we can infer

$$\begin{aligned}&\frac{\partial V_1}{\partial x}f_1+\gamma_1^2C_1^TC_1+\gamma_1^2(\frac{1}{2\gamma_1^2}\frac{\partial V_1}{\partial x}B_1+C_1^TD_1)R_1^{-1}\\&\cdot(\frac{1}{2\gamma_1^2}\frac{\partial V_1}{\partial x}B_1+C_1^TD_1)^T-\beta V_2=-\frac{315}{16}x^2\le0,\end{aligned}\tag{32}$$

and

$$\begin{aligned}&\frac{\partial V_2}{\partial x}f_2+\gamma_2^2C_2^TC_2+\gamma_2^2(\frac{1}{2\gamma_2^2}\frac{\partial V_2}{\partial x}B_2+C_2^TD_2)R_2^{-1}\\&\cdot(\frac{1}{2\gamma_2^2}\frac{\partial V_2}{\partial x}B_2+C_2^TD_2)+\alpha V_1=-\frac{182}{15}x^2\le0,\end{aligned}\tag{33}$$

Let $\mu=2,\lambda^*=0.3$, we have $\tau^*=\frac{\ln\mu}{\alpha}=0.8664$ and the activation ratio of stabilizable subsystems to unstabilizable subsystems is $\frac{\Pi^-(t_0,t)}{\Pi^+(t_0,t)}=\frac{\beta+\lambda^*}{\alpha-\lambda^*}=1.6$. Using the switching strategy provided by Theorem 1, we obtained that the robust $H_\infty$ control problem of (1) is solvable, the simulation results are depicted in Figs. 1-2.

## V. Conclusion

In this paper, we have investigated the problem of robust $H_\infty$ control for a class of uncertain nonlinear switched systems based on ADT. Uncertainties are considered to be nonlinearly relied on state and state derivative and allowed to appear in the state, control input and disturbance input. Under the condition that the activation time ratio between stabilizable subsystems and unstabilizable ones is not less than a specified constant, we have derived sufficient conditions for the stabilization and weighted $L_2$-gain property of the switched system. The feasibility of the developed results have been proved by using a numerical example.

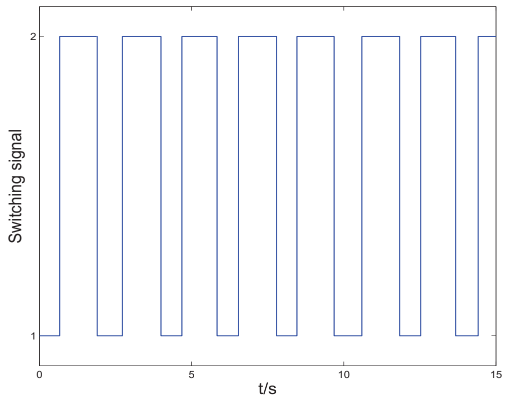


Fig. 1. The switching signals for the switched system (1).

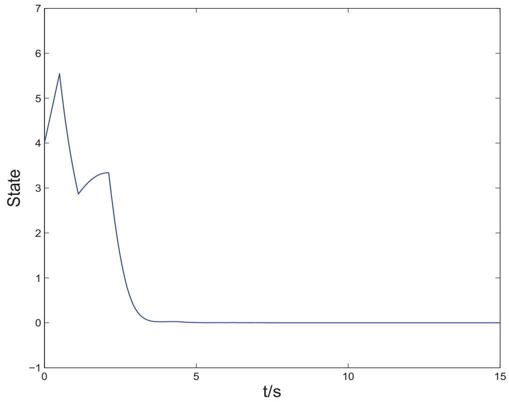


Fig. 2. The state responses of the switched system (1).

## Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grants 61174073 and 90816028. The authors would like to thank the anonymous reviewers for their valuable comments and suggestions for further improving the quality of this article.

## References

[1] D. Liberzon, Switching in Systems and Control, Birkh$\ddot{a}$user, Boston, 2003.

[2] D. Liberzon, A. S. Morse, Basic problems in stability and design of switched systems, *IEEE Control Systems Magazine* 19 (5) (1999) 59-70.

[3] P. Mhaskar, N. H. El-Farra, P. D. Christofides, Predictive control of switched nonlinear systems with scheduled mode transitions, *IEEE Transactions on Automatic Control* 50 (11) (2005) 1670-1680.

[4] Z. D. Sun, S. S. Ge, Analysis and synthesis of switched linear control systems, *Automatica* 41 (2) (2005) 181-195.

[5] J. Daafouz, P. Riedinger, C. Iung, Stability analysis and control synthesis for switched systems: a switched Lyapunov function approach, *IEEE Transactions on Automatic Control* 47 (11) (2002) 1883-1887.

[6] M. Chadli, M. Darouach, Robust admissibility of uncertain switched singular systems, *International Journal of control* 84 (10) (2011) 1796-1806.

[7] M. S. Branicky, Multiple Lyapunov functions and other analysis tools for switched and hybrid systems, *IEEE Transaction on Automatic Control* 43 (4) (1998) 475-482.

[8] J. P. Hespanha, A. S. Morse, Stability of switched systems with average dwell-time, *in: Proceedings of the 38th IEEE Conference on Decision and Control*, Phoenix, AZ, 1999, pp. 2655-2660.

[9] X. M. Sun, J. Zhao, G. M. Dimirovski, State feedback control for discrete delay systems with controller failures based on average dwell-time method, *IET Control Theory and Applications* 2 (2) (2008) 126-132.

[10] A. Y. Aleksandrov, A. A. Kosov, A. V. Platonov, On the asymptotic stability of switched homogeneous systems, *Systems and Control Letters* 61 (1) (2012) 127-133.

[11] K. Sofiane, R. Micky, L. Philippe, Combining $H_\infty$ approach and interval tools to design a low order and robust controller for systems with parametric uncertainties: application to piezoelectric actuators, *International Journal of control* 85 (3) (2012) 251-259

[12] J. Kaloust, Z. Qu, Robust control design for nonlinear uncertain systems with an unknown time-varying control direction, *IEEE Transactions on Automatic Control* 42 (3) (1997) 393-399 .

[13] S. M. Hoseini, M. Farrokhi, A. J. Koshkouei, Robust adaptive control of nonlinear non-minimum phase systems with uncertainties, *Automatica* 47 (1) (2011) 1-13.

[14] W. Z. Su, M. Y. Fu, Robust stabilization of nonlinear cascaded systems, *Automatica* 42 (4) (2006) 645-651

[15] L. Zhang, P. Shi, E. K. Boukas, C. Wang, Robust $l_2$-$l_\infty$ filtering for switched linear discrete time-delay systems with polytopic uncertainties, *IET Control Theory and Applications* 1 (3) (2007) 722-730.

[16] Y. S. Fu, Z. H. Tian, S. J. Shi, Robust $H_\infty$ control of uncertain nonlinear systems, Automatica 42 (9) (2006) 1547-1552.

[17] G. S. Deaecto, J. C. Geromel, J. Daafouz, Dynamic output feedback $H_\infty$ control of switched linear systems, *Automatica* 47 (8) (2011) 1713-1720.

[18] L. Rodrigues, E. K. Boukas, Piecewise-linear $H_\infty$ controller synthesis with applications to inventory control of switched production systems, *Automatica* 42 (8) (2006) 1245-1254.

[19] Z. R. Xiang, C. H. Qiao, M. S. Mahmoud, Finite-time analysis and Robust $H_\infty$ control of uncertain nonlinear systems control for switched stochastic systems, *Journal of the Franklin Institute* 349 (3) (2012) 915-927.

[20] K. Hu, J. Yuan, Finite sum equality approach to $H_\infty$ output-feedback control for switched linear discrete-time systems with time-varying delay, *IET Control Theory and Applications* 3 (8) (2009) 1006-1016.

[21] P. Niamsup, Controllability approach to $H_\infty$ control problem of linear time-varying switched systems, *Nonlinear Analysis: Hybrid Systems* 2 (3) (2008) 875-886.

[22] W. M. Xiang, J. Xiao, $H_\infty$ finite-time control for switched nonlinear discrete-time systems with norm-bounded disturbance, *Journal of the Franklin Institute* 348 (2) (2011) 331-352.

[23] X. Z. Liu, S. Yuan, On designing $H_\infty$ estimator for switched nonlinear systems of neutral type, *Communications in Nonlinear Science and Numerical Simulation* 16 (11) (2011) 4379-4389.

[24] B. Niu, J. Zhao, Stabilization and L2-gain analysis for a class of cascade switched nonlinear systems: An average dwell-time method, *Nonlinear Analysis: Hybrid Systems* 5 (4) (2011) 671-680.

[25] B. Niu, J. Zhao, Robust $H_\infty$ control for a class of switched nonlinear cascade systems via multiple Lyapunov functions approach, *Applied Mathematics and Computation* 218 (11) (2012) 6330-6339.

[26] G. S. Zhai, B. Hu, K. Yasuda, A. N. Michel, Disturbance attenuation properties of time-controlled switched systems, *J. Franklin Inst.* 338 (7) (2001) 765-779.

[27] M. Wang, J. Zhao, $H_\infty$ control for a class of cascade switched nonlinear systems, *Asian Journal of Control* 10 (6) (2008) 724-729.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# An Investigation on the Design and Performance Assessment of double-PID and LQR Controllers for the Inverted Pendulum

Wende Li, Hui Ding
School of Mechanical-Electrical Engineering
Harbin Institute of Technology
Harbin, China
liwende.yecao@gmail.com

Kai Cheng
School of Engineering and Design
Brunel University
London, United Kingdom
Kai.Cheng@brunel.ac.uk

*Abstract*— **The widespread application of inverted pendulum principles requires better dynamic performance and steady state performance of the inverted pendulum system. The objective of this paper is to design and investigate the time specification performance of the inverted pendulum controllers. Two control methods are proposed in this paper, an innovative double PID control method and a modern LQR (liner quadratic regulator) control method. Dynamic performance and steady state performance are investigated and compared of the two controllers. This paper proves that the LQR controller can guarantee the inverted pendulum a faster and smoother stabilizing process and with better robustness and less oscillation than the double-PID controller. The novelty of this paper is the design of the two controllers, and the adoption of limits cycles as the performance assessment method for the inverted pendulum, which not only makes the steady state performance assessment available, but also, provides an effective way for the evaluation of any equilibrium control problem with friction involved.**

*Keywords- inverted pendulum; performance assessment; double-PID controller; LQR controller; limits cycles*

## I. INTRODUCTION (HEADING 1)

The inverted pendulum system, as a typical experimental device for the research and application of control theory, is a single input and multiple outputs, nonlinear and unstable system. The inverted pendulum problem is one of the most important problems in control theory and has always been appealing among researchers [1]. Because the control of the inverted pendulum is similar to other position and equilibrium control tasks, such as the control of the electric two-wheeled self-balancing vehicles, the stabilization of aircraft in the turbulent air flow, the arm position control of open loop robots etc. All these tasks bring new challenges to control engineering, and therefore, much attention has been given to explore better solutions for the inverted pendulum system and to acquire better control performance, include dynamic performance and steady state performance.

In this paper, an innovative double-PID control method and a modern LQR (liner quadratic regulator) control method are designed respectively, and the dynamic performance and steady state performance are assessed and compared based on virtual prototype simulation and experimental setup. The virtual prototype is built which consists of control model, mechanical model, and comprehensive friction model with obtained parameters to fully investigate the steady state performance. Experimental setup validates the virtual prototype simulation and especially the predicted limit circle, which is caused by friction as is proposed in this paper and is adopted as steady state performance assessment indices. The paper proposes a comprehensive assessment on the two designed controllers for an inverted pendulum based on both traditional performance indices and the novel limits cycles.

## II. SESTEM DESCRIPTION AND MODELING

### A. Sestem Description

A schematic of the inverted pendulum is shown in Fig. 1. In this paper, the simulation and experiment are both based on a direct driven inverted pendulum system, in which a pendulum is mounted on a stage driven by an ironless permanent magnet linear motor. Linear motor is a new type driving device which can directly transform electric energy to mechanical linear motion. Sensors are attached to the cart and the pivot in order to measure the cart position and pendulum inclined angle respectively.

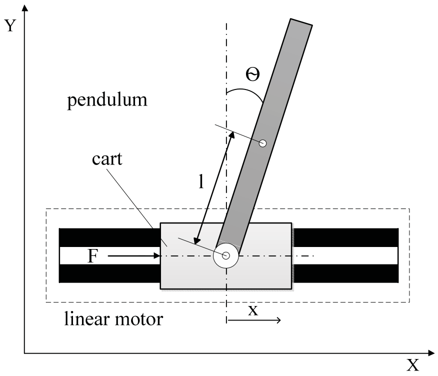


Figure 1. Schematic of the inverted pendulum

Usually the pendulum is in its pendant configuration and there are many methods to swing up the pendulum from its pendant position [2]; when the pendulum approaches the upright unstable equilibrium the control is switched to using a stabilizing controller. This paper focuses on designing the stabilizing controllers. A control force exerted by the linear motor is required on the cart to maintain the pendulum in its upright equilibrium. The control objective in this paper is to maintain the upright unstable equilibrium of the inverted pendulum system with better dynamic qualities, robustness and steady state performance.

### B. *Nonlinear Differential Equations of Seytem Motion*

The motion equations of the cart and pendulum in Fig. 1 can be obtained using Hamilton's principle, shown in (1), (2):

$$(M+m)\ddot{x}(t)+ml\ddot{\theta}(t)\cos\theta(t)-ml\dot{\theta}^2(t)\sin\theta(t)=F_M(t)+F_{fric}(t) \quad (1)$$

$$\frac{4}{3}ml^2\ddot{\theta}(t)+ml\ddot{x}(t)\cos\theta(t)-mgl\sin\theta(t)=f_{fric}(t) \quad (2)$$

where *x* is the position of the cart, $\theta$ is the pendulum angle, measuring from the upright position, *F* is the force applied to the cart, $F_{\text{fric}}$(t)is the friction force between cart and the track, and $f_{\text{fric}}$(t) is the friction force between the pendulum and the pivot, which is very small thus ignored. In this section, since we concern the design of the inverted pendulum controllers, especially the dynamic performance of the controllers, the friction forces are simplified as viscous frictions, as in (3).

$$F_{fric}(t)=-\varepsilon\dot{x} \quad (3)$$

In (3) $\varepsilon$ is coefficient of viscous friction. The definition and value of the inverted pendulum system parameters are given in Table 1.

TABLE I. PARAMETERS IN THE INVERTED PENDULUM SYSTEM

| Parameter | Description | Value |
|---|---|---|
| *M* | Mass of the cart | 1.336kg |
| *m* | Mass of the pendulum | 0.083kg |
| *l* | Distance from the pivot to mass center of the pendulum | 0.1685m |
| *g* | Gravitational constant | 9.8m/s$^2$ |
| *K*$_\text{f}$ | Current to force conversion factor | 1N/A |

### C. *Mathematical Model of Sestem Input Force*

In experimental setup, the effective electromagnetic thrust applies by the linear motor to the cart is given by (4)

$$F_M(t)=\frac{3\pi}{2\tau}\left[\lambda_{PM}i_q+(L_d-L_q)i_qi_d\right] \quad (4)$$

where $\lambda_{PM}$ is flux linkage generated by permanent magnet, $L_\text{d}$ and $L_\text{q}$ is flux linkage of *d*-axis and *q*-axis respectively. The linear motor adopts the zero *d*-axis current control method, thus the applied force can be simplified as (5).

$$F_M(t)=K_fi_q \quad (5)$$

In (5) $i_\text{q}$ is the *q*-axis current after coordinate conversion from the external current supplied to the linear motor, and $K_\text{f}$ is current to force conversion factor, which is given in Table 1 in experimental apparatus.

### D. *Mathematical Model of Inverted Pendulum System*

Combine the equations (1), (2) and the input force equation (5), solve for $\ddot{x}$ and $\ddot{\theta}$ from the differential equations, and introduce the variable: $y=(y_1,y_2,y_3,y_4)^T=(x,\dot{x},\theta,\dot{\theta})$ then after linearization the first-order mathematical model of the inverted pendulum system can be obtained as described in (6).

$$\dot{\mathbf{y}}(t)=\begin{pmatrix}0 & 1 & 0 & 0\\ 0 & \dfrac{-4\varepsilon}{(4M+m)} & \dfrac{\text{-}3mg}{(4M+m)} & 0\\ 0 & 0 & 0 & 1\\ 0 & \dfrac{3\varepsilon}{(4M+m)l} & \dfrac{3(M+m)g}{(4M+m)l} & 0\end{pmatrix}\mathbf{y}(t)+\begin{pmatrix}0\\ \dfrac{4K_f}{4M+m}\\ 0\\ \dfrac{-3K_f}{4M+ml}\end{pmatrix}i_q(t) \quad (6)$$

$$=A\mathbf{y}+\mathbf{b}i_q(t)$$

## III. DESIGN AND TUNING OF DOUBLE-PID CONTROLLER

Taking the Laplace transform of the system first-order equation, we can obtain the transfer function of the system as

$$G_A(s)=\frac{\Theta(s)}{I(s)}$$
$$=\frac{mlK_fs}{[m^2l^2-\frac{4}{3}ml(M+m)]s^3-\frac{4}{3}\varepsilon ml^2s^2+ml(M+m)gs+\varepsilon mgl} \quad (7)$$

$$G_P(s)=\frac{X(s)}{I(s)}$$
$$=\frac{4ml^2K_fs^2-3mglK_f}{[4m(M+m)l^2-3m^2l^2]s^4+4\varepsilon ml^2s^3\text{-}3m(M+m)gls^2-3\varepsilon mgls} \quad (8)$$

Several methods have been proposed to control the inverted pendulum, such as traditional PID control [3], fuzzy control [4], genetic algorithm optimizing control [5], and linear quadratic regulator (LQR) control [6]. Although a lot of control algorithms are researched in the designing of the inverted pendulum system controller, PID controller is still the most widely used controller structure in the realization of a control system. However, the inverted pendulum system is a one input and two output system which contradicts to the one input and one output control characteristic of the single PID controller. The steady state error of the pendulum angle using single PID controller results in the one direction displacement of the cart. In [3], the displacement of the cart cannot be controlled very well because the PID controller can control only one variable, and in [7], the cart position control problem is ignored and only pendulum angle is focused. In this paper, we attempt to use double PID control structure to solve the multi-output problem. Block diagram of the designed double-PID control method is shown as follows in Fig. 2.

The angle reference signal is zero at the upright equilibrium position. The position reference signal is given by a pulse signal, with pulse amplitude 0.5m, pulse period 20s and pulse width 50% of period, to test the position tracking ability of the controller. After tuning the control parameters, we obtain a group of parameters for double-PID controller, with which the controller is capable of controlling the inverted pendulum system and providing a very robust performance.

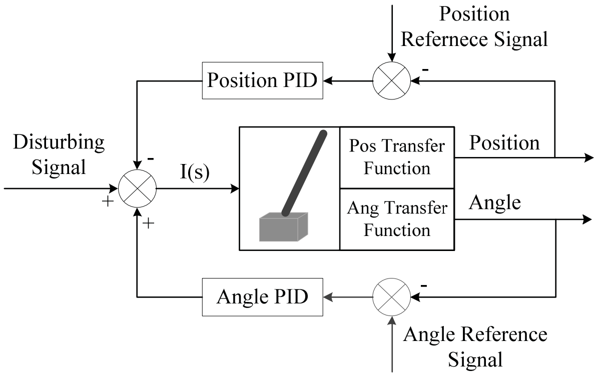


Figure 2. Block diagram of double-PID control method

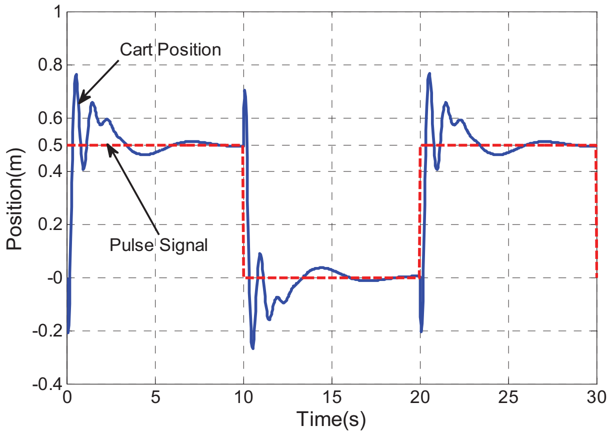


Figure 3. Cart position when tracking pulse signal

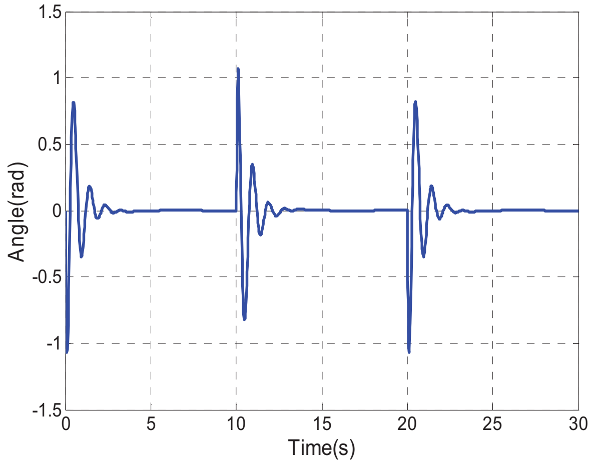


Figure 4. Pendulum angle keep balancing while cart tracking reference signal

The control parameters are:

[$K_{pp}$=20, $K_{pi}$=30, $K_{pd}$=14; $K_{ap}$=50, $K_{ai}$=90, $K_{ad}$=8].

Simulation result shows that the cart can track the position reference signal while the pendulum angle still balancing at the upright equilibrium point with occasional tracking disturbing, as shown in Fig. 3 and Fig. 4. Dynamic performance of the inverted pendulum using double-PID controller will be discussed later.

## IV. Design and Tuning of LQR Controller

### A. *Design of LQR Controller*

Still considering the first-order mathematical model of inverted pendulum system (6), using linear quadratic control theory to design the control $i_q$(t) so that the upright position becomes a stable equilibrium point as described in [8]. Consider the linear quadratic cost function (9)

$$J = \inf_{i_q(t)\in L(0,\infty)} \int_0^\infty [\mathbf{y}(t)^T Q\mathbf{y}(t) + r i_q^2(t)]dt \qquad (9)$$

The weights $Q \geq 0$ , $r \geq 0$ are then chosen to reflect the relative importance error loss of the state **y**, and energy loss of the current $i_q$(t). $Q$ and $r$ are weighting matrixes that penalize certain states and control inputs of the system. $Q$ is positive semi-definite matrix, while $r$ is positive definite matrix. If this system is disturbed and offset the zero state, the control $i_q$(t) can make the system come back to zero state and $J$ is minimal at the same time [9]. The solution to this problem is the feedback law as shown in (10)

$$i_q(t) = -r^{-1}\mathbf{b}^T P(t)\mathbf{y}(t) \qquad (10)$$

where $P$(t) is the solution of Riccati equation, which is

$$A^T P + PA - P\mathbf{b}r^{-1}\mathbf{b}^T P + Q = 0 \qquad (11)$$

Then we can obtain the value of $P$, and the linear optimal feedback matrix can be obtained using (12)

$$K = r^{-1}B^T P = [k_1, k_2, k_3, k_4]^T \qquad (12)$$

The Riccati equation may be solved numerically for given values for $A$, **b**, $r$, and $Q$. Fig. 5 shows the full states feedback representation of inverted pendulum system.

### B. *Controller Solving and System Simulation*

According to the LQR optimal control law, its optimality is totally depended on the selection of $Q$ and $r$. However, there is no resolving method to choose these two matrices. The widespread method used to choose $Q$ and $r$ is by means of simulation and trial [10, 11, and 12]. Since we consider the control of the cart position and pendulum angle, so in this section, the LQR controller is tuned by changing the nonzero elements of the $Q_{11}$ and $Q_{33}$ elements, which, corresponding to and closely influence the two concerned states: cart position and pendulum angle. We choose the weighting matrix $Q$=diag[4000, 0, 5100, 0], and $r$=1 after trial and error.

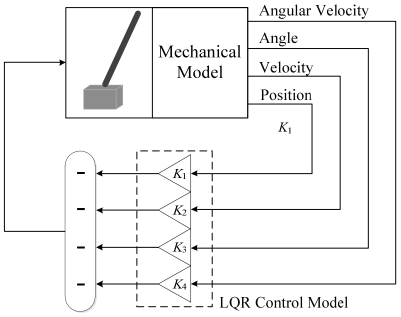


Figure 5. Block diagram of LQR controller possesses full states feedback

The linear optimal feedback matrix $K$ can be calculated using the weighting matrix $Q$, which is

$$K = [-63.2456, -40.0071, 137.2913, 17.6830]$$

Dynamic performance of LQR controller can be assessed by means of step response of the cart and the corresponding disturbing response of the pendulum. Suppose that the cart position reference input is 0.5m (step signal), and the pendulum angle reference input is zero, Simulation Result in Fig. 6 shows that the cart position can be controlled and pendulum can be balanced smoothly using LQR optimal controller, and the system possess good performance.

## V. Dynamic Performance Assessment on the Designed Controllers

Dynamic performance indices are chosen to reflect practical control effect. Performance indices include *rise time*, *transition time*, *steady state error*, and *maximum overshoot*, for the response of cart position and pendulum angle.

For comparison of the system dynamic response, plot the step response (cart position reference input is 0.5m for double-PID and LQR) of the cart and the corresponding disturbing response of the pendulum about the double-PID and LQR controller in one figure, see Fig. 7.

Table 2 and Table 3 show the summary of the dynamic performance indices for the double-PID controller and LQR controller.

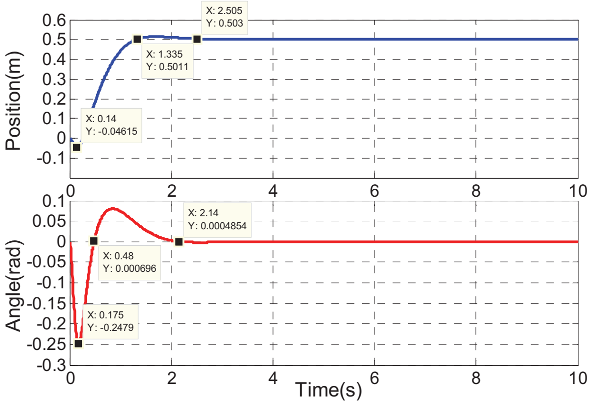


Figure 6. Dynamic response of the system using LQR controller

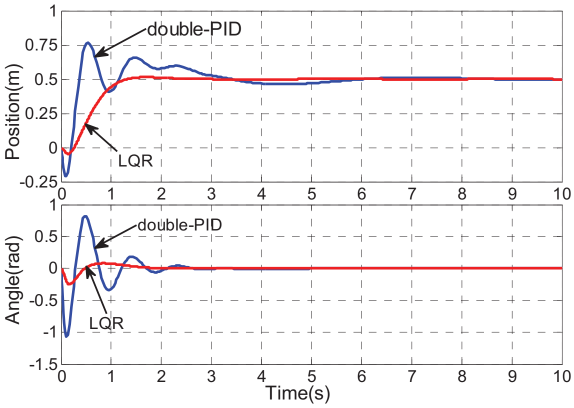


Figure 7. Dynamic response comparison of designed controllers

TABLE II. PERFORMANCE INDICES OF CART POSITION

| Time Response Specification | Double-PID | LQR |
|---|---|---|
| Rise Time | 0.36s | 1.34s |
| Transition Time | 8.65s | 2.51s |
| Steady State Error | ≈0.00 | ≈0.00 |
| Maximum Overshoot | 0.265m | 0.046m |

TABLE III. PERFORMANCE INDICES OF PENDULUM ANGLE

| Time Response Specification | Double-PID | LQR |
|---|---|---|
| Rise Time | 0.28s | 0.48s |
| Transition Time | 3.13s | 2.14s |
| Steady State Error | ≈0.00 | ≈0.00 |
| Maximum Overshoot | 1.07rad | 0.248rad |

Based on the performance indices tabulated in the two tables, LQR controller has the faster transition time both in cart position and pendulum angle, and with the longer rise time of cart position and pendulum angle, the balancing process will be more smooth. Furthermore, the pendulum angle maximum overshoot using double-PID method is bigger than using LQR method, which is because of the shorter rise time of the cart position when using double-PID contributes to larger pendulum angle shocking and maximum overshoot.

In conclusion, for dynamic response, the inverted pendulum controlled by LQR controller 1) balances faster because of the shorter settling time; 2) has a smoother response because of longer rise time; 3) has better robustness due to less maximum overshoot and shocking numbers. So, the LQR controller can guarantee the inverted pendulum system possess better dynamic performance.

## VI. Assessment on Steady State Performance of the Designed Controllers

In the former section, friction is simplified as viscous friction, and this is valid in the designing of the double-PID and LQR controllers, especially for evaluating the dynamic performance of the designed controllers. However, this kind of simplification leads to near zero steady state error, and obviously it is not the case. In this section, the steady state performance, with the influence of friction, can be investigated by means of virtual prototype and experimental setup.

### *A. Comprehensive Friction Model*

The viscous friction, as described before, is described as $F_{fric}(t) = -\varepsilon \dot{x}$. The Coulomb friction $F_C$ is proportional to the normal load, i.e. $F_C = \mu_s F_N$. The static friction $F_{\text{static}}$ counteracts the external applied linear motor forces $F_M$(t). Define $F_S$ as the maximum static friction, which is higher than the Coulomb friction, however we simplify it as $F_S$=$F_C$. Thus the static friction can be described as in (13).

$$F_{static} = \begin{cases} -F_M & \text{if } \dot{x} = 0 \text{ and } |F_{applied}| < F_S \\ \mu_s F_N \,\text{sgn}(F_M) & \text{if } \dot{x} = 0 \text{ and } |F_{applied}| \geq F_S \end{cases} \quad (13)$$

The friction components (viscous friction, Coulomb friction and static friction) can be combined in different ways and such combination is referred as the classical model. However, the classical combination results of a discontinuous friction force at the transition segment from static friction to the Coulomb and viscous friction. The discontinuity contradicts to experimental observation [13]. A common way that can smooth the transition segment is to add an exponential factor [14]. This leads to the comprehensive friction model in (14).

$$F_{fric} = \begin{cases} F_{static} & \text{if } \dot{x} = 0 \\ -(\mu_c + (\mu_s - \mu_c)e^{-(\dot{x}/v_s)^{\gamma}})F_N \operatorname{sgn}(\dot{x}) - \varepsilon \dot{x} & \text{if } \dot{x} \neq 0 \end{cases} \quad (14)$$

where $\varepsilon$ , $\mu_s$ and $\mu_c$ are the coefficient of viscous, Coulomb, and static frictions, respectively. $F_N$ is the magnitude of the normal force equal to gravitation of the cart plus the pendulum, $v_s$ is called the Stribeck velocity, and $r$ is the form factor. The using of the comprehensive friction models confronts the problem of detecting zero velocity. Karnopp proposed the creation of a range of values and within the range the movement velocity is zero [15]. For velocities within this range, the friction is static friction. In the simulation we set ZR=0.005m/s.

To estimate $\mu_c$ and $\varepsilon$ we applied two constant current values $i_q$ to the linear motor and let the cart to reach a steady state, and the corresponding constant final velocity, $\dot{x}$ , can be obtained. Then we can use the steady state equation of motion to solve for $\mu_c$ and $\varepsilon$ . When estimating the value of th static friction coefficient $\mu_s$ , the Armstrong-Helouvry procedure [16] was used, that is to gradually increase the current applied to the linear motor and measure the current (and hence force) necessary to make the cart start to move. The Stribeck parameters, $v_s$ and $r$ can be chose by trial and error, so that the simulations of the model can best match experimental data. The value of the friction parameters are given in Table 4.

TABLE IV. PARAMETERS OF THE COMPREHENSIVE FRICTION MODEL

| Parameter | Description | Value |
|---|---|---|
| $\mu_s$ | Coefficient of static friction | 0.042 |
| $\mu_c$ | Coefficient of Coulomb friction | 0.027 |
| $\varepsilon$ | Coefficient of viscous friction | 0.1N/m/sec |
| $\upsilon_s$ | Stribeck velocity | 0.083m/s |
| $\gamma$ | Form factor | 2 |

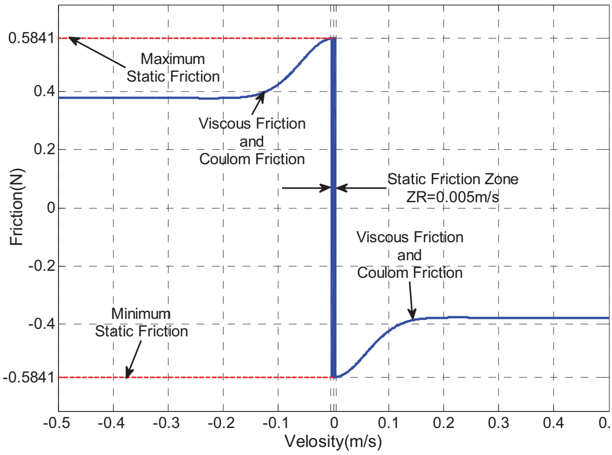


Figure 8. Force-Velocity plot of the friction model

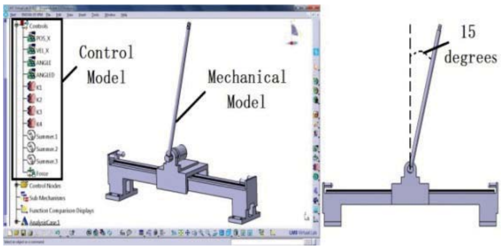


Figure 9. Virtual prototype of the inverted pendulum system

A Force-Velocity plot for the comprehensive friction model with parameters obtained from experiment is shown in Fig. 8. The figure shows that the maximum friction is 0.5841N.

### B. Assessment on the Steady State Performance based on Virtual Prototype Simulation

A virtual prototype of the inverted pendulum system is developed in Virtul.Lab environment, as shown in Fig. 9. The virtual prototype allows us easily inputting the conditions in real environment, such as frictions. The virtual prototype consists of mechanical model and control model. The mechanical model simulates the experimental setup of the inverted pendulum in size and mass. The control model simulates the double-PID and LQR controller respectively. The initial condition is that the pendulum starts with an inclined angle 15 degrees from upright position.

The steady state performance indices, as well as the dynamic performance indices, of the two controllers are shown in Fig. 10 and Fig. 11. The figures show that the steady-state response of the system has oscillation, the so called limit cycle, which is validated both in theory and in experiment, and can be seen as steady state error. Reference [17] proposed that the presence of friction produces limits cycles. Considering one of the practical use of inverted pendulum, the electric two-wheeled self-balancing vehicle, lots of efforts have been spent to reduce the existed vibration by improved control algorithm [18] or available friction compensation model. Thus for practical application of the inverted pendulum, the smaller oscillation amplitude is the better.

Table 5 shows a comparison between the performance indices of double-PID controller and LQR controller. The oscillation amplitude of cart position is 0.095m and pendulum angle is 5.73deg when using double-PID controller, while using LQR controller the amplitude is 0.025m and 1.15deg respectively, which means that the LQR controller can guarantee smaller steady state error. The oscillation frequency of cart position and pendulum angle is 0.417Hz when using double-PID controller, while using LQR controller the frequency is 0.250Hz, which means that the oscillation is slower, this indicates that the cart or the pendulum will oscillate with more smooth.

TABLE V. STEADY STATE PERFORMANCE INDICES

| Performance Index | double-PID | LQR |
|---|---|---|
| Cart Position *Oscillation Amplitude* | 0.095m | 0.025m |
| Pendulum Angle *Oscillation Amplitude* | 5.73deg | 1.15deg |
| Cart Position *Oscillation Frequency* | 0.417Hz | 0.250Hz |
| Pendulum Angle *Oscillation Frequency* | 0.417Hz | 0.250Hz |

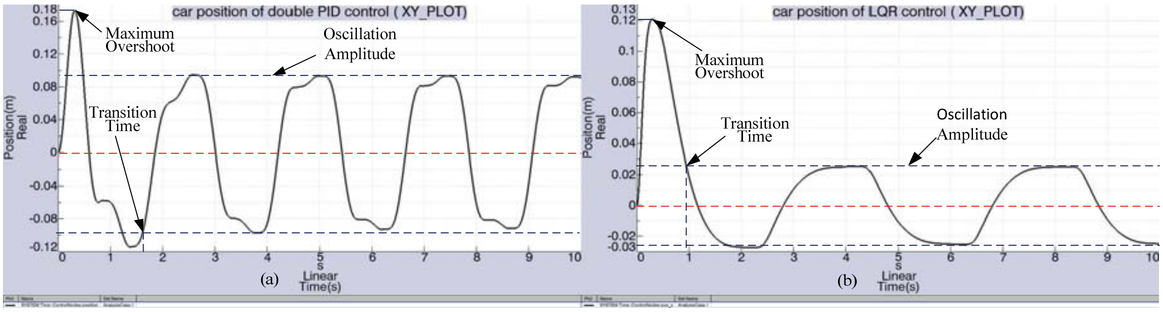


Figure 10. Comparison of cart position between double-PID controller and LQR controller based on virtual prototype

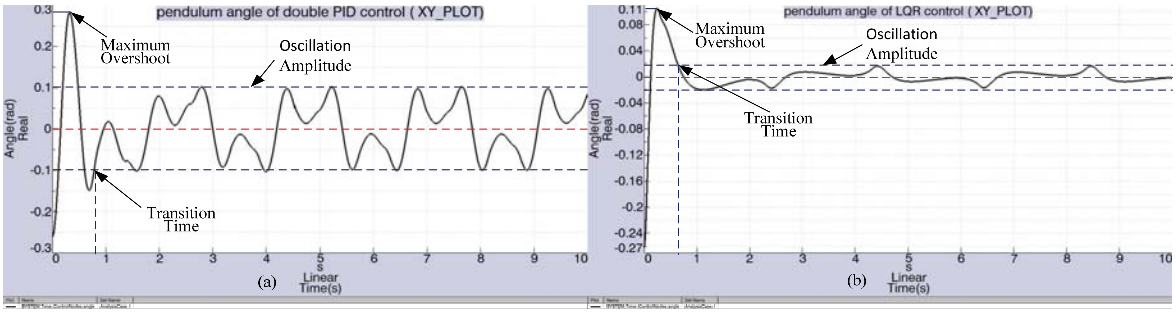


Figure 11. Comparison of pendulum angle between double-PID controller and LQR controller based on virtual prototype

### C. *Assessment on the Steady State Performance based on Experimental Setup*

Experimental setup based on real-time control in dSPACE environment provides the interface between the Simulink environment variables and the real system. Experimental setup, as shown in Fig. 12, intends to investigate the steady state performance of the two controllers. Results are shown in Fig. 13, Fig. 14 and Fig. 15. Table 5 summarizes the steady state performance indices of the two double-PID and LQR controllers. Note that Figure 15 plots the control error signal which equals the control force applied by linear motor. We can see that smaller control force is needed for maintaining the equilibrium position at steady state when using LQR controller. Thus we can conclude that the steady state performance of the LQR controller is better for its smaller oscillation amplitude, oscillation frequency and maintaining force. The data also shows that the simulation results of virtual prototype and that of the experiments is very close.

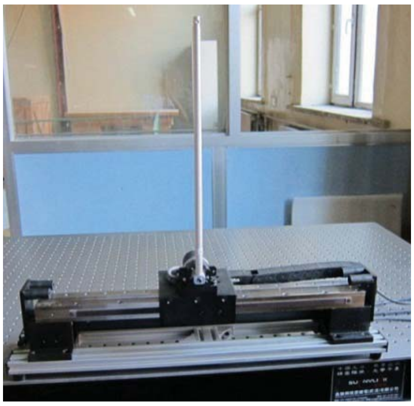


Figure 12. Experimental setup of inverted pendulum

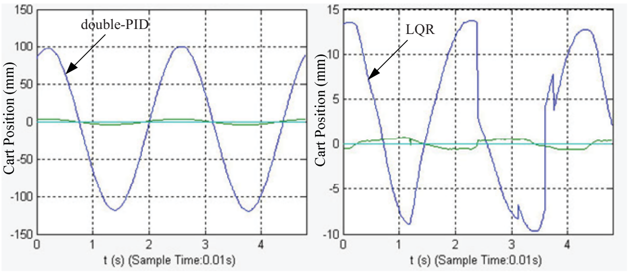


Figure 13. Cart position in experimental setup

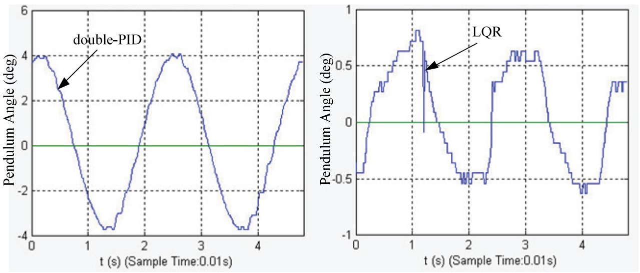


Figure 14. Pendulum angle in experimental setup

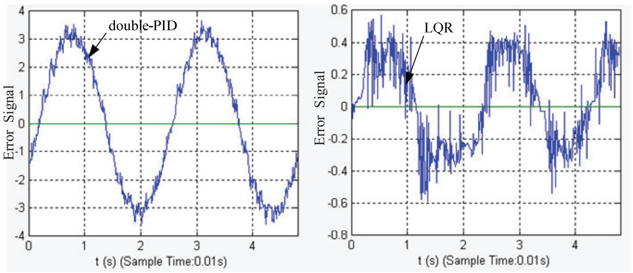


Figure 15. Control error signal in experimental setup

TABLE VI. STEADY STATE PERFORMANCE OF THE DESIGNED CONTROLLERS BASED ON EXPERIMENTAL SETUP

| **Performance Index** | **double-PID** | **LQR** |
|---|---|---|
| Cart Position *Oscillation Amplitude* | 110mm | 10mm |
| Pendulum Angle *Oscillation Amplitude* | 4deg | 0.6deg |
| Cart Position *Oscillation Frequency* | 0.455Hz | 0.431Hz |
| Pendulum Angle *Oscillation Frequency* | 0.455Hz | 0.431Hz |
| Control *Error (Force) Amplitude* | 3N | 0.6N |

## VII. CONCLUSIONS

This paper proposes two control methods for the inverted pendulum, an innovative double-PID control method and modern LQR control method. Control parameters are tuned based on simulation, and results show the two proposed controllers are all capable of controlling the cart position and pendulum angle of the inverted pendulum system with acceptable performance. Dynamic performance and steady state performance are investigated of the two proposed control methods. Dynamic performance proves that the proposed LQR controller can guarantee the inverted pendulum a faster and smoother stabilizing process and with better robustness than the double-PID controller. Steady state performance adopts limits cycles as assessment indices, which not only makes the assessment available, but also, provides an effective way for the evaluation of any equilibrium control problem with friction involved. Results show that the steady state performance of the proposed LQR controller has smaller oscillation amplitude than that of the double-PID controller. Further improvement need to be done for both of the controllers. The double-PID controller should be improved to guarantee better dynamic and steady state performance, and the LQR controller can be improved, such as using friction compensation, to further reduce the oscillation amplitude and frequency.

## ACKNOWLEDGMENT

The authors would like to thank Winner Motor Ltd for kindly supporting the experimental equipment for this research.

## REFERENCES

[1] X. M. Chen, H. X. Zhou, R. H. Ma, F. C. Zuo, G. F. Zhai, M. L. Gong, "Linear motor driven inverted pendulum and LQR controller design," Proceedings of IEEE, International Conference on Automation and Logistics. Jinan. China, pp. 1750–1754, August 2007.

[2] D. Chatterjee, A. Patra, H. K.Joglekar, "Swing-up and stabilization of a cart-pendulum system under restricted cart track length," system & control letters, 2002, vol.47, pp. 335-364.

[3] Y. Xin, B. Xu, H. Xin, J. Xu, L. Y. Hu, "The computer simulation and real-time control for the inverted pendulum system based on PID," Communication Systems and Information Technology, Lecture Notes in Electrical Engineering, 2011, vol.100, pp. 729-736.

[4] J. Yi, N. Yubazaki, "Stabilization fuzzy control of inverted pendulum systems," Artificial Intelligence in Engineering, 2000, pp.153–163.

[5] T. K. Liu, C. H. Chen, Z. S. Li, J. H. Chou, "Method of inequalities-based multi-objective genetic algorithm for optimizing a cart-double-pendulum system," International Journal of Automation and Computing, 2009, vol. 6, pp. 29-37.

[6] H. M. Wang, S. J. Jiang, "LQR control of single inverted pendulum based on square root filter," Advanced Materials Research, 2012, pp. 433-440.

[7] J. J. Wang, "Simulation studies of inverted pendulum based on PID controllers," Simulation Modelling Practice and Theory, 2011, vol. 19, pp. 440-449.

[8] M. Ge, M. S. Chiu, Q. G. Wang, "Robust PID controller design via LMI approach," Journal of Process Control, 2002, vol.12, pp. 3-13.

[9] Y. Z. Yin, H. S. Zhang, "Linear quadratic regulation for discrete-time systems with single input delay," Proceedings of IEEE, China Control Conference, Shenzhen, China, pp. 672-677, Angust 2006.

[10] K. Barya, S. Tiwari, R. Jha, "Comparison of LQR and robust controllers for stabilizing inverted pendulum system," Proceedings of IEEE, International Conference on Communication Control and Computing Technologies, Jalandhar, India, pp. 300-304, October 2010.

[11] Z. Y. Xu, X. D. Huang, "Based on linear quadratic regulator optimal control of single inverted pendulum system Co-simulation of ADAMS and SIMULINK," Advanced Materials Research, 2011, vol. 383-390, pp. 7258-7264.

[12] L. Y. Hu, G. P. Liu, X. P. Liu, H. Zhang, "The computer simulation and real-time stabilization control for the inverted pendulum system based on LQR," Proceedings of IEEE, Fifth International Conference on Natural Computation, Nanchang, China, pp. 438-442, August 2009.

[13] S. A. Campbell, S. Crauford, K. Morris. "Friction and the Inverted Pendulum Stabilization Problem," Journal of Dynamic system, Measurement, and Control, 2008, vol. 130, pp. 054502 (1-7).

[14] B. Armstrong Helouvry, P. Dupont, C. Canudas de Wit, "A survey of Models, Analysis Tools and Compensation Methods for the control of Machines with Friction," Automatica, 1994, vol. 30, pp. 1083-1138.

[15] H. Olsson, K. J. Astrom, C. Canudas de Wit, M. Cafvert, P. Lischinsky, "Friction Models and Friction Compensation," European Journal of Control, 1998, vol. 3, pp. 176-195.

[16] B. A. Helouvry, "Control of Machines with Friction," Journal of Tribology, 1992, vol. 114.

[17] H. T. Teixeira, Victor Semedo de Mattos Siqueira, Celso Jose Munaro, "Comparison of Methods for Estimation and Compensation of Friction Applied to an Inverted Pendulum," Proceedings of the IEEE, International Conference on Control and Automation, Santiago, Chile, pp. 818-823, December 2011.

[18] D. Chol, J. Oh, "Human-friendly Motion Control of a Wheeled Inverted Pendulum by Reduced-order Disturbance Observer," Proceedings of IEEE, International Conference on Robotics and Automation, Pasadena, USA, pp. 2521-2526May 2008.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# An adaptive PI controller for room temperature control with level-crossing sampling

Burkhard Hensel, Volodymyr Vasyutynskyy, Joern Ploennigs, Klaus Kabitzsch
Dresden University of Technology
Chair for Technical Information Systems, 01062 Dresden, Germany
Email: {burkhard.hensel, volodymyr.vasyutynskyy, joern.ploennigs, klaus.kabitzsch}@tu-dresden.de

***Abstract*—Event-based sampling allows saving energy in the sensor transmitter by avoiding unnecessary messages. One important application is room temperature control with wireless sensors. Optimizing the controller parameters of a PI controller for this application is a difficult task, because usually no process model is available and challenging issues like actuator saturation have to be taken into account. Adaptive controllers offer the possibility to tune themselves automatically. In this paper, an adaptive PI controller based on pattern recognition is proposed, designed for room temperature control, sensor energy efficiency, and level-crossing sampling. The implementation is much easier than that of most other adaptive controllers and robustness to disturbances and noise is high. The focus of this paper lies rather on the basic idea, simulations and practical issues than on theoretical investigations.**

## I. Introduction

Room temperature control loops using wireless sensor networks allow high quality control with lower costs than using wired sensors, especially if the building automation system is not installed in the construction phase of the building. Reduction of the energy consumption of the nodes is one of the most investigated research issues in that field. Much energy can be saved by reducing the message count because sending messages requires much more energy than computing [1]. It has been shown that level-crossing sampling (also called send-on-delta, deadband, or Lebesgue sampling) allows a reduction of messages compared to periodic sampling while assuring the same control quality [2]. Therefore, level-crossing sampling can be used in common commercial building automation system technologies, e. g. LonWorks and EnOcean. The main idea of level-crossing sampling is that a new message from the sensor to the controller is only sent if the controlled (measured) signal has changed from the last sent value at least by a threshold $\Delta_{lc}$:

$$|y_m(t_n) - y_m(t_{n-1})| \geq \Delta_{lc} \tag{1}$$

where $y_m(t)$ is the measured signal, and $t_n$ and $t_{n-1}$ are two subsequent time instances at which a message is sent. Usually the sensor wakes up periodically, measures then the current value of the controlled variable (here: room temperature) and decides according to (1) whether a message has to be sent [2].

Adaptive control allows near-optimal control without manual process identification because the controller sets its parameters itself based on available information from past control actions. Additionally, adaptive controllers change their parameters automatically if the process changes. This allows to install an untuned controller in each room of a building, and after start-up each controller optimizes itself according to the room it has to control. This allows cost reductions compared to manual tuning and higher control performance than using always-working, safe, but conservative settings.

Possible reasons for differences between the rooms are the varying room size, wall material, window area, leaking doors or windows, heating and cooling equipment, duct architecture, sensor/actuator location and sensor inertia. Reasons for process changes are variations of the flow temperature from the central heat generator, larger changes of furniture (energy storages), refurbishments (e. g. new windows, new façade insulation), sensor/actuator replacement, variations of the air flow, and changes in the HVAC (heating, ventilation and air conditioning) system.

The main contribution of this paper is an adaptive control algorithm for usage with level-crossing sampling and special emphasis on typical problems of room temperature control, i. e. actuator saturation, typical disturbances, and processes of unknown order. The goal of the adaptation is good control performance in combination with energy efficiency of the sensor. The focus of this paper is the basic idea, practical problems and simulation results; a more theoretical investigation is currently done by the authors.

This article is structured as follows. Section II gives an overview on possibilities for adaptive control based on nonuniformly sampled signals. Section III defines precisely the objective of the adaptive controller. Improvements over an older tuning rule are presented in section IV. The new adaptive controller is explained in section V. The approach is verified using simulations in section VI. While the algorithm is based on reference step responses, in section VII disturbance compensation is briefly discussed. Finally, section VIII draws the conclusions.

## II. Overview: Possibilities for adaptive control with level-crossing sampling

To the authors' knowledge, there are only few works towards adaptive control with nonuniform sampling. Pawlowski et al. used a gain-scheduling controller based on outside temperature and outside wind speed together with several event-based sampling schemes for controlling the temperature in a greenhouse [3]. Dormido et al. published an autotuner

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

based on a relay test with level-crossing sampling [4]. But, it cannot be used without interrupting the usual control action. Wang and Hovakimyan presented an $\mathcal{L}_1$ Adaptive Controller for an event-based, but not level-crossing sampling scheme [5].

A usual adaptive controller consists of one component identifying the process and another component for tuning the controller based on this process model [6]. In combination with level-crossing sampling the task can be split in the process identification for nonuniformly sampled signals and a separate PI controller tuning rule based on the identified model. A tuning rule for the latter point has been given by the authors [7].

In literature, there exist three basic approaches for process identification using nonuniformly sampled signals:

1) Resampling/interpolation followed by the application of methods for uniformly sampled signals.
2) Approximate Fourier transform and application of methods based on the frequency spectrum.
3) Identification of continuous-time ARMA (autoregressive moving average) models.

All these possibilities are quite complex and computationally expensive [8]–[10]. Besides, the usual adaptation methods for periodically sampled signals have strong disadvantages [6], [11], e. g. if the time delay or process order is not known in advance or if there are larger disturbances. Unfortunately, these issues are unavoidable in room temperature control. Even critical effects like *bursting* [12] or instability are possible. In addition to the methods above, there are many well known simple graphical approaches based on the open-loop step response, but these cannot be applied in closed loop what is necessary for an adaptive controller.

However, there are some approaches for adaptive control which do not need a detailed model and are thus also independent of the sampling scheme: Model free adaptive control and pattern recognition based adaptive control.

In *Model-free adaptive control* a sensitivity function is estimated from measured data and the controller output is computed using a simple learning algorithm [13]–[15] or an artificial neural network [16], [17]. Tuning the parameters of these methods (e. g. special learning parameters of the prevailing method, or security parameters for avoiding division by zero) is difficult, especially without a process model, because the optimal learning parameter settings depend on the controlled process while the final adapted parameters depend on the learning parameters. The meaning and magnitude of these parameters is not comprehensible for non-experts.

*Pattern recognition based adaptive control* imitates the procedure which a skilled control engineer would perform if there would be the task to improve a running but insufficiently tuned control loop without building a model of the process. The common part of these algorithms is that the first step is to extract special *features* from the measured signals. Pattern recognition based adaptive controllers usually update the controller parameters not permanently but only after significant events like set-point changes or larger disturbances. Some authors avoid therefore the name "adaptive" and use "auto-tuning" instead [18]. However, the problem with the term "auto-tuning" is that most commercially available auto-tuning controllers update their parameters not regularly but only when the user starts a procedure, mostly an open-loop experiment.

Note that the term "pattern recognition" is more known for classification and grouping of patterns [19] what is not done in any of the algorithms itemized below. However, as most authors of such control algorithms use the term "pattern recognition" [18], [20]–[24], that is also done in this paper.

The first one who made pattern based adaptive control popular was Bristol in 1977 [20]. This controller ignores the dynamics of the process and is thus relatively slow.

Seif modeled the transients by a set of elementary patterns without giving details about the adaptation rules [21].

Seem proposed a control algorithm where the changes of the PI controller parameters are described as functions of two features of the closed-loop set-point or disturbance step response [22], [23]. Some details of Seem's algorithm need periodic information of the current process output and are therefore not suitable for nonuniform sampling, but the basic idea can be transferred to level-crossing sampling. Seem's algorithm is optimized to minimize IAE (integral of absolute error) instead of considering also sensor energy efficiency. Seem spent much attention on reaching robustness for many practically important special cases of HVAC control. The algorithm has been used successfully in more than 500,000 controllers [23].

Morilla et al. published a multi-step approach for PID controllers [18]. The results are good, but unfortunately some practically important aspects like disturbances and actuator saturation have not been considered. Some of their used features of the step-responses (e. g. decay ratio and oscillation period) are practically not measurable with level-crossing sampling and sensitive to disturbances and measurement noise.

INTUNE is a commercial adaptive controller using pattern recognition for updating the parameters of a PID controller [25]. To the authors' knowledge, the details of the adaptation rules have not been published.

Segovia et al. proposed a simple iterative pattern based PID controller scheme [24] roughly based on the Ziegler/Nichols tuning rule, but they gave only a tuning rule for oscillatory step responses. Furthermore, they did not write anything about actuator saturation or disturbances.

## III. Assumptions and objective

This paper deals with the control loop shown in Fig. 1 where $y_m(t)$ is the signal measured by the sensor (process output $y_p(t)$ with disturbance $d(t)$) and $y_c(t)$ the signal which is sent to the controller. $w(t)$ is the set-point or reference signal, $e_c(t) = w(t) - y_c(t)$ the control error and $u_c(t)$ the control signal (manipulated variable) which is sent to the actuator. The actuator uses a zero-order hold and its output is called $u_p(t)$. For ease of notation, a distinction between continuous-time and discrete-time signals has been avoided here.

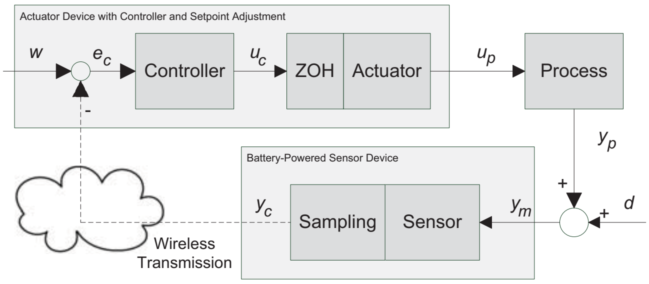


Fig. 1. Control loop with battery-powered sensor device.

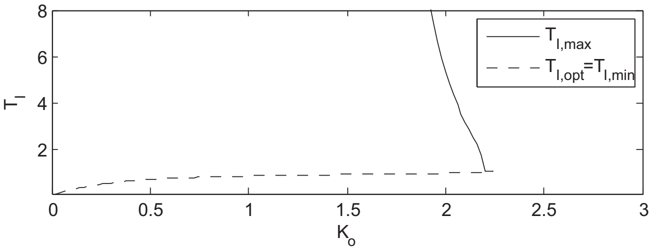


Fig. 2. Minimum (optimal) and maximum value of $T_I$ as a function of $K_o$ for avoiding oscillations for the example of $T_m = 1$ and $\tau = 0.2$, with continuous-time control.

In contrast to this control loop, some other authors integrate sensor and controller in one device and transmit the control signal to the actuator using level-crossing sampling [26]. The advantage of separating the sensor from the controller is that the sensor does not need a receiver or display (for editing the current set-point, weekly schedules or controller parameters) and can therefore save much energy. The actuator usually needs far more energy anyhow as well as it must power a receiver for getting the sensor messages and so the energy consumption of the controller and display is less critical.

For describing the algorithm, the process is assumed to be a FOLPD (first order lag plus time delay) process

$$G(s) = \frac{Y_p(s)}{U_p(s)} = \frac{K_m}{1+sT_m}e^{-s\tau} \tag{2}$$

with proportional action coefficient $K_m$, time constant $T_m$, and time delay (deadtime) $\tau$. The ratio

$$\eta := \frac{\tau}{T_m} \tag{3}$$

is called *degree of difficulty* [27]. Practical ranges of these parameters for room temperature processes can be found e. g. in [27] where one should keep in mind that variations could be larger, especially for $K_m$ caused by variations of flow (supply) temperature. Nevertheless, the proposed algorithm is not limited to this process type because it does not need to know or identify any time constant directly. The physics of rooms (room airflow) is such complex that it theoretically cannot be modeled adequately by a transfer function (or a differential equation) [28]. The proposed controller has to deal with these issues.

PI and PID controllers are the by far mostly used controllers in industry [29], [30]. The basic continuous-time PI controller is given by

$$R(s) = \frac{U_c(s)}{E_c(s)} = K_P\left(1+\frac{1}{T_I s}\right) \tag{4}$$

with proportional action coefficient $K_P$ and reset time $T_I$.

PI(D) controllers have also been used successfully in an event-based fashion [2], [7], [31]. The PI algorithm which is used in this paper is taken from [7], only enhanced by anti-reset windup. Derivative action has not been used. This will be explained later using results of the presented investigations.

A cost function which helps to design the adaptive controller is the product

$$J_{Prod} := N_{sc} \cdot J_{ISE} \tag{5}$$

of the number $N_{sc}$ of messages from the sensor to the controller and the control quality measured as Integral of Squared Error (ISE) $J_{ISE}$. The adaptation should reach small values of $J_{Prod}$ as this is a hint for a good trade-off between message number and control performance. More detailed reasons for this choice are discussed in [7].

## IV. Improved tuning rule

In [7] it has been shown that for minimzing $J_{Prod}$ at step responses, large oscillations should be avoided and—in an ideal case (no actuator saturation, FOLPD process)—a suitable tuning rule is

$$K_P = \frac{0.468}{K_m \cdot \eta}, \tag{6a}$$
$$T_I = T_m. \tag{6b}$$

But, typical challenges in room temperature control are actuator saturation, large time-variable (but not step-wise) disturbances, and processes of unknown order. As stated in [7], the tuning rule (6) is not optimal in these cases. Simulations confirm that ignoring these problems leads to poorly tuned control loops. This section thus presents some improvements over tuning rule (6).

### A. Oscillating step responses

As the goal is to avoid oscillating step responses, it is interesting to know for which $K_P$ that is possible at all. If the open-loop gain $K_o$, which is defined as

$$K_o := K_P \cdot K_m, \tag{7}$$

is greater than a limit $K_{o,max}(\eta)$, there are oscillations, independent of the selection of the reset time $T_I$, see an example in Fig. 2. Fig. 3 shows the maximum $K_{o,max}$ as a function of $\eta$, found by simulations. The simulation uses continuous-time control to reach idealized results, what is similar to an infinitesimally small threshold $\Delta_{lc}$. For numerical reasons a deadband around the set-point should be defined in which the control loop is allowed to oscillate for avoiding too conservative results. This has been set to 1 % of the step width.

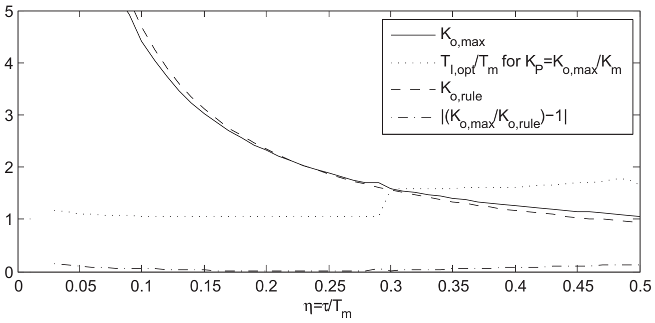


Fig. 3. $K_{o,max}$ as a function of $\eta$ for a deadband of 1 % and the optimal setting of $T_I$, normalized to $T_m$. Also $K_{o,rule}$ according to tuning rule (6) is shown as well as the ratio of $K_{o,max}$ and $K_{o,rule}$.

The conclusion for avoiding oscillating step responses is

$$K_P \overset{!}{\leq} \frac{K_{o,max}(\eta)}{K_m}. \tag{8}$$

$K_P$ according to (6a) lies near to the curve for a deadband of 1 % (see Fig. 3) and gives therefore nearly the fastest response with no more than 1 % overshoot.

### B. Actuator saturation

The immediate change of the controlled variable $u_p(t)$ caused by proportional action after a stepwise reference change $\Delta w$ can be calculated by

$$\Delta u_{prop} = K_P \cdot \Delta w. \tag{9}$$

Since in real applications $u_p(t)$ is bounded between 0 and 1 (valve fully closed to valve fully opened), it is useless to set $K_P$ higher than $1/\Delta w$ because of actuator saturation. That means

$$K_P \overset{!}{\leq} \frac{1}{|\Delta w|}. \tag{10}$$

An illustrative example: Assumed that the reference change for night setback is $\Delta w = -5\,\mathrm{K}$. So, each $K_P > 0.2$ would result in immediate actuator saturation. Of course, increasing $K_P$ does not degrade control performance at step responses (because it has little influence if the stability limit is not exceeded), but high $K_P$ will lead to many messages in "steady state" without improving control performance as will be discussed in section VII.

Note that the fulfillment of (10) does not guarantee that there is no actuator saturation, because that depends on the value of $u_p$ before the set-point change as well as on integral action.

Besides, derivative action (a PID controller) increases actuator action and hence also actuator saturation. Thus, it is at step responses reasonable to do without derivative action.

### C. New tuning rule

The proposed tuning rule for $K_P$ is the combination of (8) and (10)

$$K_P = \min\left(\frac{1}{|\Delta w|}, \frac{K_{o,max}(\eta)}{K_m}\right) \tag{11}$$

As (6a) approximates $K_{o,max}(\eta)/K_m$ quite well, the rule can be approximated by

$$K_P \approx \min\left(\frac{1}{|\Delta w|}, \frac{0.468}{K_m \cdot \eta}\right). \tag{12}$$

No rule for setting $T_I$ is given here because iterative optimization is used in the proposed adaptive controller.

## V. Proposed adaptation strategy

This section presents the new adaptation strategy.

### A. Used patterns

Only reference step responses are taken to analyze the process behavior. In a typical office building, reference changes occur twice a day: Once in the morning and once in the evening because of night setback. In residential buildings, there are often four reference changes because the set-point is higher only in the morning and in the evening. So, there are at least two step responses a day at which the controller parameters can be optimized.

The precondition for benefiting from an adaptive controller is that the process does not change faster than the adaptation can follow. Faster changes are regarded as belonging to the disturbances. However, the sources of process changes which have been itemized in section I do not change significantly during one day (or they change fast but only seldom, like on refurbishment). Only considering the variation of the flow temperature may be advantageous, what could be done via gain scheduling [3], [32].

Some other authors used also step-wise disturbances for updating the controller parameters [22], [24]. However, since significant step-wise load changes do not often occur in practical room temperature control, only set-point changes are considered here.

### B. Proportional action coefficient $K_P$

As

$$K_m = G(0) = \frac{Y_p(0)}{U_p(0)}, \tag{13}$$

and assuming a constant disturbance $d$, $K_m$ can be calculated from two pairs ($u_p[k]$, $y_p[k] + d$) measured in "steady state" with different values of $u_p[k]$. So, after each closed-loop step response, when the process is again in "steady state", $K_m$ is estimated by

$$K_m = \frac{(y_p(t_s) + d) - (y_p(t_0) + d)}{u_p(t_s) - u_p(t_0)} \approx \frac{y_c(t_s) - y_c(t_0)}{u_c(t_s) - u_c(t_0)} \tag{14}$$

where $t_0$ is the step time (or shortly before) and $t_s$ is the time when steady state is (assumed to be) reached after the step response. This method is relatively robust to measurement noise but in the case of slowly varying loads $d(t)$ the estimation gets inaccurate. If $(t_s - t_0)$ is chosen too small, the "steady state" may not yet have been reached, leading to more inaccurate estimation of $K_m$. Contrary, if $(t_s - t_0)$ is chosen too large, load changes can get more influence on the results. In the simulations of section VI $(t_s - t_0)$ has been set to one hour.

After estimating $K_m$, (12) is applied to compute $K_P$. Unfortunately, $\eta$ depends on $\tau$ and $T_m$ which are hard to estimate in closed loop under noisy conditions with time-variable load and significant threshold $\Delta_{lc}$. Instead, an upper limit $\eta_{max}$ can be used which is the greatest degree of difficulty which is expected to occur. According to [27], for room temperature control $\eta_{max}$ is 0.3. If $\eta$ of the real process is lower than $\eta_{max}$, $K_P$ is lower than necessary, but the reset time $T_I$ will be adjusted (reduced) to improve the control loop performance, see section V-C. The authors are working on more sophisticated solutions, but simulations show that this simple solution works well, too.

### C. Reset time $T_I$

As announced in section IV-C, an iterative method is used for updating $T_I$, i. e. the reset time is updated after each set-point change based on the measured overshoot. The overshoot $h_r$ is a monotonically decreasing function of $T_I$, see some examples in Fig. 4(a). This is intuitively clear because the higher the reset time $T_I$, the slower is the response and the less the set-point is exceeded before the controller can react on the overshoot. The basic idea for adaptation is:

- If there is (too much) overshoot, increase the reset time.
- If there is no (or too little) overshoot, decrease the reset time.

Because of (12) it is guaranteed that there is a setting of $T_I$ without oscillations (in the case of an ideal FOLPD process, continuous-time control and no disturbances).

The step width for updating $T_I$ must be defined. Simple learning algorithms (similar to first order low-passes) could be used, comparable to [13], [22], [24], but it is possible to apply more specific algorithms. An initial attempt is given in the following; more sophisticated solutions as well as theoretical analysis are part of current research of the authors.

Let $T_{I,opt}$ define the smallest possible reset time without overshoot. Fig. 4(a) shows some examples for the overshoot $h_r$ as a function of $T_I/T_{I,opt}$ using continuous-time control. The qualitative curves of Fig. 4(b) substantiate the assumption that the overshoot can be approximated by

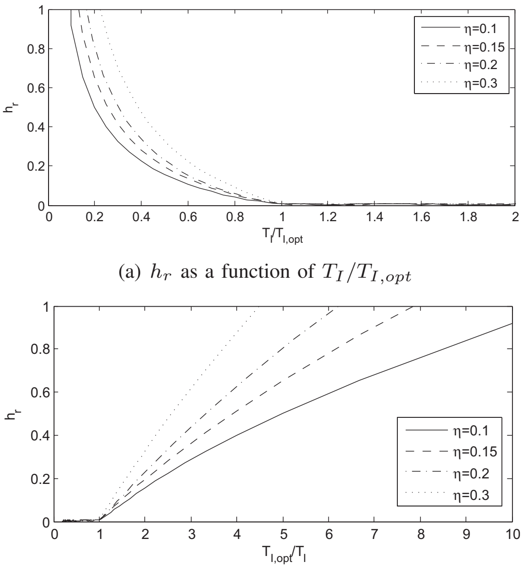
$$h_r \approx \begin{cases} \gamma \cdot \left(\frac{T_{I,opt}}{T_I} - 1\right), & \text{if } T_I < T_{I,opt} \\ 0, & \text{otherwise,} \end{cases} \tag{15}$$

The proportionality constant $\gamma$ is discussed later.

Equation (15) can be used to calculate $T_{I,opt}$ based on the measured overshoot $h_{r,old} > 0$ and the appropriate reset time $T_{I,old}$:

$$T_{I,new} = T_{I,opt} = \left(1 + \frac{h_{r,old}}{\gamma}\right) \cdot T_{I,old}. \tag{16}$$

If there is no overshoot, it is only known that the reset time should be reduced, but not how much. However, from (15) a maximum step width can be calculated with the requirement that the overshoot after the update must not exceed a given limit $h_{r,max}$. The important case is $T_{I,old} = T_{I,opt}$, because a reduction of $T_I$ will then result in the highest overshoot. The rule for updating is



(a) $h_r$ as a function of $T_I/T_{I,opt}$

(b) $h_r$ as a function of $T_{I,opt}/T_I$

Fig. 4. Overshoot $h_r$ as a function of reset time $T_I$, normalized to $T_{I,opt}$ (and inverse) for several $\eta$ and $K_o = 1.5$.

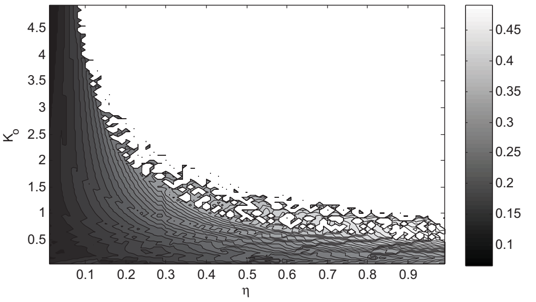
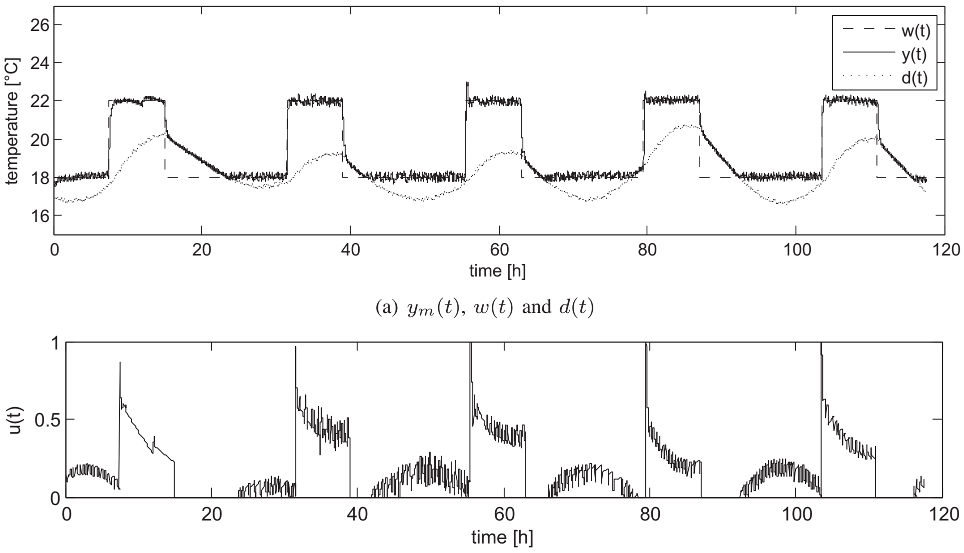


Fig. 5. $\gamma$ as a function of $\eta$ and $K_o$ for idealized conditions: Continuous-time control, FOLPD process, no disturbances, no actuator saturation.

$$T_{I,new} = \frac{1}{1 + \frac{h_{r,max}}{\gamma}} \cdot T_{I,old}. \tag{17}$$

The greater $h_{r,max}$ is chosen, the faster the algorithm converges to less conservative settings. $h_{r,max}$ can be reduced after finding a rough estimation of $T_{I,opt}$, e. g. after the first step-response with overshoot happened.

The remaining task is to find the appropriate value of $\gamma$. This parameter depends on $\eta$, $K_o$, $\Delta_{lc}$, the real process order, and actuator limits. Additionally, the measured overshoot $h_{r,old}$ can be adulterated with disturbances and other effects, in particular when using level-crossing sampling where the exact overshoot cannot be measured. Similarly to [22] a low-pass filter could be used for reducing influences of disturbances, but filter design is difficult if the process parameters can vary over a wide range, especially with nonuniform sampling. Fortunately, it is not very important to know an exact value



(a) $y_m(t)$, $w(t)$ and $d(t)$

(b) $u_p(t)$

Fig. 6. Simulation over five days with level-crossing sampling and measurement noise.

of $\gamma$ because of the iterative method—if $\gamma$ is chosen too high, only the convergence rate is degraded. Thus, a worst case estimation is enough. Fig. 5 shows $\gamma$ as a function of $K_o$ and $\eta$ under the idealized assumptions of continuous-time control, a FOLPD process, no disturbances and no actuator saturation. According to this graphic and assuming that in room temperature control $\eta < 0.3$ holds [27], the worst case to be expected is $\gamma \approx 0.3$. This value has also been found to work well in simulations.

Step responses with too little overshoot *and* actuator saturation must not be used for updating $T_I$, because the too little overshoot may result from actuator saturation and not from sluggish tuning. This is no problem, since the controller knows whether the actuator limits (0 or 1) have been reached in the last step response.

This iterative method allows adaptation without estimating $\tau$ and $T_m$ what would be difficult with level-crossing sampling and disturbances. It also improves the tuning results in the case of unknown process order where (6) leads to suboptimal results [7].

### D. Initial settings

The initial settings of $K_P$ and $T_I$ should be based on the most critical process to be expected. This is the maximum of each parameter $\eta$, $K_m$, and $\tau$, i. e. for room temperature control according to [27] $\eta = 0.3$, $\tau = 0.05\,\mathrm{h}$, and $K_m = 10\,\mathrm{K}$. $K_P$ can then be set using (12) and (with regard to [7]) $T_I$ to $T_m$, i. e. $T_I = 0.5\,\mathrm{h}$.

In result, the first step response is stable. In most cases (all but the most critical case) the response will be too sluggish. After each step response, $K_P$ is updated using (12) and (14) as well as $T_I$ is reduced according to (16). The responses get faster until overshoots occur. Then $T_I$ is increased using (17).

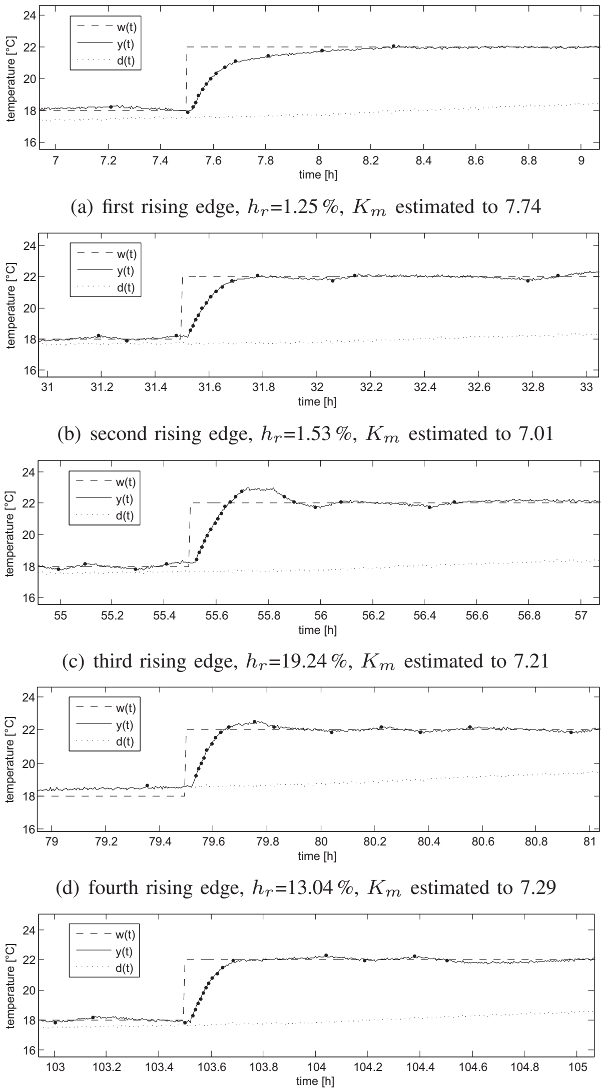
## VI. Simulations

Fig. 6 presents a simulation over five days using a process with $T_m = 0.15\,\mathrm{h}$, $K_m = 7\,\mathrm{K}$, and $\tau = 0.01\,\mathrm{h}$. The disturbances (this is the room temperature without heating, i. e. $u_p(t) \equiv 0$) are based on measurements taken from an office building which was built in 2005, having a heat energy consumption of 34 kWh/m$^2$a. By setting $\Delta_{lc}$ to 0.3 K (see section VII) and applying night setback of 4 K, an overshoot of 7.5 % (0.3K/4K) must be accepted because of level-crossing sampling. The internal sampling period of the sensor is 0.005 h.

As this is not the "most critical" of expectable processes, the first step response is slower than necessary, shown detailed in Fig. 7(a). Fighting against that, the adaptation algorithm estimates $K_m$, updates $K_P$ and reduces $T_I$. The negative set-point changes cannot be used for pattern recognition, because actuator saturation without overshoot occurs (the valve is fully closed). The second rising edge is significantly faster than the first, Fig. 7(b). The third even has (too much) overshoot, Fig. 7(c). So, the adaptation algorithm increases the reset time. The fourth step response has less overshoot, Fig. 7(d), but due to the inexact adaptation method, also this step response has too much overshoot and the reset time is further increased. The fifth set-point change is as desired.

## VII. Disturbance compensation

The tuning rules (6) and (12) are exclusively based on step responses. But, it is interesting how the parameters should be chosen in "steady state" when only disturbances have to be compensated. To the authors knowledge, until now only limit cycles have been studied in several publications (e. g. [2], [33]). Limit cycles are periodic movements of the controlled variable between two or more sampling levels which often occur in "steady state", especially because of integral action,



(a) first rising edge, $h_r$=1.25 %, $K_m$ estimated to 7.74

(b) second rising edge, $h_r$=1.53 %, $K_m$ estimated to 7.01

(c) third rising edge, $h_r$=19.24 %, $K_m$ estimated to 7.21

(d) fourth rising edge, $h_r$=13.04 %, $K_m$ estimated to 7.29

(e) fifth rising edge, $h_r$=7.12 %, $K_m$ estimated to 9.53

Fig. 7. The five first rising edge step responses with level-crossing sampling and measurement noise. The dots are sampling instants, i. e. level crossings.

and that results in an unpleasant high message rate without improving control performance. But, if disturbances lead to level crossings (and therefore messages) anyhow, the importance of avoiding limit cycles is reduced.

The disturbance is equal the room temperature without heating or cooling, i. e. $u_p(t) \equiv 0$. The main reasons for disturbances are time-varying heat flows to or from the outside which depend on the outside temperature, sunlight, and room utilization (how many persons and machines are creating heat). These influences do not lead to any step-wise changes of the room temperature. It is well known from experience and measured in the office building considered in section VI that without heating the room temperature is a slowly varying signal, usually periodic with the minimum after midnight and the maximum after noon. The amplitude depends on the outside temperature, sunlight, and room utilization and does thus change from day to day. In this section, for theoretical investigations disturbances are assumed to be of the form

$$d(t) = A \cdot \sin\left(\frac{t}{24\,\mathrm{h}}\right), \tag{18}$$

that means a sinus curve with a period of one day and an amplitude $A$.

Simulations show that the PI controller parameter settings minimizing (5) depend on the ratio $a$ of the disturbance amplitude $A$ and the threshold $\Delta_{lc}$ which is defined as

$$a := \frac{A}{\Delta_{lc}}. \tag{19}$$

The practical bounds of $a$ are of interest. Since humans usually do not feel temperature changes smaller than 0.3 K there is no need to set $\Delta_{lc}$ significantly smaller than this value—even if it would allow a smaller value for the cost function (5) because of smaller ISE, the occupants would not notice it while the message rate (and therefore energy consumption) would be unnecessarily high, and also the costs for a sensor measuring such exactly would be high. Besides, the temperature variations inside one room because of stratification and bordering spaces are significantly greater than 0.3 K [28]. The amplitude $A$ can be expected to be lower than 5 K (very old buildings with poor insulation), for modern, well insulated buildings lower than 1.5 K. So, $a$ can be expected to be lower than 17 (5 K/0.3 K), often lower than 5 (1.5 K/0.3 K).

Simulations show that for such small values of $a$ the cost function (5) can be optimized by considerably reducing $K_P$ compared to the value got by (6a) because reducing $K_P$ increases the period of limit cycles and hence decreases the message number and energy consumption. As the temperature set-point trajectory correlates roughly with the disturbance (at noon higher than at midnight) the necessary control action is further reduced.

This fact can be used to improve the tuning rule in "steady state" by simply adding a factor $\kappa$, getting

$$K_P = \frac{0.468 \cdot \kappa}{K_m \cdot \eta}. \tag{20}$$

Optimal values of $\kappa$ as a function of $a$ and $\eta$ found by simulations can be seen in Fig. 8. The optimal $\kappa$ rises with $a$ because for higher $a$ a higher $K_P$ can improve the ISE more significantly than for lower values of $a$.

$\kappa$ could be realized in the controller by using a second degree of freedom (set-point weighting).

Note that using derivative action of a PID controller would lead to stronger actuator action and faster oscillations (more messages) without improving control performance.

Unfortunately, since $a$ and $\eta$ are neither known a priori nor estimated on-line, the formula cannot be applied. Automatically finding values of $\kappa$ suitable for each controlled room is another part of current research of the authors. Additional sensors, e. g. for outdoor temperature, room occupancy, and illumination, can help estimating $a$. Weather forecasts promise even better results. However, these solutions are much more

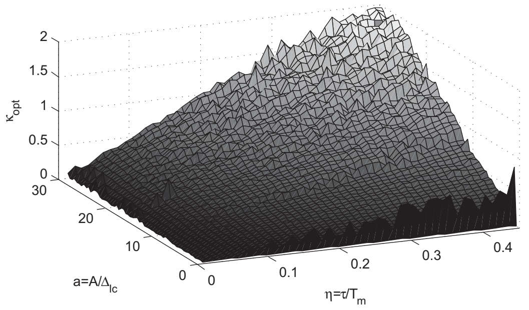


Fig. 8. Optimal setting of $\kappa$ in (20) as a function of degree of difficulty $\eta$ and disturbance to threshold ratio $a$. Simulation parameters are $T_m = 0.3\,\mathrm{h}$, $K_m = 10$, $T_I = T_m$ and $\Delta_{lc} = 0.3\,\mathrm{K}$. Minimum inter sample intervals [7] and actuator saturation have not been taken into account.

expensive than a simple single-loop controller like the proposed one.

## VIII. Conclusion

A pattern-based adaptive controller designed for level-crossing sampling and room temperature control has been presented. Simulations show that the algorithm can deal with typical problems of this kind of control loop. Also several points for future research have been pointed out, including practical improvements, theoretical analysis and less conservative assumptions.

## Acknowledgment

This work was funded by the European Commission, the European Social Funds and the Sächsische Aufbaubank in the junior research group ZESSY.

## References

[1] R. Zheng, J. C. Hou, and N. Li, “Power management and power control in wireless networks,” in *Ad Hoc and Sensor Networks*. Nova Science Publishers, 2004, pp. 1–25.

[2] J. Ploennigs, V. Vasyutynskyy, and K. Kabitzsch, “Comparative study of energy-efficient sampling approaches for wireless control networks,” *IEEE Transactions on Industrial Informatics*, vol. 6, no. 3, pp. 416–424, August 2010.

[3] A. Pawlowski, J. L. Guzmán, F. Rodríguez, M. Berenguel, J. Sánchez, and S. Dormido, *Factory Automation*. InTech, March 2010, ch. Study of event-based sampling techniques and their influence on greenhouse climate control with Wireless Sensor Networks, pp. 289–312.

[4] S. Dormido, L. Grau, J. L. Fdez-Marron, M. A. Canto, and J. M. de la Cruz, “A PID autotuner based on the Åström and Hägglund method’s using non-uniform sampling,” in *Proceedings of the 6th Mediterranean Electrotechnical Conference*, vol. 2, Ljubljana, Slovenia, May 1991, pp. 860–863.

[5] X. Wang and N. Hovakimyan, “$\mathcal{L}_1$ adaptive control of event-triggered networked systems,” in *Proceedings of the 2010 American Control Conference*, Marriott Waterfront, Baltimore, MD, USA, June–July 2010, pp. 2458–2463.

[6] K. J. Åström and B. Wittenmark, *Adaptive Control*. Reading: Addison-Wesley Publishing Company, 1989.

[7] B. Hensel, J. Ploennigs, V. Vasyutynskyy, and K. Kabitzsch, “A simple PI controller tuning rule for sensor energy efficiency with level-crossing sampling,” in *Proceedings of the 9th International Multi-Conference on Systems, Signals & Devices*, Chemnitz, Germany, March 2012.

[8] J. Gillberg and L. Ljung, “Frequency domain identification of continuous-time ARMA models: Interpolation and non-uniform sampling,” Linköpings universitet, Linköping, Sweden, Tech. Rep., September 2004.

[9] E. Müller, H. Nobach, and C. Tropea, “Model parameter estimation from non-equidistant sampled data sets at low data rates,” *The measurement of scientific and technological activities*, vol. 9, pp. 435–441, 1998.

[10] S. Ahmed, B. Huang, and S. L. Shah, *Identification of continuous-time models from Sampled Data*. Springer London, March 2008, ch. Process Parameter and Delay Estimation from Non-uniformly Sampled Data, pp. 313–337.

[11] V. J. VanDoren, *Techniques for Adaptive Control*. Amsterdam: Butterworth-Heinemann, 2003, ch. Introduction, pp. 1–21.

[12] B. D. O. Anderson, “Adaptive systems, lack of persistency of excitation and bursting phenomena,” *Automatica*, vol. 21, no. 3, pp. 247–258, 1985.

[13] Z.-G. Han and X. Yu, “An adaptive model free control design and its applications,” in *Proceedings of the 2nd International Conference on Industrial Informatics*, Berlin, Germany, June 2004, pp. 243–248.

[14] Y. Ma, X. Chen, and X. xiaohua, “A novel model free adaptive controller with tracking differentiator,” in *Proceedings of the 2009 IEEE International Conference on Mechatronics and Automation*, Changchun, China, August 2009, pp. 4191–4196.

[15] Z. Hou and S. Jin, “Data-driven model-free adaptive control for a class of MIMO nonlinear discrete-time systems,” *IEEE Transactions on Neural Networks*, vol. 22, no. 12, pp. 2173–2188, December 2011.

[16] G. S. Cheng, *Techniques for Adaptive Control*. Amsterdam: Butterworth-Heinemann, 2003, ch. Model-Free Adaptive Control with CyboCon, pp. 145–202.

[17] X. Aidong, Z. Yangbo, S. yan, and L. Mingzhe, “An improved model free adaptive control algorithm,” in *Proceedings of the Fifth International Conference on Natural Computation*, Tianjian, China, August 2009, pp. 70–74.

[18] F. Morilla, A. González, and N. Duro, “Auto-tuning PID controllers in terms of relative damping,” in *Proceedings of the IFAC Workshp on Digital Control*, Terrassa, Spain, April 2000.

[19] A. K. Jain, R. P. Duin, and J. Mao, “Statistical pattern recognition: A review,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 1, pp. 4–37, January 2000.

[20] E. H. Bristol, “Pattern recognition: An alternative to parameter identification in adaptive control,” *Automatica*, vol. 13, pp. 197–202, 1977.

[21] A. A. Seif, “On the adaptive pattern recognition control,” in *Proceedings of the IEEE Conference on Computer Systems and Software Engineering*, May 1992, pp. 706–709.

[22] J. E. Seem, “A new pattern recognition adaptive controller with application to HVAC systems,” *Automatica*, vol. 34, no. 8, pp. 969–982, 1998.

[23] ——, “An improved pattern recognition adaptive controller,” in *Proceedings of IFAC Conference on Energy Saving Control in Plants and Buildings*, Bansko, Bulgaria, October 2006, pp. 117–122.

[24] J. P. Segovia, D. Sbarbaro, and E. Ceballos, “An adaptive pattern based nonlinear PID controller,” *ISE Transactions*, vol. 43, pp. 271–281, 2004.

[25] T.-L. Chia and I. Lefkowitz, *Techniques for Adaptive Control*. Amsterdam: Butterworth-Heinemann, 2003, ch. Expert-Based Adaptive Control: ControlSoft’s INTUNE Adaptive and Diagnostic Software, pp. 203–232.

[26] M. Beschi, S. Dormido, J. Sánchez, and A. Visioli, “On the stability of an event-based PI controller for FOPDT processes,” in *Proceedings of the IFAC Conference on Advances in PID Control*, Brescia, Italy, March 2012.

[27] Control technology. Siemens Switzerland Ltd, Building Technology Group. Technical brochure, Reference number 0-91913-en.

[28] P. Riederer, *Thermal room modelling adapted to the test of HVAC control systems*, 2002, PhD thesis.

[29] K. J. Åström and T. Hägglund, “The future of PID control,” *Control Engineering Practice*, vol. 9, pp. 1163–1175, 2001.

[30] A. O’Dwyer, *Handbook of PI and PID Controller Tuning Rules*, 2nd ed. London: Imperial College Press, 2006.

[31] K. E. Årzén, “A simple event-based PID controller,” in *Preprints 14th World Congress of IFAC*, January 1999.

[32] F. Rodríguez, *Modeling and hierarchical control of greenhouse crop production (in Spanish)*. University of Almería, Spain, 2002, PhD thesis.

[33] M. Beschi, A. Visioli, S. Dormido, and J. Sánchez, “On the presence of equilibrium points in PI control systems with send-on-delta sampling,” in *Proceedings of the 50th IEEE Conference on Decision and Control and European Control Conference*, Orlando, FL, USA, December 2011, pp. 7843–7848.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Fault Detection with MAC Delay Compensation in Wireless Sensor Actuator Networks

Xuewu Dai*, Yang Yang[†], Yu Zhang[‡], Zhiwei Gao[§] and Hong Wang[‖]
* School of Electronic & Information Engineering, Southwest University, Chongqing, 400715, China
Email: daixuewu@swu.edu.cn
[†] Shanghai Institute of Microsystem and Information Technology, Shanghai Research Center for Wireless Communications, Chinese Academy of Sciences, Shanghai 200335, China Email: yang.yang@shrcwc.org
[‡]National Engineering Research Center for Broadband Networks and Applications, Shanghai, China
[§]School of Computing, Engineering and Information Sciences, Northumbria University, Newcastle NE2 1XE, UK
Email: zhiwei.gao@northumbria.ac.uk
[‖]Control Systems Centre,University of Manchester, Manchester, M60 1QD, UK Email: hong.wang@manchester.ac.uk

*Abstract*—**Although the Wireless Sensor and Actuator Networks (WSANs) have many advantages than the wired networks, the nature of sharing wireless media and the complicated behavior of Media Access Control (MAC) introduce adverse impacts on the control system. In contrast to most work on networked control systems using simplified models of network induced delays, this paper considers the study the random access delay caused by a contention-based MAC scheme slotted ALOHA in WSANs. An improved fault detection observer with access delay compensation is proposed to improve the fault detection performance against the MAC delays.**

## I. Introduction

As one of the main challenges in Wireless Sensor Actuator Networks (WSANs) and Networked Control Systems (NCSs), the network-induced delay degrades the fault detection performance [1], [2],[3], [4] and much of attention have been paid on designing a fault detection system robust to network-induced delays [5], [6]. In most existing studies, the characteristics of a communication network are usually represented by simplified analytical models without considering the details of different communication protocols. For instance, it is popular to adopt a finite state Markov chain to represent the dynamics of the network-induced delays $\{\tau_k\}$, without consideration of the complicated behaviours of various communication protocols. As a result, the control system is modelled as a Markov Jumping System (MJS) [7] and various methods were proposed to design the fault detection filters, including Riccati equation methods [3] and linear matrix inequalities (LMIs) for $H_\infty$ fault detection observer design. In [8], [3], a Takagi-Sugeno (T-S) is built to represent network-induced delays and a fuzzy fault detection observer is proposed. [9] models the delay as a stochastic process with known mean and variance, and treats the impacts of delay as a parameter change in the framework of stochastic systems. Paper [10] used the Taylor approximation to analysis the impacts of delay and proposed a parity space-based residual generator. Eigen-decomposition and the Pade approximation were used in [11] under the assumption that the state matrix is diagonal.

However, the communication network is a complicated dynamic system and the simplified model of network delays cannot represent the communication network behaviour very well. The network-induced delays are caused by many reasons and there have been studies on the estimate the packet delays in a network [12]. For instance, paper [13] adopted the network calculus theory to determine the bounds of delays and adjust the threshold of fault detection observer's residual accordingly. In wireless networks, the contention-based MAC schemes are widely adopted to support multiple access. Unfortunately, it is still not clear how a random MAC scheme and its associated delay statistics influence the FD performance of NCSs. To bridge this gap between the communication protocols and fault detection observer designs, this paper brings the statistical properties of the access time of a contention-based MAC scheme (the so-called slotted-ALOHA) into the FD observer design. The main contribution of this paper are, (1) According to the analysis of the slotted ALOHA MAC scheme and its associated access delay (namely MAC-delay), a statistical estimation of the MAC-delay in slotted ALOHA induced is proposed. (2) With the estimate of MAC-delays, a new FD observer is proposed to compensate the MAC-delay such that the adverse impacts of the delays are attenuated and the fault detection performance is improved.

The rest of this paper is organized as follows: Section II is an introduction to a industrial WSAN. The slotted ALOHA is analyzed and its MAC-delay is estimated in Section III, followed by the proposal of a novel FD observer design with MAC-delay compensation in Section IV. The fault detection performance is then optimised in Section V. Finally, Section VI evaluates the performance of the proposed FD observer on a MATLAB/OMNeT hybrid simulation platform.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

## II. System Model and Problem Formulation

As shown in Fig. 1, this paper considers a WSAN consisting of a set of distributed sensors and controllers organized in a group of control loops, an access point (AP) linking together the wireless network and the wired Ethernet and a Fault Detection (FD) system. All controllers and sensors are connected through the wireless network with a star topology, and the wireless network takes two non-overlapped radio channels, namely uplink and downlink channels, in 2.4G ISM band.

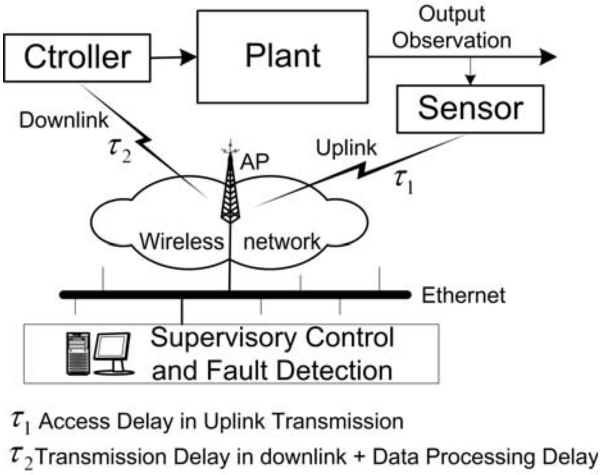

Ctroller
Plant
Output Observation
Sensor
Downlink
$\tau_2$
AP
Uplink
$\tau_1$
Wireless network
Ethernet
Supervisory Control and Fault Detection
$\tau_1$ Access Delay in Uplink Transmission
$\tau_2$ Transmission Delay in downlink + Data Processing Delay

Fig. 1. A wireless sensor actuator network for supervisory control

Consider a plant working at some operation point

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) + B_f f(t) + B_d d(t) \\ y(t) = Cx(t) + v(t) \end{cases} \tag{1}$$

where $x(t) \in \mathbb{R}^n$ is the state variable of the plant, $u(t) \in \mathbb{R}^m$ the input, $y(t) \in \mathbb{R}^p$ the output. $v(t)$ and $d(t)$ are unknown inputs with appropriate dimensions. $f(t) \in \mathbb{R}^q$ is the fault to be detected. $A, B, C, B_f$ and $B_d$ are known constant matrices with appropriate dimensions, and $C$ is of full row rank.

The output $y_k$ measured by the sensor at sampling interval $T_s$ is sent to the AP through the uplink channel by using the slotted ALOHA media access protocol [12]. Once receiving a packet of $y_k$, the AP forwards it to the corresponding controller (via downlink channel) and to the remote FD system (via Ethernet). And the controllers reply with its control signal $u_k$ to the AP. When a packet of $u_k$ arrives from a controller through the downlink channel, the AP forwards it to the FD system only. The sensors $S_i$ are time-driven with local sampling interval $T_s$ and the controllers $Ctr_i$ is event-driven.

*Remark 1* The uplink is shared by all sensors, and a sensor has to compete for the use of the uplink channel. Several sensors may send their packets at the same time, which results in a 'collision'. Thus the transmission fails, re-transmission has to be scheduled and some delay is introduced. As one of the widely used random access scheme, a slotted ALOHA protocol is considered in this paper. In the downlink communication, however, the access to the downlink is fully coordinated by the AP, there is no competition for the use of downlink and the collision is avoided.

It is further assumed that the disturbances and fault components have slow dynamics $f(t) = f_k$, $d(t) = d_k$ and $v(t) = v_k$ for all $t \in [kT_s, (k+1)T_s]$. Similar to the discretization of dynamic system, at the $k$-th sampling instants, an equivalent discrete system of (1) can be represented by [10][14]

$$\begin{cases} x_{k+1} = \bar{A}x_k + B_0(\tau_k)u_k + B_1(\tau_k)u_{k-1} + \bar{B}_f f_k + \bar{B}_d d_k \\ y_k = Cx_k + v_k \end{cases} \tag{2}$$

where $\bar{A}$, $B_0(\tau_k)$, $B_1(\tau_k)$, $\bar{B}_f$ and $\bar{B}_d$ are defined in a similar way as [10].

Let

$$\bar{B} = \int_0^{T_s} e^{At} B dt = B_0(\tau_k) + B_1(\tau_k), \tag{3}$$

and

$$\Delta u_k = u_k - u_{k-1}, \tag{4}$$

system (2) can be rewritten under the following form:

$$\begin{cases} x_{k+1} = \bar{A}x_k + \bar{B}u_k - B_1(\tau_k)\triangle u_k + \bar{B}_f f_k + \bar{B}_d d_k \\ y_k = Cx_k + v_k \end{cases} \tag{5}$$

Note that $B_1(\tau_k) \in \mathbb{R}^{n \times m}$ is a matrix of functions with respect to $\tau_k$. Due to the fact that $\tau_k$ is a stochastic process determined by the communication network behaviors, the input matrix $B_1(\tau_k)$ varies at different time instants and the system under consideration actually behaves as a stochastic system,

## III. MAC Delay Analysis and Estimation

The network-induced delay $\tau_k$ is the delay relating the sensor sending out $y_k$ to the controller receiving it. Due to the nature of media sharing in wireless communication, if two or more devices sends their packet to the radio media during the transmission, a collision occurs, the receiver cannot receive the packet successfully, and the packets have to be re-transmitted. Thus a significant delay is induced. Such a delay is refereed to as 'MAC delay' or 'access delay'. In this section, it is our main interests to study the characteristics of MAC delay, and to estimate its mean value.

Let $\tau_1$ denote the MAC-delay in the uplink, $\tau_1$ is a stochastic variable and can be decomposed into two parts:

$$\tau_1 = E(\tau_1) + \Delta\tau_1 \tag{6}$$

where $E(\tau_1)$ is the mean value of the MAC-delay, and $\Delta\tau_1$ represents the jitter (the variation in the time between packets arriving at the AP). Generally, $\Delta\tau_1$ is a stochastic process with zero mean and variance $var(\Delta\tau_1) = \sigma^2$. Considering the reality of canceling re-transmissions in case of a too many re-transmissions, $\Delta\tau_1$ will be bounded.

Since the access to the downlink is fully coordinated by the AP, the delay from the AP to controllers consists of the transmission delay and data processing delay only. Let $\tau_2$ denote the downlink delay from the AP to controllers, $\tau_2$ is a deterministic constant value.

$$\tau_2 \approx T \tag{7}$$

where $T$ is the transmission time of a packet. Here, we have implicitly assumed that the the sum of processing time at the AP and corresponding controller is smaller and can be ignored. When this is not true, only a fixed constant term needs to be added to the final delay expression.

Let $\tau_k$ denote the network-induced delay relating the sensor to the actuator during sampling period $[kT_s, (k+1)T_s)$, $\tau_k$ can be expressed as

$$\tau_k \approx \tau_1 + T = T + E(\tau_1) + \Delta\tau_1 \tag{8}$$

and the mean value of the network-induced delay is

$$\bar{\tau} = E(\tau_k) \approx T + E(\tau_1) \tag{9}$$

In the followings, the focus is on the analysis of the slotted ALOHA scheme and the estimation of the mean value of access delay $\tau_1$ in the uplink channel. Borrowing the models and notations of the slotted ALOHA scheme in [12], the slotted ALOHA MAC protocol is summarized as follows:

1) When a sensor generates a new packet $y_k$, it accesses the uplink channel at the beginning of the next slot. This is called immediate-first transmission (IFT). Let $D_0$ be the access delay from the sampling time to the end of the initial transmission.
2) All packets are of the same size with transmission time $T$ and the length of a time slot is equal to $T$.
3) For a given sampling interval $T_s$, the total number of available slots for packet re-transmission $n$ is determined by:
$$n = \lfloor \frac{T_s}{T} \rfloor - 3 \tag{10}$$
4) A retransmission is scheduled after a random *backoff delay*, when a packet transmission fails. Let $W_i$ be the $i$-th backoff delay in unit of slots, that is refereed to as *backoff window size*. Then, the $i$-th retransmission takes place at the beginning of the $W_i$-th available slot. Note that $W_i \geqslant 1$. In this paper, a *uniform backoff* (UB) policy is adopted. That is all $W_i$'s are uniformly distributed in the same range , say $[1, w]$. The statistics of $W_i$ under UB policy are
$$\begin{aligned} P\{W_i = j\} &= \frac{1}{w}, \quad j = 1, 2, \ldots, w && (11)\\ E[W_i] &= \frac{1+w}{2} && (12)\\ \text{var}[W_i] &= \frac{[w^2-1]}{12} && (13) \end{aligned}$$
5) Let $R$ denote the number of retransmission needed for a successful packet transmission, and $D_i$ be the delay time due to the $i$-th unsuccessful transmission.

Thus, the access delay $\tau_1$ in the uplink is the time duration from its generation to the moment it is successfully transmitted to the AP, that is

$$\tau_1 = \sum_{i=0}^{R} D_i \tag{14}$$

and

$$D_i = (W_i + 1)T, \qquad i = 1, 2, ... \tag{15}$$

As required by the control criteria, the total network-induced delay should be less than $T_s$, that is $\tau_1 + \tau_2 < T_s$. Since the delay $\tau_2$ in the downlink is always $T$, the access delay $\tau_1$ should less than $T_s - T$. This requirement implies that the up bound of the access delay in the slotted ALOHA is $D_0 + nT$, where $n$ is given by (10). Hence, the value of $\tau_1$ can only be selected from the following set

$$\{D_0,\ D_0 + 2T,\ D_0 + 3T, \ldots, D_0 + nT\} \tag{16}$$

It is worth noting that $\tau_1$ can not be $D_0 + T$, because at least two slots are required for retransmission. According to the delay distributions of Slotted ALOHA [12] with some modification, the mean value of the access delay $\tau_1$ with the upper bound $D_0 + nT$ can be estimated as

$$\mathrm{E}(\tau_1) = D_0 + \frac{-1 + \sum_{i=1}^{n} \mathrm{P}\{\tau_1 \geqslant D_0 + iT\}}{\mathrm{P}\{\tau_1 \leqslant D_0 + nT\}} + 1 \tag{17}$$

where $\mathrm{P}\{\bullet\}$ represents the possibility of an event. For instance, $\mathrm{P}\{\tau_1 \leqslant D_0 + nT\}$ denote the possibility of $\tau_1 \leqslant D_0 + nT$ [12].

Hence, by selecting a proper initial backoff window size $w$, the distribution of the re-transmission $R$ can be calculated and the mean value of the access delay $E(\tau_1)$ in the uplink channel can be computed as (17). Thus the mean value of the total network-induced delay can be estimated as

$$\hat{\bar{\tau}} = T + E(\tau_1) \tag{18}$$

## IV. Fault Detection Observer design

Recalling the model of delay (4), one can rewrite (2c) as

$$\begin{aligned} B_1(\tau_k) &= \int_{T_s - \bar{\tau} - \Delta\tau_k}^{T_s} e^{At} B dt \\ &= A^{-1}[I - e^{-A(\bar{\tau} + \Delta\tau_k)}] e^{AT} \cdot B && (19) \end{aligned}$$

Applying the Taylor approximation of $e^{At} = [I + At] + g(t)$ to (19) gives

$$\begin{aligned} B_1(\tau_k) &= A^{-1}\big[I - [I - A(\bar{\tau} + \Delta\tau_k) + g(\bar{\tau} + \Delta\tau_k)]\big] \bar{A} \cdot B \\ &= \bar{A}(\bar{\tau} + \Delta\tau_k) - g(\tau_k)] \cdot B \\ &= \bar{A} \cdot \bar{\tau} \cdot B + \bar{A} \cdot \Delta\tau_k \cdot B - \bar{A} g(\tau_k) \cdot B \\ &= \bar{A}B \cdot \bar{\tau} + \bar{A}B \cdot \Delta\tau_k - \bar{A} g(\tau_k) B && (20) \end{aligned}$$

Substitute (20) into the state equation of (5), the system model can be written as

$$\begin{cases} x_{k+1} = \bar{A}x_k + [\bar{B} \quad -\bar{A}B\bar{\tau}] \begin{pmatrix} u_k \\ \Delta u_k \end{pmatrix} \\ \qquad\qquad +[-\bar{A}B\Delta u_k \quad \bar{B}_d\ ] \begin{pmatrix} \Delta\tau_k \\ d_k \end{pmatrix} + \bar{B}_f f_k + \bar{g}(\tau_k) \\ y_k = Cx_k + D_2 v_k \end{cases}$$

where $\bar{g}(\tau_k) = -\bar{A}g(\tau_k)B\Delta u_k$ denotes modeling errors due to the approximation. Define

$$\begin{array}{ll} \Gamma(\bar{\tau}) = [\bar{B} \quad -\bar{A}B\bar{\tau}], & \bar{u}_k = \begin{pmatrix} u_k \\ \Delta u_k \end{pmatrix} \\ \Gamma_d = [-\bar{A}B\Delta u_k \quad \bar{B}_d], & \bar{d}_k = \begin{pmatrix} \Delta\tau_k \\ d_k \end{pmatrix} \end{array}, \tag{21}$$

the system model now is expressed in the following form:

$$\begin{cases} x_{k+1} = \bar{A}x_k + \Gamma(\bar{\tau})\bar{u}_k + \Gamma_d\ \bar{d}_k + \bar{B}_f f_k + \bar{g}(\tau_k) \\ y_k = Cx_k + v_k \end{cases} \tag{22}$$

A networked fault detection observer in the following discrete Luenberger observer form is used:

$$\begin{cases} \hat{x}_{k+1} = \bar{A}x_k + \Gamma(\hat{\bar{\tau}})\bar{u}_k + L(y_k - \hat{y_k}) \\ \hat{y}_k = C\hat{x}_k \\ r_k = W(y_k - \hat{y}_k) \end{cases} \tag{23}$$

where $r_k \in \mathbb{R}^l$ is the so-called residual for indicating the fault occurrence, $L \in \mathbb{R}^{n\times p}$ and $W \in \mathbb{R}^{l\times p}$ are the observer gain matrix and weighting matrix to be designed, respectively.

Let the observer error be $e_k = x_k - \hat{x_k}$ and $\hat{\bar{\tau}}$ the estimation of $\bar{\tau}$, the the overall dynamics of residual generator (23) is governed by

$$\begin{cases} e_{k+1} = (\bar{A} - LC)e_k + (\Gamma(\bar{\tau}) - \Gamma(\hat{\bar{\tau}}))\bar{u}_k \\ \qquad\qquad +\Gamma_d\bar{d}_k + \bar{B}_f f_k - Lv_k + \bar{g}(\tau_k) \\ r_k = WCe_k + Wv_k \end{cases} \tag{24}$$

In equation (24), it appears that the residual dynamics depends on the amplitude of the terms $\Gamma(\bar{\tau})$, $\Gamma(\hat{\bar{\tau}})$ and $\Delta\tau_k$ which are functions of network-induced delay $\tau_k$.

Recalling $\Gamma(\bar{\tau}) = [\bar{B} \quad -\bar{A}B\bar{\tau}]$ in (21), one can get

$$\begin{array}{rl} & [(\Gamma(\bar{\tau}) - \Gamma(\hat{\bar{\tau}}))]\bar{u}_k \\ = & [[\bar{B} \quad -\bar{A}B\bar{\tau}] - [\bar{B} \quad -\bar{A}B\hat{\bar{\tau}}]] \begin{bmatrix} u_k \\ \Delta u_k \end{bmatrix} \\ = & \begin{bmatrix} 0 & \bar{A}B(\hat{\bar{\tau}} - \bar{\tau}) \end{bmatrix} \Delta u_k \\ = & \bar{A}B(\hat{\bar{\tau}} - \bar{\tau}) \cdot \Delta u_k \end{array} \tag{25}$$

Then $(\Gamma(\bar{\tau}) - \Gamma(\hat{\bar{\tau}}))\bar{u}_k + \Gamma_d\bar{d}_k$ in (24) can be re-formed as

$$[\bar{A}B\Delta u_k \quad \bar{A}B\Delta u_k \quad \bar{B}_d] \begin{bmatrix} (\hat{\bar{\tau}} - \bar{\tau}) \\ \Delta\tau_k \\ d_k \end{bmatrix}, \tag{26}$$

and the state estimation error (24) now is

$$\begin{cases} e_{k+1} = (\bar{A} - LC)e_k + [\bar{A}B \quad \bar{A}B \quad \bar{B}_d] \begin{bmatrix} (\hat{\bar{\tau}} - \bar{\tau})\Delta u_k \\ \Delta\tau_k\Delta u_k \\ d_k \end{bmatrix} \\ \qquad\qquad +\bar{B}_f f_k - Lv_k + \bar{g}(\tau_k) \\ r_k = WCe_k + Wv_k \end{cases} \tag{27}$$

It shows that the network-induced delay introduces an extra unknown input $\begin{bmatrix} (\hat{\bar{\tau}} - \bar{\tau})\Delta u_k \\ \Delta\tau_k\Delta u_k \end{bmatrix}$ into the dynamics of the residual generator, and the residual may deviate from zero even if no fault has happened. If this unknown input is not carefully taken into account in the fault detection observer design, residual $r_k$ could not reflect the fault's occurring properly. This phenomenon causes either false alarms or failure to detect faults.

Different from the decoupling unknown inputs in some fault detection observer designs, in this paper, the FD observer design problem is turned into an eigenvalue assignment and performance optimization problem. The objectives are to select an gain matrix $L \in \mathbb{R}^{n\times p}$, such that the following three criteria are met:

- Stability: The poles of $\bar{A} - LC$ in (27) is within the unit circle in the $z$-plane;
- Sensitivity to faults: The residual $r_k$ should be sensitive to faults $f_k$, that is the transfer function matrix (TFM) relating $f_k$ to $r_k$ should be maximized;
- Robustness to disturbances and delays: The residual $r_k$ should be insensitive to disturbances $d_k$ and network-induced delays $\tau_k$. That is the TFM relating $d_k$ and $\tau_k$ to $r_k$ should be minimized.

## V. Fault Detection Performance Optimisation

In the following, the impacts of the delays on the residual $r_k$ is analyzed in terms of TFMs and an the gain matrix $L$ in (27) is optimised to ensure the residual's sensitivity to the faults and enhance the robustness against the delays.

The $z$-transform of (27) gives the following transfer function matrices (TFMs)

$$\begin{cases} G_1(z) = G_2(z) = C(zI - \bar{A} + LC)^{-1}\bar{A}B \\ G_f(z) = C(zI - \bar{A} + LC)^{-1}B_f \end{cases} \tag{28}$$

where $G_1(G_2)$ and $G_f$ are TFM relating $r_k$ to $(\hat{\bar{\tau}} - \bar{\tau})\Delta u_k$, $(\Delta\tau_k\Delta u_k)$ and $f_k$ , respectively.

### A. Optimization in the frequency domain

As shown in the FD problem definition, there are two objective functions, namely sensitivity index and robustness index. Considering the access delays of the ALOHA in section III, we propose the following two performance indices.

*1) Robustness Index:* Observe that the TFMs $G_1(G_2)$ relating $r_k$ to $\Delta\tau_k\Delta u_k$ ($[\hat{\bar{\tau}} - \bar{\tau}]\Delta u_k$) are the same, but the signals $\Delta\tau_k$ and $(\hat{\bar{\tau}} - \bar{\tau})$ have different characteristics in the frequency domain. By optimizing the fault detection performance at the frequency of interest, instead of the whole frequency range, our observer is tailored for attenuating the network-induced delay.

$$\min_L J_1 = \sum_{z\in\Omega} \|\beta(z)G_1(z)\| \tag{29}$$

where $\Omega$ is the frequency range of interest, and $\beta(z)$ is a weighting function for $G_1(z)$ over the frequency range $\Omega$. $\Omega$ and $\beta(z)$ depend on the frequency characteristics of the MAC-delay $\tau$ and the the difference between successive input samples $\Delta u$. Generally, since it is the residual we are interested in, $\Omega$ and $\beta(z)$ are determined by finding the significant frequency components of $r_k$.

*2) Sensitivity Index:* Not like a random noise, a fault signal is usually associate with some pattern, and, its energy distribution is not uniformly distributed over the whole frequency range of interest. In general, two common faults are considered, namely, the incipient fault (a ramp signal) and the abrupt fault (a step signal). A ramp fault mainly consists of low-frequency components. For an abrupt fault, high-frequency contents exist only at the time instant when the fault starts, and it is almost constant (zero frequency) content thereafter. In order to increase the fault significance in residual $r$, it is proposed that the sensitivity index is maximized at $z = 1$ (corresponding to zero frequency in the continuous frequency domain).

$$\max_L J_2 = \|G_f(z)\|_{z=1} \tag{30}$$

Combining robustness index (29) and sensitivity index (30) yields the performance index

$$\min_{Q,\Lambda} J = \frac{J_1}{J_2} = \frac{\sum_{z\in\Omega}\|\beta(z)G_1(z)\|}{\rho + \|G_f(z)\|_{z=1}} \tag{31}$$

where $\rho$ is an arbitrary small positive number to ensure a non-zero denominator. Since the frequency information has now incorporated into the new index (31), the resulting FD observer is optimal for attenuating the negative impacts caused by the MAC-delay. In most applications, such a FD observer has a better performance in terms of attenuating the MAC-delay induced disturbance.

## VI. Simulation Platform and Results Analysis

In order to demonstrate the the proposed delay compensation approach, the hybrid simulation platform and fault detection results are presented in this section. The control system and the FD system are programmed in MATLAB/SIMULINK, and the wireless network is emulated by using OMNeT++. Compared with MATLAB/SIMULINK, the open source OMNeT++ platform is able to mimic the detailed behaviour of a wireless network, and imitate the features, such as SNIR(Singal-Noise-Interference-Ratio), SNR(Singal-Noise-Ratio), MAC-delay and packet loss due to either low SNR or collision.

In the simulation, the system matrices of the plant are as follows:

$$A = \begin{bmatrix} 30.7643 & 36.0164 \\ -30.8287 & -35.9486 \end{bmatrix}, \quad B = \begin{bmatrix} 2.2991 \\ -0.0668 \end{bmatrix},$$

$C = [1 \quad 0]$ and $B_d = B_f = B$. The sampling interval $T_s = 0.03\ sec$. The control input and plant output are subject to input disturbances $d_k$ and output measurement noises $v_k$, respectively, where $d_k$ and $v_k$ are independent bounded noises uniformly distributed between $[-0.2, 0.2]$ and $[-2, 2]$. Two kind of faults are concerned in the simulation, namely, step fault and slope fault. The step fault associated with the actuator occurs at time $t_1 = 10s$ with magnitude 0.2, and the slope fault starts at 10 s with slope rate 0.05.

As shown in Figure 1, we consider a networked control system composed with five local control loops in total and one AP. The total number of clock-driven sensors is ten and there are five event-driven controllers. A FD observer is to be designed for one of the control loops. All the communication devices works at 2Mbps on 2.4GHz ISM band. Data, either from sensors or from controllers, is transmitted with a single packet whose length is 1712 bits including physical premier(192 bits), protocol header length (272bits for MAC, 32bits for Network layer), which gives an air frame with a transmission duration of $856us$.

The simulation results of packet delay given by OMNeT++ are shown in Figure 2. The mean value of the packet delay from sensor to controller is $\bar{\tau} = 0.0136$s. With the aid of the MAC-delay estimate, a MAC-delay compensated FD observer can be constructed as (23) and the optimisation of index (31) yields the optimal gain matrix

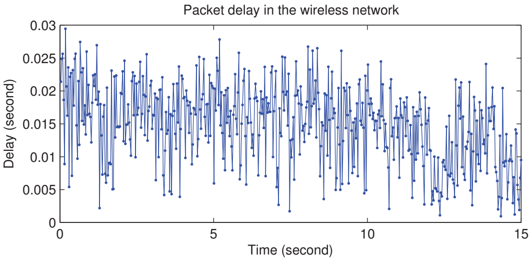


Fig. 2. Packet delay $\tau_k$ from sensor to controller

$$L = \begin{bmatrix} 0.1123 \\ -0.1027 \end{bmatrix} \tag{32}$$

The residuals of the proposed FD observer are given by Figure 3 (for step fault) and Figure 4 (for slope fault), respectively. From Figure 3-4, one can seen that the proposed delay compensation techniques is able to detect these faults clearly.

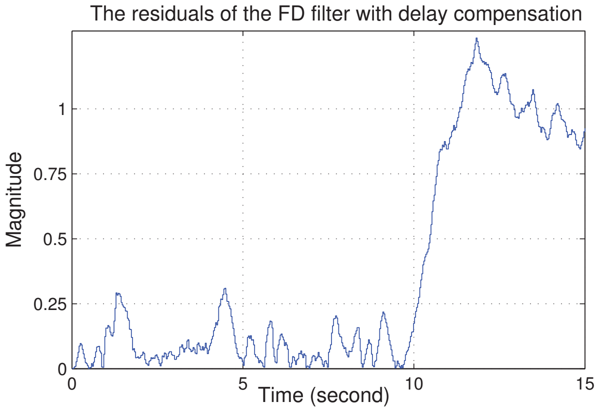


Fig. 3. Residuals $r_k$ subject to step fault occurring at 10 second with step size 0.2

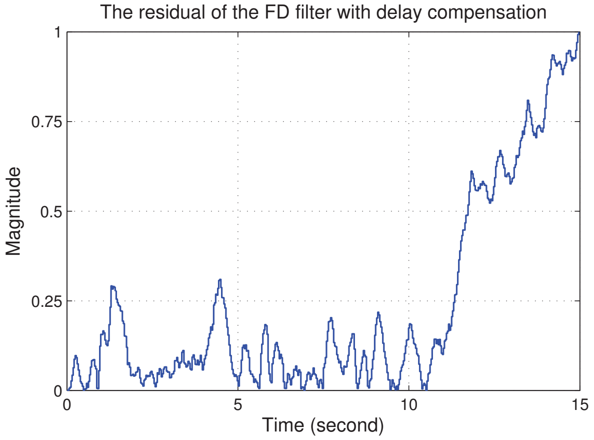


Fig. 4. Residuals $r_k$ subject to slope fault occurring at 10 second with rate of 0.05

## VII. Conclusions

This paper addresses the fault detection problem of WSANs-based wireless networked control systems, where the MAC-access delay of the slotted ALOHA is considered. With the aid of the MAC-access delay estimation, a FD observer with MAC-delay compensation is proposed. These advantages of the proposed FD observer with MAC delay compensation have been demonstrated through extensive computer simulations. As to future work, it is certainly possible to extend our cross-discipline analytical approach by considering more sophisticated MAC schemes (such as Carrier Sense Multiple Access(CSMA) and 802.11 Distributed Coordination Function), random backoff policies (such as geometric backoff and binary exponential backoff) and physical channel models.

## Acknowledgment

This work was supported by the National Natural Science Foundation of China (NSFC) under grant 61101135, the Energy Systems and Advanced Materials Research Group at Northumbria University and the Natural Science Foundation of Chongqing CSTC (2010BB2007).

## References

[1] V. D. Gupta, A. F. Hespanha, J. P. Murray, and R. M. B. Hassibi, “Data transmission over networks for estimation and control,” *IEEE Transactions on Automatic Control*, vol. 54, no. 8, pp. 1807–1819, Aug. 2009.

[2] T. Yang, “Networked control system: a brief survey,” *Control Theory and Applications, IEE Proceedings* -, vol. 153, no. 4, pp. 403 – 412, July 2006.

[3] H. Fang, H. Ye, and M. Zhong, “Fault diagnosis of networked control systems,” *Annual Reviews in Control*, vol. 31, no. 1, pp. 55–68, 2007.

[4] X. He, Z. Wang, and D. Zhou, “Robust fault detection for networked systems with communication delay and data missing,” *Automatica*, vol. 45, pp. 2634–2639, 2009.

[5] S. Ding and P. Zhang, “Observer-based monitoring of distributed networked control systems,” in *Proceedings of the IFAC Symposium SAFEPROCESS*, 2006.

[6] D. Llanos, M. Staroswiecki, J. Colomer, and J. Melendez, “$H_\infty$ detection filter design for state delayed linear systems,” in *Proceedings of the IFAC Symposium SAFEPROCESS*, 2006.

[7] N. P. Y. X. Hespanha, J.P., “A survey of recent results in networked control systems,” *Proceedings of the IEEE*, vol. 95, no. 1, pp. 138 –162, Jan. 2007.

[8] Y. Zheng, H. Fang, and H. O. Wang, “Takagisugeno fuzzy-model-based fault detection for networked control systems with markov delays,” *IEEE Trans. Syst. Man Cybern. B: Cybern*, vol. 36, no. 4, p. 924929, 2006.

[9] I. M. Al-Salami, S. Ding, and P. Zhang, “Statisical based residual evaluation for fault detection in networked control systems,” 2006.

[10] D. Sauter, S. Li, and C. Aubrun, “Robust fault diagnosis of networked control systems,” *Int. J. Adapt. Control Signal Process*, vol. 23, pp. 722–736, 2009.

[11] Y. Wang, Y. Hao, Y. Cheng, and G. Wang, “Fault detection of ncs based on eigendecomposition and pade approximation,” in *Proceedings of the IFAC Symposium SAFEPROCESS*, 2006, pp. 937–941.

[12] Y. Yang and T. S. P. Yum, “Delay distributions of slotted ALOHA and CSMA,” *IEEE Transactions on Communications*, vol. 51, no. 11, pp. 1846–1857, Nov. 2003.

[13] C. Aubrun, D. Sauter, and J. J. yame, “Fault diagnosis of networked control systems,” *Applied Mathematics and Computer Science*, vol. 18, no. 4, pp. 525–53, 2008.

[14] Y. Wang, S. X. Ding, H. Ye, and G. Wang, “A new fault detection scheme for networked control systems subject to uncertain time-varying delay,” *IEEE TRANSACTIONS ON SIGNAL PROCESSING*, vol. 56, no. 10, pp. 5258–5268, 2008.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Stability Analysis of Dynamic Quantized Feedback System With Packet Loss

Mu Li, Jian Sun, Lihua Dou
School of Automation
Beijing Institute of Technology
Beijing 100081, China
limu.bit@gmail.com, helios1225@yahoo.com.cn, doulihua@bit.edu.cn

*Abstract*—**This article is concerned with stability analysis of a linear discrete-time dynamic quantizer system with packet loss. First, several modifications are made to the original dynamic quantizer to make it easier to realize. Then communication channel subject to packet loss of Bernoulli distribution from the quantizer to the plant input is considered. Moreover, based on Lyapunov function approach, a sufficient condition for mean square stability of the closed-loop system is derived. Finally, a numerical simulation is given for effectiveness of the proposed method.**

***Keywords-Discrete-time; Dynamic quantizer; Mean square stability; Linear discrete system; Pactket loss;***

## I. Introduction

In the past few years, quantized control has been of great significance in the research field of control systems. In practical discrete-time systems, owing to limited network bandwidth, extensive use of encoders and decoders, command-driven actuators and discrete-level sensors, it becomes necessary for the signals to be quantized before transmission.

Numerous results on this subject have been obtained in recent years. Early studies of quantized control systems mainly focus on construction of static quantizers which can guarantee the stability of the system. For example, global asymptotic stabilization of continuous-time systems is considered in [1], [9], where a special uniform quantizer with scaling factor is used for quantization. Stabilization of the given system with the coarsest quantization density is analyzed in [2], [3], [7] using the sector bound approach. Furthermore, stabilization problem for systems with one-dimensional input using quantized feedback with a memory structure is analyzed in [4]-[6], focusing on the tradeoff between static quantizer complexity and system performance. And the least amount of information needs to be communicated between the quantizer and the controller in order to stabilize an unstable linear system is addressed in [8], [10]-[13]. Moreover, the coarsest quantization density to stabilize the system with networked packet losses is considered in [17]-[18], and input-to-state stability of systems with time-varying delays is analyzed in [19] in terms of LMIs.

For early works of quantized control systems, the parameters of quantizers stay invariant as the system evolves, which will generate large quantized error. This means that part of the system performance has to be sacrificed to cope with such a quantized error. For this reason, in [14], [15], a novel optimal dynamic quantizer was given whose parameters can vary as the system evolves, which is able to handle such problem. The optimal dynamic quantizer in [14], [15] is constructed to minimize the quantized error according to the output of the plant, which can sacrifice fewer system's performance. In further study of this work in [16], the stability problem was considered under a dynamic quantized LFT system in terms of the poles/zeros. However, these works cannot deal with networked problems such as packet losses and time-delays effectively.

In this paper, stability of optimal dynamic quantized system is analyzed based on Lyapunov function approach. Besides, networked packet loss of the closed-loop system is also taken into consideration, which cannot be effectively solved in the LFT form. Moreover, some modifications are made to the original dynamic quantizer. The static part of the dynamic quantizer is replaced by the static quantizer with scaling factor and saturation value, which makes it easier to be realized and can guarantee smaller quantized error when the state comes close to the equilibrium point.

The whole paper is organized as follows. In section 2, we review the original dynamic quantized system and introduce the improved dynamic quantizers used in this paper. Next in section 3, main result on stability analysis is derived. A numerical simulation is given in section 4 and section 5 concludes this article.

## II. Problem Formulation and Preliminaries

### A. Original Dynamic Quantized System

Consider the discrete-time systems shown in Fig.1, where the linear plant $P$ is given by

$$P:\begin{cases} x(k+1)=Ax(k)+Bv(k) \\ y(k)=Cx(k) \end{cases} \tag{1}$$

where $x\in R^n$ is the state, $v\in R^m$ is the control input, $y\in R^p$ is the output, $A\in R^{n\times n}$, $B\in R^{n\times m}$, and $C\in R^{p\times n}$ are system matrices. The initial state of the plant $P$ is

This work was supported by the National Natural Science Foundation of China (Grant No. 61104097, No. 61120106010), Research Fund for the Doctoral Program of Higher Education of China (Grant No. 20111101120027).

$x(0)=x_0$ for $x_0 \in R^n$ , the pair $(A,B)$ is stabilizable and $A$ is unstable. Assume the following assumption holds:

**Assumption 1.** The plant $P$ satisfies that the dimensions of $v$ and $y$ are the same $(m=p)$ and the matrix $CB$ is nonsingular.

The original optimal dynamic quantizer $Q$ in Fig.1 ( $\Sigma^{**}$ ) is given by [14]

$$Q:\begin{cases}\xi(k+1)=A\xi(k)-Bu(k)+Bv(k)\\ v(k)=q(-(CB)^{-1}CA\xi(k)+u(k))\end{cases} \tag{2}$$

where $\xi \in R^n$ , $u \in R^m$ , and $v \in R^m$ are the state, input, and output of this quantizer respectively, $A,B$ and $C$ of the quantizer are the same as the plant $P$ , $q: R^m \to R^m$ is a static uniform quantizer.

The static uniform quantizer $q$ in (2) satisfies that

$$q(x)=\left\lfloor \frac{x}{\Delta}+\frac{1}{2}\right\rfloor \Delta \tag{3}$$

that is

$$abs(q(x)-x)=abs(x-q(x))\le \frac{\Delta}{2} \tag{4}$$

where $\tilde{a}=\lfloor a \rfloor$ denotes the biggest integer satisfying $\tilde{a}\le a$ , and $\Delta$ is the quantized interval of the quantizer.

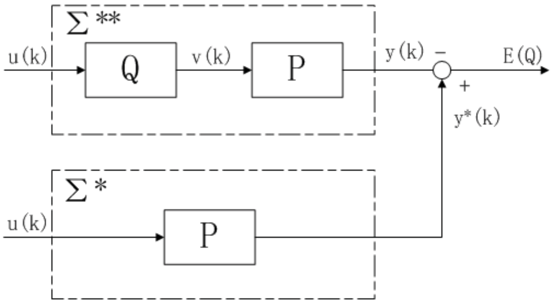


Figure 1. Two controlled plants: ( $\Sigma^{*}$ ) usual controlled plant; ( $\Sigma^{**}$ ) controlled plant with dynamic quantizer

**Remark 1.** The dynamic quantizer $Q$ in Fig.1 ( $\Sigma^{**}$ ) has been proved to be an optimal quantizer according to the output of the plant in [14]. Which indicates the parameters of the dynamic quantizer $Q$ are chosen such that the output error $E(Q)=\sup_{k\in Z_+}\left\| y(k)-y^{*}(k)\right\|$ between the system $\Sigma^{**}$ and $\Sigma^{*}$ is minimized. For the static quantizer $q$ , the quantized error is $\frac{\Delta}{2}$ as (4) indicates. It is clear that such open-loop dynamic quantized system cannot always guarantee stability of the whole system for the quantizer itself may not be stable.

*B. Closed-loop System*

In this paper, consider that packet losses occur with probability $\alpha$ in the input channel of the plant as is shown in Fig.2. The plant is described as

$$P:\begin{cases}x(k+1)=Ax(k)+B\theta(k)v(k)\\ y(k)=Cx(k)\end{cases} \tag{5}$$

where $\theta(k)$ is a Bernoulli random variable with a probability distribution given by

$$\Pr(\theta(k)=i)=\begin{cases}\alpha, & i=0,\\ 1-\alpha, & i=1,\end{cases} \quad 0\le\alpha<1 \tag{6}$$

Consider the closed-loop system $\Sigma$ in Fig.2 using sate feedback control law $u(k)$ which is given by

$$u(k)=Kx(k) \tag{7}$$

where $K \in R^{m\times n}$ is the feedback gain.

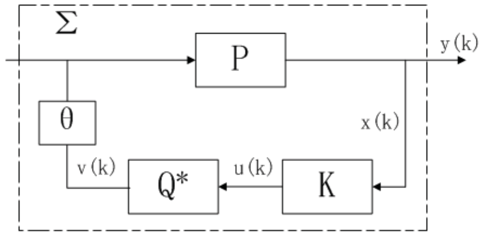


Figure 2. Dynamic quantized system $\Sigma$ with packet loss

In this paper we replace static uniform quantizer $q$ in (2) with $q_\mu$

$$q_\mu(x)=\begin{cases}\mu(k)M, & if \dfrac{x}{\mu(k)}>M-\dfrac{1}{2}\Delta\\ -\mu(k)M, & if \dfrac{x}{\mu(k)}\le -M+\dfrac{1}{2}\Delta\\ \left\lfloor \dfrac{x}{\Delta\mu(k)}+\dfrac{1}{2}\right\rfloor \Delta\mu(k), & if \dfrac{|x|}{\mu(k)}\le M-\dfrac{1}{2}\Delta\end{cases} \tag{8}$$

where $\mu(k)$ is the scaling factor, M is the saturation value and $\Delta$ is the sensitivity.

Obviously, the following conditions can be obtained according to the quantizer:

Ⅰ.If $|x|\le M\mu(k)$ , then $|q_\mu(x)-x|\le\frac{\Delta\mu(k)}{2}$ ;

Ⅱ.If $|x|>M\mu(k)$ , then $|q_\mu(x)|>M\mu(k)-\frac{\Delta\mu(k)}{2}$ .

**Remark 2.** The uniform quantizer $q_\mu$ here is different from quantizer $q$ in that it brings in a scaling factor $\mu(k)$ , and a saturation value $M$ . The former can guarantee $q_\mu$ not saturate by adjusting scaling factor $\mu(k)$ properly, which will be considered in next section. The latter makes the static part of the dynamic quantizer easier to realize. Moreover, it is clear that by bringing in this scaling factor $\mu(k)$ , when the quantizer does not meet with saturation, the quantized error varies according to $\mu(k)$ .

Therefore, an improved dynamic quantizer $Q^*$ in Fig.2 is given by

$$Q^*:\begin{cases}\xi(k+1)=A\xi(k)-Bu(k)+Bv(k)\\ v(k)=q_\mu(-(CB)^{-1}CA\xi(k)+u(k))\end{cases} \tag{9}$$

**Remark 3.** It should be noticed that although the static quantizer $q_\mu$ used here is different from $q$ in [14], the dynamic quantizer $Q^*$ is still the optimal quantizer for the system $\Sigma$ . It has been proved in [14] that for the dynamic quantizer (2) with static quantizer (3), minimum value of output error $E(Q)$ can be given as

$$E(Q)=\|abs(CB)\|\frac{\Delta}{2}$$

where for the matrix $M:=\{M_{ij}\}$ , $abs(M):=\{|M_{ij}|\}$ .

When it comes to $q_\mu$ for our system, the quantized error becomes $\frac{\Delta\mu(k)}{2}$ , obviously, similar conclusion can be expressed as

$$E(Q)=\|abs(CB)\|\frac{\Delta\overline{\mu}}{2}$$

where $\overline{\mu}=\sup_{k\in Z_+}(\mu(k))$ is upper bound of $\mu(k)$ . As a result, the improved dynamic quantizer is still optimal dynamic quantizer for the system.

Therefore, the closed-loop system $\Sigma$ in Fig.2 can be written as

$$\begin{cases}\begin{bmatrix}x(k+1)\\ \xi(k+1)\end{bmatrix}=\begin{bmatrix}A+BK & -\theta B(CB)^{-1}CA\\ 0 & A-B(CB)^{-1}CA\end{bmatrix}\begin{bmatrix}x(k)\\ \xi(k)\end{bmatrix}\\ \qquad\qquad+\begin{bmatrix}\theta B\\ B\end{bmatrix}\left[q_\mu(\phi)-\phi\right]\\ y(k)=[C\quad 0]\begin{bmatrix}x(k)\\ \xi(k)\end{bmatrix}\end{cases} \tag{10}$$

where $\phi=-(CB)^{-1}CA\xi(k)+Kx(k)$ .

Rewrite the closed-loop system $\Sigma$ as

$$\begin{cases}z(k+1)=\overline{A}z(k)+\overline{B}e(\Gamma z(k))\\ y(k)=\overline{C}z(k)\end{cases} \tag{11}$$

where $z(k)=\left[x(k)\quad \xi(k)\right]^T$ , $e(\Gamma z(k))=q_\mu(\phi)-\phi$ denotes the quantized error, and $\Gamma=\left[-(CB)^{-1}CA\quad K\right]$ .

Matrices $\overline{A}\in R^{2n\times 2n}$ , $\overline{B}\in R^{2n\times m}$ , $\overline{C}\in R^{p\times 2n}$ are defined as

$$\overline{A}=\begin{bmatrix}A+\theta BK & -\theta B(CB)^{-1}CA\\ 0 & A-B(CB)^{-1}CA\end{bmatrix},\overline{B}=\begin{bmatrix}\theta B\\ B\end{bmatrix},\overline{C}=[C\quad 0].$$

When $\theta(k)=1$ , we let

$$A_1=\overline{A}=\begin{bmatrix}A+BK & -B(CB)^{-1}CA\\ 0 & A-B(CB)^{-1}CA\end{bmatrix},B_1=\overline{B}=\begin{bmatrix}B\\ B\end{bmatrix}.$$

When $\theta(k)=0$ , we can get

$$A_2=\overline{A}=\begin{bmatrix}A & 0\\ 0 & A-B(CB)^{-1}CA\end{bmatrix},B_2=\overline{B}=\begin{bmatrix}0\\ B\end{bmatrix}.$$

**Definition 1.** System (11) is said to be mean square stable if

$$\lim_{k\to\infty}E\left\{\left|z(k)\right|^2\right\}=0 \tag{12}$$

for any initial state $z(0)\in R^{2n}$ .

**Lemma 1.** The following inequality holds for any positive definite matrix $G$ and matrices $E$ and $F$

$$E^TGF+F^TGE\le E^TGE+F^TGF \tag{13}$$

## III. STABILITY ANALYSIS

The following theorem presents a sufficient condition for the stability of the closed-loop system $\Sigma$ :

**Theorem 1.** For a given feedback gain $K$ and packet loss probability $\alpha$ , if there exists a positive definite symmetric matrix $P$ satisfying that

$$(1-\alpha)A_1^TPA_1+\alpha A_2^TPA_2-\frac{1}{2}P<0 \tag{14}$$

then the system (11) is mean square stable under control law $u(k)=Kx(k)$ .

*Proof*: The closed-loop system (11) can be expressed as

$$\begin{cases}z(k+1)=(1-\alpha)\left(A_1z(k)+B_1e(\Gamma z(k))\right)\\ \qquad\qquad+\alpha\left(A_2z(k)+B_2e(\Gamma z(k))\right)\\ y(k+1)=\overline{C}z(k)\end{cases} \tag{15}$$

Choose a Lyapunov function, $V(z(k))=z^T(k)Pz(k)$ , where $P\in R^{2n\times 2n}$ is a positive definite matrix, $\Delta V(z(k))$ is given by the following expression

$$\Delta V(z(k))=E\{V(z(k+1))\}-V(z(k))$$

by using Lemma 1, we can get

$$\begin{aligned}
&\Delta V(z(k)) \\
&= E\left\{z^T(k+1)Pz(k+1)\right\} - z^T(k)Pz(k) \\
&= (1-\alpha)[z^T(k)(A_1^T PA_1 - P)z(k) + z^T(k)A_1^T PB_1 e(\Gamma z(k)) \\
&\quad + e^T(\Gamma z(k))B_1^T PA_1 z(k) + e^T(\Gamma z(k))B_1^T PB_1 e(\Gamma z(k))] + \\
&\quad + \alpha[z^T(k)(A_2^T PA_2 - P)z(k) + z^T(k)A_2^T PB_2 e(\Gamma z(k)) \\
&\quad + e^T(\Gamma z(k))B_2^T PA_2 z(k) + e^T(\Gamma z(k))B_2^T PB_2 e(\Gamma z(k))] \\
&\le 2z^T(k)[(1-\alpha)A_1^T PA_1 + \alpha A_2^T PA_2 - \frac{1}{2}P]z(k) \\
&\quad + 2e^T(\Gamma z(k))[(1-\alpha)B_1^T PB_1 + \alpha B_2^T PB_2]e(\Gamma z(k)) \\
&\le 2z^T(k)[(1-\alpha)A_1^T PA_1 + \alpha A_2^T PA_2 - \frac{1}{2}P]z(k) \\
&\quad + 2e^T(\Gamma z(k))B^{*T} PB^* e(\Gamma z(k))
\end{aligned} \tag{16}$$

where $B^* = \left\{\overline{B} : \max(B_1^T PB_1, B_2^T PB_2)\right\}$.

Define $D = -\left[(1-\alpha)A_1^T PA_1 + \alpha A_2^T PA_2 - \frac{1}{2}P\right]$, if the matrix $D$ is positive definite, (14) can be obtained directly, and (16) can be rewritten as

$$\begin{aligned}
&\Delta V(z(k)) \\
&\le -2z^T(k)Dz(k) + 2e^T(\Gamma z(k))B^{*T} PB^* e(\Gamma z(k)) \qquad (17) \\
&\le -2\left[\lambda_{\min}(D)|z(k)|^2 - \left\|B^{*T} PB^*\right\| \Delta^2 \mu^2(k)\right]
\end{aligned}$$

The last expression of (17) is negative if the state of $\Sigma$ is outside the ball

$$H = \left\{z(k) : z(k) \le \Theta\Delta\mu(k)\right\} \tag{18}$$

where $\Theta = \sqrt{\dfrac{\left\|B^{*T} PB^*\right\|}{\lambda_{\min}(D)}}$.

Define the scaling factor $\Omega$ as

$$\Omega = \sqrt{\frac{\lambda_{\max}(P)}{\lambda_{\min}(P)}}\sqrt{\Theta^2 + \varepsilon}\,\|\Gamma\|\Delta M^{-1} \tag{19}$$

where $\varepsilon > 0$ is a fixed real number.

We will analyze the control strategy by two stages according to the variation of the scaling factor $\mu(k)$ as [1], [9]:

The "zooming-out" stage of the quantizer.

Set $u(k) = 0$, $\mu(0) = 1$, $\mu(k) = \|A\|^k$ and increase $k$ fast enough, then a positive integer $k$ can be found such that

$$\left|\frac{\Gamma z(k)}{\mu(k)}\right| \le M\sqrt{\frac{\lambda_{\min}(P)}{\lambda_{\max}(P)}} - 2\Delta$$

In the view of condition Ⅰ in the former section

$$\left|q_\mu\left(\frac{\Gamma z(k)}{\mu(k)}\right)\right| \le \left|\frac{\Gamma z(k)}{\mu(k)}\right| + \Delta \le M\sqrt{\frac{\lambda_{\min}(P)}{\lambda_{\max}(P)}} - \Delta$$

Define

$$k_0 = \min\left\{k \ge 1 : \left|q_\mu\left(\frac{\Gamma z(k)}{\mu(k)}\right)\right| \le M\sqrt{\frac{\lambda_{\min}(P)}{\lambda_{\max}(P)}} - \Delta\right\} \tag{20}$$

Hence it follows that

$$\left|\frac{\Gamma z(k_0)}{\mu(k_0)}\right| \le \left|q_\mu\left(\frac{\Gamma z(k_0)}{\mu(k_0)}\right)\right| + \Delta \le M\sqrt{\frac{\lambda_{\min}(P)}{\lambda_{\max}(P)}}$$

Therefore

$$|\Gamma z(k_0)| \le M\mu(k_0)\sqrt{\frac{\lambda_{\min}(P)}{\lambda_{\max}(P)}}$$

Which means

$$|z(k_0)| \le \frac{M}{\|\Gamma\|}\mu(k_0)\sqrt{\frac{\lambda_{\min}(P)}{\lambda_{\max}(P)}}$$

Hence $z(k_0)$ belongs to an ellipsoid

$$R_1 = \left\{z(k) : z^T(k)Pz(k) \le \frac{M^2}{\|\Gamma\|^2}\mu^2(k)\lambda_{\min}(P)\right\} \tag{21}$$

It is obvious that $\Gamma z(k) \le M\mu(k)$ holds with $\mu(k) = \mu(k_0)$ for all $z(k) \in R_1$.

Take $M$ and $\Delta$ in (19) properly to guarantee $\Omega < 1$, it follows that $R_1 \supset H$. Moreover, if $k \ge k_0$, $z(k)$ will never leave $R_1$.

The "zooming-in" stage of the quantizer.

Define

$$\tilde{\tau} = \frac{M^2\lambda_{\min}(P) - \Delta^2\Theta^2\|T\|^2\lambda_{\max}(P)}{\|T\|^2\lambda_{\min}(D)\Delta^2\varepsilon} \tag{22}$$

We can have $\tilde{\tau} > 0$ as $\Omega < 1$. Define $\tau = \lceil\tilde{\tau}\rceil$, where $\lceil\tilde{\tau}\rceil$ denotes the smallest integer satisfy $\tau \ge \tilde{\tau}$, $\tau \in Z_{0+}$.

Assume an inequality can be get according to $\tau \in Z_{0+}$

$$E[z^T(k_0+\tau)Pz(k_0+\tau)] \le \Delta^2\mu^2(k_0)(\Theta^2+\varepsilon)\lambda_{\max}(P) \tag{23}$$

Suppose (23) is not true, then we can have that

$$E[z^T(k_0+\tau)Pz(k_0+\tau)] > \Delta^2\mu^2(k_0)(\Theta^2+\varepsilon)\lambda_{\max}(P) \tag{24}$$

That is

$$E[|z(k_0+\tau)|^2] > \Delta^2\mu^2(k_0)(\Theta^2+\varepsilon) \tag{25}$$

for all $k \in [k_0, k_0+\tau]$.

Based on (18) and $\Omega < 1$, it is clear that

$$\begin{aligned} &\Delta V[z(k_0+\tau-1)] \\ &= E[z^T(k_0+\tau)Pz(k_0+\tau)] - \\ &\quad E[z^T(k_0+\tau-1)Pz(k_0+\tau-1)] \qquad (26) \\ &\le -\lambda_{\min}(D)E[|z(k_0+\tau-1)|^2] + \lambda_{\min}(D)\Theta^2\Delta^2\mu^2(k_0) \\ &< -\lambda_{\min}(D)\Delta^2\mu^2(k_0)\varepsilon \end{aligned}$$

Furthermore, it can be obtained that

$$\begin{aligned} &\Delta V[z(k_0+\tau-i)] \\ &= E[z^T(k_0+\tau-i+1)Pz(k_0+\tau-i+1)] - \\ &\quad E[z^T(k_0+\tau-i)Pz(k_0+\tau-i)] \qquad (27) \\ &\le -\lambda_{\min}(D)E[|z(k_0+\tau-i)|^2] + \lambda_{\min}(D)\Theta^2\Delta^2\mu^2(k_0) \\ &< -\lambda_{\min}(D)\Delta^2\mu^2(k_0)\varepsilon \end{aligned}$$

where $i \in \{1,2,3,\cdots,\tau\}$.

Then we have

$$\begin{aligned} &E[z^T(k_0+\tau)Pz(k_0+\tau)] - z^T(k_0)Pz(k_0) \\ &< -\lambda_{\min}(D)\Delta^2\mu^2(k_0)\varepsilon\,\tau \\ &\le -\lambda_{\min}(D)\Delta^2\mu^2(k_0)\varepsilon\,\tilde{\tau} \qquad (28) \\ &= \lambda_{\max}(P)\Delta^2\Theta^2\mu^2(k_0) - \frac{M^2}{\|\Gamma\|^2}\lambda_{\min}(P)\mu^2(k_0) \end{aligned}$$

However, the following inequality can be obtained from (21) and (24)

$$\begin{aligned} &E[z^T(k_0+\tau)Pz(k_0+\tau)] - z^T(k_0)Pz(k_0) \\ &\ge \Delta^2\mu^2(k_0)(\Theta^2+\varepsilon)\lambda_{\max}(P) - \mu^2(k_0)\frac{M^2}{\|\Gamma\|^2}\lambda_{\min}(P) \qquad (29) \\ &\ge \lambda_{\max}(P)\Delta^2\Theta^2\mu^2(k_0) - \frac{M^2}{\|\Gamma\|^2}\lambda_{\min}(P)\mu^2(k_0) \end{aligned}$$

Obviously (28) and (29) contradict with each other, which implies the validity of (23).

Based on (23) and $\Omega < 1$, it follows that

$$\begin{aligned} E[z^T(k_0+\tau)Pz(k_0+\tau)] &\le \Delta^2\mu^2(k_0)(\Theta^2+\varepsilon)\lambda_{\max}(P) \\ &< (\Omega\mu(k_0))^2\frac{M^2}{\|\Gamma\|^2}\lambda_{\min}(P) \qquad (30) \end{aligned}$$

Thus it is clear that $z(k_0+\tau)$ belongs to

$$R_2 = \left\{ z(k) : E[z^T(k)Pz(k)] \le (\Omega\mu(k))^2\frac{M^2}{\|\Gamma\|^2}\lambda_{\min}(P) \right\} \qquad (31)$$

Let $\mu(k) = \Omega\mu(k_0)$ for $k_0+\tau \le k \le k_0+2\tau$, a similar result can be obtained

$$E[z^T(k_0+2\tau)Pz(k_0+2\tau)] < (\Omega^2\mu(k_0))^2\frac{M^2}{\|\Gamma\|^2}\lambda_{\min}(P) \qquad (32)$$

Let $\mu(k) = \Omega^{i-1}\mu(k_0)$ for $k_0+(i-1)\tau \le k \le k_0+i\tau$, it can be given as

$$E[z^T(k_0+i\tau)Pz(k_0+i\tau)] < (\Omega^i\mu(k_0))^2\frac{M^2}{\|\Gamma\|^2}\lambda_{\min}(P) \qquad (33)$$

By repeating this procedure, we can obviously obtain that $\mu(k) \to 0$ as $k \to \infty$, and $\lim_{k\to\infty} E[|z(k)|^2] \to 0$, which means the closed-loop system $\Sigma$ is mean square stable.

## IV. A NUMERICAL EXAMPLE

In this section, a numerical example is given to show the effectiveness of the proposed method. Consider the following system

$$P: \begin{cases} x(k+1) = \begin{bmatrix} 1.1 & \\ & 0.5 \end{bmatrix} x(k) + \begin{bmatrix} 1 \\ 2 \end{bmatrix} \theta(k)v(k) \\ y(k) = \begin{bmatrix} 1 & 0.2 \end{bmatrix} x(k) \end{cases}$$

where $\theta(k)$ is a Bernoulli random variable with probability distribution given by

$$\Pr(\theta(k) = i) = \begin{cases} 0.2, & i = 0 \\ 0.8, & i = 1 \end{cases}$$

which means the packet-loss rate is 0.2.

Obviously, the plant $P$ is unstable as one of its eigenvalues is 1.1. And it is stabilizable as $rank[B\ AB] = 2$. Set the feedback gain $K = [-1.53\ 0.13]$. The dynamic quantizer is given as

$$Q: \begin{cases} \xi(k+1) = \begin{bmatrix} 1.1 & \\ & 0.5 \end{bmatrix} \xi(k) - \begin{bmatrix} -1.53 & 0.13 \\ -3.06 & 0.26 \end{bmatrix} x(k) \\ \qquad + \begin{bmatrix} 1 \\ 2 \end{bmatrix} v(k) \\ v(k) = q_\mu(\begin{bmatrix} -0.7857 & 0.0714 \end{bmatrix} \xi(k) \\ \qquad + \begin{bmatrix} -1.53 & 0.13 \end{bmatrix} x(k)) \end{cases}$$

whose parameters are determined by the plant $P$ and the feedback gain $K$.

The closed-loop system $\Sigma$ can be expressed as

$$\begin{cases} z(k+1) = \begin{bmatrix} -0.43 & 0.13 & -0.7857\theta & -0.0714\theta \\ -3.06 & 0.76 & -1.5714\theta & -0.1429\theta \\ & & 0.3143 & -0.0714 \\ & & -1.5714 & 0.3571 \end{bmatrix} z(k) \\ \qquad + \begin{bmatrix} \theta & 2\theta & 1 & 2 \end{bmatrix}^T e(\Gamma z(k)) \\ y(k+1) = \begin{bmatrix} 1 & 0.2 & 0 & 0 \end{bmatrix} z(k) \end{cases}$$

We can find there is a positive definite matrix

$$P=\begin{bmatrix} 7.7602 & -1.2269 & 2.1796 & -0.4196 \\ -1.2269 & 0.3699 & -0.5016 & 0.1113 \\ 2.1796 & -0.5016 & 9.8224 & 0.1117 \\ -0.4196 & 0.1113 & 0.1117 & 0.5190 \end{bmatrix}$$

satisfying (14), then the system is mean square stable under control law (7).

The static quantizer is given by

$$q_\mu(x)=\begin{cases} 4\mu(k), & if\, x>3.95\mu(k) \\ -4\mu(k), & if\, x\le -3.95\mu(k) \\ \left\lfloor \dfrac{x}{\mu(k)}+0.05 \right\rfloor \mu(k), & if\, |x|\le 3.95\mu(k) \end{cases}$$

with $M=4$ and $\Delta=0.1$.

Set $\varepsilon=0.1$, we can get $\Omega=0.7497<1$. Let the initial state of the system be $z(0)=[5\ \ 7\ \ 8\ \ 6]^T$. Then the trajectories of state $z(k)$ are shown in Fig.3, where $z(i)$ denotes the $i$ -th component of $z(k)$. It is clear that system (11) is mean square stable as $z(k)\to 0$ when $k\to\infty$.

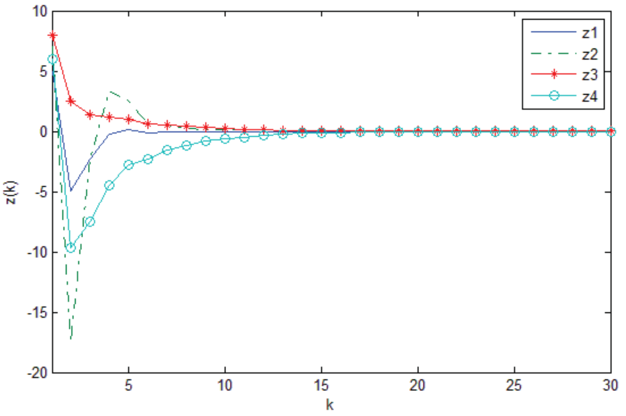


Figure 3. Trajectories of state $z(k)$

## V. Conclusion

This paper has discussed the stability of the optimal dynamic quantizer system with packet loss subject to Bernoulli distribution. In order to make the quantizer more practical and have better performance, traditional optimal dynamic quantizer has been improved here by replacing its static part with another one which contains saturation value and scaling factor. The communication channel has been considered subject to packet loss from the quantizer to the plant input. Based on Lyapunov function approach, a sufficient condition for mean square stability of the closed-loop system has been given.

## References

[1] R. W. Brockett and D. Liberzon, "Quantized feedback stabilization of linear systems", IEEE Transactions on Automatic Control, Vol. 45, No. 7, pp. 1279–1289, 2000.

[2] N. Elia and S. K. Mitter, "Stabilization of linear systems with limited information", IEEE Transactions on Automatic Control, Vol. 46, No. 9, pp. 1384–1400, 2001.

[3] M. Fu and L. Xie, "The sector bound approach to quantized feedback control", IEEE Transactions on Automatic Control, Vol. 50, No. 11, pp. 1698–1711, 2005.

[4] F. Fagnani and S. Zampieri, "Performance evaluations of quantized stabilizers", Proceedings of the 42nd IEEE Conference on Decision and Control, pp. 1897–1901, 2003.

[5] F. Fagnani and S. Zampieri, "Stability Analysis and Synthesis for Scalar Linear Systems With a Quantized Feedback", IEEE Transactions on Automatic Control, Vol. 48, No. 9, pp. 1569–1584, 2003.

[6] F. Fagnani and S. Zampieri, "Quantized Stabilization of Linear Systems : Complexity Versus Performance", IEEE Transactions on Automatic Control, Vol. 49, No. 9, pp. 1534–1548, 2004.

[7] H. Hatmovich, M.M.Seron and G.C.Goodwin, "Geometric characterization of multivariable quadratically stabilizing quantizers" , International Journal of Control, Vol. 79, No. 8, pp. 845–857, 2006.

[8] W. S. Wong and R. W. Brockett, "Systems with finite communication bandwidth constraints, ii: Stabilization with limited information feedback".IEEE Transactions on Automatic Control, Vol. 44, 1049–1053, 1999.

[9] D. Liberzon, "Hybrid feedbackstabilization of systems with quantized signals", Automatica, Vol. 39, pp. 1543–1554, 2003.

[10] G.N. Nair and R.J. Evans, "Stabilizability of stochastic linear systems with finite feedback data rates", SIAM J. Control Optim., Vol. 43, No.2, pp. 413–436, 2004.

[11] S. Tatikonda and S. Mitter, "Control under communication constraints," IEEE Trans. Automat. Control, Vol. 49, No.7, pp. 1056–1068, 2004.

[12] G.N.Nair, F. Fagnani, Sandro Zampieri, and R.J. Evans, "Feedback Control Under Data Rate Constraints: An Overview" , Proceedings of the IEEE, Vol. 95, No. 1, pp. 108–137, 2007.

[13] G.N. Nair and R.J. Evans, "Exponential stabilisability of finite dimensional linear systems with limited data rates" , Automatica, Vol. 44, No. 9, pp. 2364–2369, 2008.

[14] S. Azuma and T. Sugie, "Optimal dynamic quantizers for discrete valued input control" , Automatica, Vol. 44, No. 2, pp. 396–406, 2008.

[15] S. Azuma and T. Sugie, "An Analytical Solution to Dynamic Quantization Problem of Nonlinear Control Systems" , Proceedings of the 48th IEEE Conference on Decision and Control, pp. 3914–3919, 2009.

[16] S. Azuma and T. Sugie, "Stability analysis of optimally quantised LFT-feedback systems", International Journal of Control, Vol. 83, No. 6, 1125–1135, 2010.

[17] H.Ishii, T.Koji , "The Coarsest Logarithmic Quantizers for Stabilization of Linear Systems with Packet Losses" , Proceedings of the 46th IEEE Conference on Decision and Control, pp. 2235–2240, 2007.

[18] H.Ishii, T.Koji , "Tradeoffs between quantization and packet loss in networked control of linear systems" , Automatica, Vol. 45, No. 12, pp. 2963–2970, 2009.

[19] E.Fridman, M.Dambrine, "On input-to-state stability of systems with time-delay: A matrix inequalities approach" , Automatica, Vol. 44, No. 9, pp. 2364–2369, 2008.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Design and simulation of fuzzy controller based (IPM) converter fed DC Motors

Abdallah A. Ahmed, Yuanqing Xia and Bo Liu

*Abstract*—**This paper proposes a DSPF2812 based 32-bit micro-controller is used to to generate PWM waveform required to switch IPM_DRIVE, by using fuzzy logic controller to control dc motor fed by AC/DC (IPM) converter. The fuzzy controller is designed in such a way that it can be implemented in a micro-controller or DSP processor based embedded system. The system designed consists of an inner ON/OFF current controller and an outer fuzzy speed controller. The fuzzy speed controller is used to change the duty cycle of the(IPM) converter and thereby, the voltage fed to the DC motor regulates its speed. The performance in respect of load variation and speed change has been reported. Simulations results are presented to demonstrate the proposed performance and then, compared with the reported results and found that the performance of fuzzy based IPM_DRIVE drive for DC motors is improved.**

*Index Terms*—**Fuzzy controller, DC motors, IPM_DRIVE converter, speed control.**

## I. INTRODUCTION

Since the first successful application of the fuzzy concept into the control of a dynamic plant several decades ago, there has been a considerable world wide interest in fuzzy controller. It has been known that it is possible to control many complex systems effectively by skilled operators who have no knowledge of their underlying dynamics, while it is difficult to achieve the same with conventional controllers [1][2]. They are used in several applications ranging from the control of power converters to speed control of motors [3]. It's suitable for applications such as the speed control of a dc motor which has nonlinearities. It is this fact which has ultimately made the fuzzy controller is more powerful to handle those un-modeled uncertainties. Meanwhile, the motion control applications can be found in almost every sector of industry, from factory automation and robotics to high-tech computer hard disk drives. They are used to regulate mechanical motions in terms of position, velocity, acceleration and/or to coordinate the motions of multiple axes or machine parts. Furthermore, DC motors drives have been widely used in such applications where the accurate speed tracking is required, and in spite of the fact that AC motors are rugged, cheaper and lighter, DC motors are still a very popular choice in particular applications. It is known as a typical plant in the teaching on the control theory and research, and many methods have been developed for DC motors, for example, see [4], [5]-[10]. In conventional control strategies were used and it comprises of fixed arrangement with fixed parameter design. Hence the tuning and optimization of these controllers is a challenging and difficult task, particularly, under varying load conditions, parameter changes, abnormal modes of operation.

[11] has demonstrated and reported the separately excited dc motor fed by a chopper (DC to DC converter) and controlled by a fuzzy logic controller. It has been reported that the fuzzy logic controller controls the duty cycle of the chopper, there by the voltage fed to the motor for regulating the speed. The experimental setup has improved the performance over PI controller. It is seen that the separately excited dc drive have low starting torque which limits its applications.

H.A.Yousef and H.M.Khalil [12] have demonstrated the dc series motor drive fed by a single phase controlled rectifier (AC to DC converter) and controlled by fuzzy logic. It has been concluded that the fuzzy logic controller provides better control over the classical PI controller which has improved the performance. It is also reported that the settling time and maximum overshoot can be reduced. Due to the inherent limitations, AC to DC converter fed drive introduces unwanted harmonic ripples in the output.

Abdallah Ahmed and Yuanqing Xia are with the Department of Automatic Control, Beijing Institute of Technology, Beijing 100081, China. E-mail: abdouahmed12@gmail.com, xia_yuanqing@bit.edu.cn; Bo Liu is with the Systems Engineering Department, King Fahd University of Petroleum and Minerals, 31261 Dhaharn, Saudi Arabia. E-mail: bo.liu.777@gmail.com

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

[13] has reported the dc series motor drive fed by a single phase full-bridge converter (DC to DC converter) controlled by fuzzy logic. It has been reported that the motor performance was simulated for different controllers like simplify fuzzy logic model (SFL), PI type fuzzy controller (FPI) and classical PI controller. The simulation result shows that the SFI provides superior performance over other controllers. It is found from the analysis that only the speed error has been taken as fuzzy input. There are some important achievements in fuzzy controller for closed loop control of DC drive (see, for example [14]-[20]).

The proposed system deals development of a speed for a dc motor fed by AC/DC (IPM) converter, utilizes the fuzzy logic controller and IPM_DRIVE (IPM power modules). The fuzzy logic based speed command is followed even under load torque disturbances.

The paper is structured as follows. Section II describes the mathematical model formulation of the proposed DC motors, while Section III presents the design of Fuzzy Logic Control . Section IV, gives a simulation results and discussion. Conclusion is given in Section V.

## II. Mathematical Model of DC Motors

The goal in the development of the mathematical model is to relate the voltage applied to the armature to the velocity or position of the motor. Two balance equations can be developed by considering the electrical and mechanical characteristics of the system. Because of the complexity of dynamic-system problems, idealizing assumptions will be made. These assumptions are:

**Assumption 2.1:** The brushes are narrow, and commutation is linear.

**Assumption 2.2:** The armature is assumed to have no effect on the total direct-axis flux because the armature-wave is perpendicular to the field axis.

**Assumption 2.3:** The effects of the magnetic saturation will be neglected [21].

The electric circuit of the armature and the free body diagram of the rotor are shown in Fig.1, where

$V_t$ - Motor terminal voltage
$Ra$ - Armature resistance
$La$- Armature inductance
$J$ - Moment of inertia
$B$-Friction coefficient
$T_l$ - Load torque
$\omega$ -Angular speed
$\theta$-Angular displacement
$K_T$- Torque constant and
$K_b$- Back emf constant.

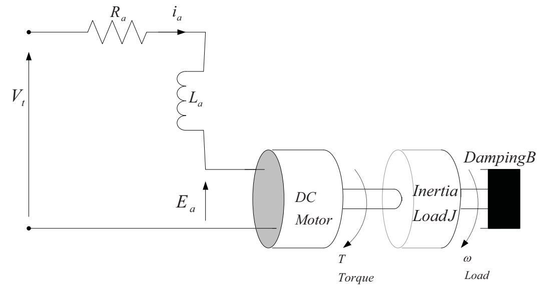


Fig. 1: DC Motor Equivalent Circuit.

By choosing $i_a$, $\omega$ and $\theta$ as the state variables and $V_t$ as input. The output is chosen to be $\omega$, and define the system matrices $A_c$, $B_c$ and $C_c$ as follows

$$A_c = \begin{bmatrix} \frac{-R_a}{L_a} & \frac{-K_b}{L_a} & 0 \\ \frac{K_T}{J} & \frac{-B}{J} & 0 \\ 0 & 1 & 0 \end{bmatrix}, \tag{1}$$

$$B_c = \begin{bmatrix} \frac{1}{L_a} \\ 0 \\ 0 \end{bmatrix}, C_c = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}, \tag{2}$$

we have the following state space description of DC motors dynamic

$$\left\{ \begin{array}{lcl} \dot{x}_t & = & A_c x_t + B_c u_t \\ y_t & = & C_c x_t \end{array} \right. \tag{3}$$

where state variables are

$$x = [i_a \ \omega \ \theta]^T \tag{4}$$

Sampling this system with step time $T_s$ yields

$$\left\{ \begin{array}{lcl} x_{k+1} & = & Ax_k + Bu_k \\ y_k & = & Cx_k \end{array} \right. \tag{5}$$

where

$$A = e^{A_c T_s}, \quad B = \int_0^{T_s} e^{A_c(T_s-\tau)} B_c d\tau, \ C = C_c \tag{6}$$

## III. Fuzzy Logic Control (FLC)

Fig.2 shows the block diagram of the complete proposed system. The proposed system consists of DC motor, IPM_DRIVE (IPM power modules) for driving the dc motor. A tacho generator(optical encoder) is used to measure the speed and which used for speed feedback. A DSPF2812 based 32-bit microcontroller is used to to generate PWM waveform required to switch Intelligent Power Module (IPM), during the experimental setup. Firstly, Matlab/simlink model of the DC motor and the IPM_Drive power converter is developed and simulated, and then, fuzzy logic controller is designed by using fuzzy toolbox, finally, the closed loop operation is simulated.

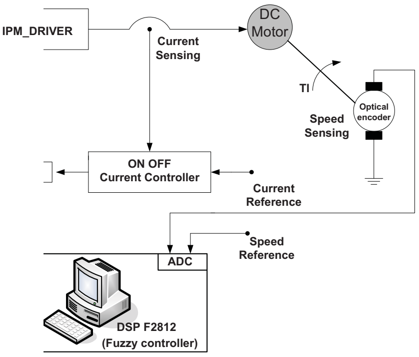


Fig. 2: Black diagram of proposed system.

**Intelligent Power Module:** Intelligent Power Module (IPM) is advanced hybrid power devices that combine high speed, low loss IGBTs with optimized gate drive and protection circuitry. System reliability is further enhanced by the IPM's integrated over temperature and under voltage lock out protection. Compact, automatically assembled Intelligent Power Modules are designed to reduce system size, cost, and time. In this paper, we use AC/DC converter based on Intelligent Power Module (IPM module, IRAM10UP60B-2) to driver DC motor, equivalent circuit of the IPM is shown in Fig.3.

The function of a fuzzy controller is to convert linguistic control rules based on expert knowledge into control strategy [23]. The effective and efficient control using fuzzy logic has emerged as a tool to deal with uncertain, imprecise or qualitative decision making problems [24]-[30]. The FLC consists of mainly four stages, namely Fuzzification, Rule-base, Inference engine and Defuzzification. The Takagi- Sugeno type controller is performed for present control for DC motor because it has singleton membership in the output variable. Moreover, it can be easily implemented and number of calculations can be reduced and works well with optimization and adaptive techniques which makes it very attractive in control problems, particularly for dynamic nonlinear systems [31]. General structure of the fuzzy controller is shown in Fig. 4, and the simulink model developed is given in Fig.5.

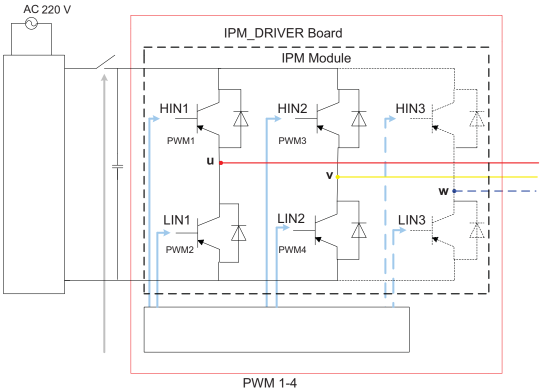


Fig. 3: Equivalent circuit of the IPM.

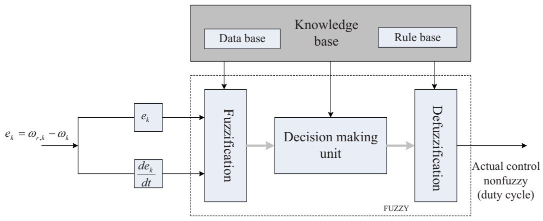


Fig. 4: General structure of the fuzzy controller .

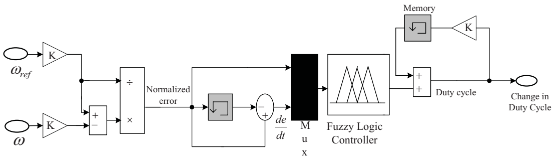


Fig. 5: Simulink scheme of the fuzzy controller.

## *A. Fuzzification*

In this work, the motor variables considered are speed $\omega$ and current $i_a$. The speed $\omega$ is the control object of FLC. Let $\omega_r$ denote the reference speed, then the definitions for error $e_k$ and change in error $\Delta e_k$ are given in (7) and (8).

$$e_k = \omega_{r,k} - w_k \tag{7}$$

$$\Delta e_k = e_k - e_{k-1} \tag{8}$$

TABLE I: Rule based of the system

| E<br>CE | NB | NS | O | PS | PB |
|---|---|---|---|---|---|
| NB | NB | NB | NB | NS | Z |
| NS | NB | NB | NS | Z | PS |
| O | NB | NS | Z | PS | PS |
| PS | NS | Z | PS | PB | PB |
| PB | Z | PS | PB | PB | PB |

Five linguistic variables are used for fuzzificating the input variable $e_k$ and $\Delta e_k$ as follows, Negative Big $(NB)$, Negative Small $(NS)$, Zero $(Z)$, Positive Small $(PS)$ and Positive Big $(PB)$. In this work only five membership functions are used for the input, i.e. error and change in error. In order to reduce the number of membership function the width of the membership functions are kept different. The membership function width for the center membership functions is considered narrow and wide towards outer.

### B. defuzzification

After fuzzy reasoning we have a linguistic output variable which needs to be translated into a crisp value. The objective is to derive a single crisp numeric value that best represents the inferred fuzzy values of the linguistic output variable. The linguistic variables are converted in to a numerical variable [32]. The defuzzified output is the duty cycle $dc_k$. The change in duty cycle $\Delta dc_k$ can be obtained by adding the pervious duty cycle $pdc_k$ with the duty cycle $dc_k$ which is given in Eq.(9).

$$\Delta dc_k = dc_k + pdc_k \tag{9}$$

### C. Rule table and Inference Engine

The Inference Mechanism provides the mechanism for invoking or referring to the rule base such that the appropriate rules are fired. The control rules related to the fuzzy output to the fuzzy inputs are derived from general knowledge of the system behavior, also the perception and experience. However, some of the control rules are developed using “trial and error” method [14]. The general rule can be written as “If $e(k)$ is $X$ and $\Delta e(k)$ is $Y$ then $\Delta dc(k)$ is $Z$”, where $X, Y$and $Z$ are the fuzzy variable for $e(k)$, $\Delta e(k)$ and $\Delta dc(k)$ respectively . The rule table for the designed fuzzy controller is given in the Table I [33].

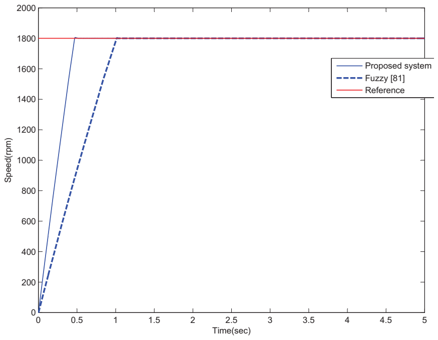


Fig. 6: Comparisons between the proposed system and fuzzy controller reference [34] for $\omega_r = 1800$ rpm .

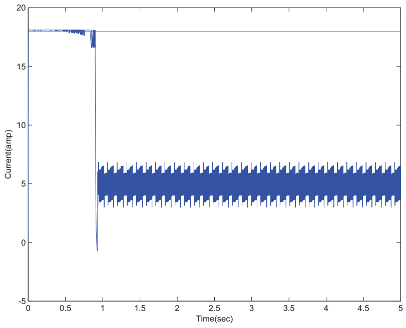


Fig. 7: Current variation with respect to time .

## IV. Simulation and Discussion of the Proposed System

In this section, the effectiveness of proposed model has been simulated using Matlab/simulink toolbox, the designed fuzzy controller (FC) and AC/DC converter based on Intelligent Power Module is tested. The simulation results of the proposed fuzzy controller(FC) are compared with fuzzy controller (FC) [34]. Fig.6 shows the comparisons between the proposed system fuzzy controller and fuzzy controller reference [34] for $\omega_r = 1800$ rpm , and Fig.7 is shown Current Variation respect to time. We can observed that, the current variation includes some harmonics due to AC to DC converter. In order to show the robustness of the proposed system, performance comparisons of the proposed system with reference [12] and reference [34] for the speed $\omega_r = 1800$ rpm is shown in the Table II.

The simulated result of speed regulation for a step change in load torque of 50% and 100% applied at $t = 2.5sec$ are shown in Fig. 8 and Fig. 9 respectively.

Form these figures, we can observed that the load

TABLE II: The performance comparisons of the proposed system with reference [12] and reference [34] .

| Controller | [12] | [34] | current |
|---|---|---|---|
| Setting time | 1.7 sec | 1sec | 0.7 |
| Max.over shoot | 3.2% | 0.36% | 0.30% |

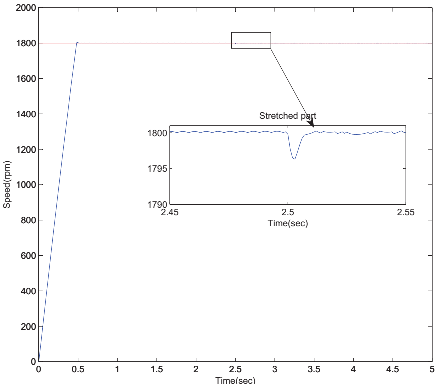


Fig. 8: Speed variation for the step change in load torque $(\Delta T_l = 50\%)$ applied at t= 2.5 sec for $\omega_r = 1800$ rpm.

influences the performance of the controller. Furthermore, it is also observed that the motor speed is function of the load torque and it seen that when load is applied the motor takes it is sufficient time to reach the reference speed.

In order to validate the proposed method, the compared between proposed method and fuzzy controller (FC) [34] for speed $\omega_r = 1800$ rpm and load torque $\Delta T_l = 5N.m$ applied at time $t = 2.5sec$ is shown in Fig. 10. Furthermore, the comparison of the proposed system with reference [12] and reference [34] for the speed $\omega_r = 1800$ rpm and $\Delta T_l = 5N.m$ applied at $t = 2.5sec$ is shown in the Table III. Fig. 11 shown current variation for step change in load torque $(\Delta T_l = 5N.M)$ applied at t= 2.5 sec for $\omega_r = 1800$

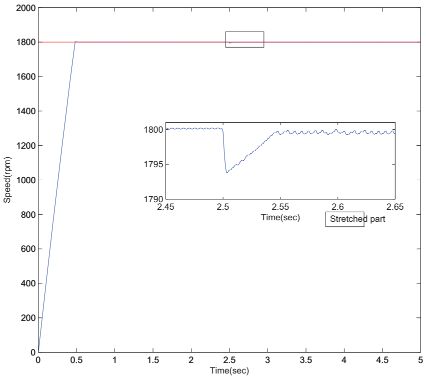


Fig. 9: Speed variation for the step change in load torque $(\Delta T_l = 100\%)$ applied at t= 2.5 sec for $\omega_r = 1800$ rpm.

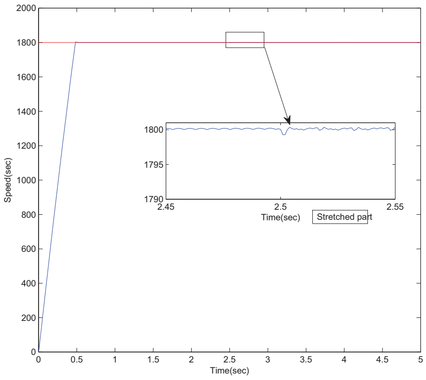


Fig. 10: Speed variation for the step change in load torque $(\Delta T_l = 5N.M)$ applied at t= 2.5 sec for $\omega_r = 1800$ rpm.

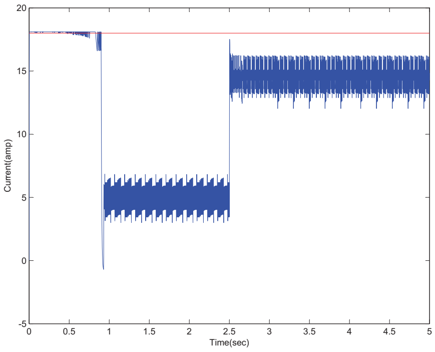


Fig. 11: Current variation for the step change in load torque $(\Delta T_l = 5N.m)$ applied at t= 2.5 sec for $\omega_r = 1800$ rpm.

## V. Conclusion

This paper consider the fuzzy logic controller for control DC motor. In the proposed method, the mathematic model of DC motor is applied to evaluate the fuzzy logic controller, and a IPM_DRIVE (IPM power modules) is used to fed dc motor . The effective results show that the

TABLE III: Comparison of the proposed system with reference [12] and reference [34] speed $\omega_r = 1800$ rpm at $\Delta T_l = 5N.m$ applied at $t = 2.5sec$ .

| Controller | [12] | [34] | Current |
|---|---|---|---|
| Max.over shoot | 3.2% | 0.36% | 0.30% |
| Max. speed drop | 3.5% | 0.36% | 0.25% |
| Recovery time | 2.4 sec | 0.01 sec | 0.007 sec |

performance of DC motor controller has obtained better results by using Fuzzy Logic Controller and (IPM power modules) fed the proposed system. The designed fuzzy logic controller also is implemented in a micro-controller.

## REFERENCES

[1] J. Qiu, G. Feng, H. Gao, "Fuzzy-model-based piecewise $H_\infty$ static output feedback controller design for networked nonlinear systems," IEEE Transation on Fuzzy Systems, vol. 18, no. 5, 919-934, Oct. 2010.

[2] J. Qiu, G. Feng, H. Gao, "Asynchronous output feedback control of networked nonlinear systems with multiple packet dropouts: T-S fuzzy affine model based approach," IEEE Transation on Fuzzy Systems, vol. 19, no. 6, 1014-1030, Dec. 2011.

[3] C.C. Lee, "Fuzzy logic in control system: Fuzzy Logic Controller- Part, I and II," IEEE Trans. on Systems Man and Cybernetics, Vol. 20, pp. 404 - 435,1990.

[4] Y. Kung and C. Liaw, "A fuzzy controller improving a linear model following controller for motor drives," IEEE Transactions on Fuzzy Systems, Vol. 2, no. 3 , pp.194-202, 1994.

[5] Muhammad H.Rashid, Power electronics circuits, devices and applications, 2d edition, Prentice-Hall, 1993.

[6] Philip.T.Krein, Elements of power Electronics, Oxford University Press, 1998.

[7] N.C. Jagan, Control System Second Edition, BS Publications 4-4-309, Giriraj Lane, Sultan Bazar, Hyderabad 2008.

[8] Prof. Dr.-Ing. Dr."h.c. , Control of Electrical Drives, Leonhard Technische University Braunschweig Institute for Regelungs technik published New Work 1997.

[9] S. Thompson, Control Systems Engineering and Design, Department of Mechanical and Manufacture Engineering, University of Belfast. UK Led1989.

[10] B. Chen, H. Uang and C. Tseng, "Robust tracking enhancement of robot systems including motor dynamics: a fuzzy-based dynamic game approach," IEEE Transactions on Fuzzy Systems, Vol. 6, no. 4 , pp.538-552, 1998.

[11] N.Senthil Kumar, V.Sadasivam, K.Prema, "Design and simulation of fuzzy controller for closed loop control of chopper fed embedded dc drives," IEEE international conference, POWERCON, Singapore, 2004.

[12] H.A.Yousef and H.M.Khalil "A fuzzy logic-based control of series DC motor drives," Proceedings of the IEEE International Symposium on Vol. 2, Issue 10-14, pp. 517 C 522, Jul1995.

[13] H.L.Tan"A simplified fuzzy logic controller for DC series motor with Improve performance," IEEE International Conference on Fuzzy System, Pp. 1523-1526, 2001.

[14] T.Gupta and R.Boudreax, "Implementation of a Fuzzy Controller for DC-DC Converters Using an Inexpensive 8-Bit micro controller," IEEE Trans. on Industrial Electronics, vol. 44, no.5, pp.661-667, October 1997.

[15] G.Uma and C.Chellamuthu, "Design and implementation of fuzzy logic control speed control system for a converter fed DC drive using 8097 micro controller," ISIE, Mexico 2000.

[16] Soliman. H. F, Mansour. M. M, Kandil. S. A, and Sharaf. A. M, "A robust tunable fuzzy logic control scheme for speed regulation of DC series motor drives," Electrical and Computer Engineering, Canadian Conference, Vol. 1, Issue 5-8, pp. 296 C 299, Sep 1995. Y.F.

[17] Li and C.C. Lau, "Development of fuzzy algorithms for servo system," IEEE Control System Magazine, April, 1989, pp 65-11.

[18] Hung and Ching Lu, "Design and implemetation of a digitalized fuzzy controller for dc sero drives", IEEE International Conference on Intelligent Processing systems ,China October 1997.

[19] Y. Tipsuwan and M. Y. Chow, "Fuzzy logic microcontroller implementation for DC motor speed control," in IEEE IECON'99, San Jose, CA, 1999.

[20] Adel E. El-kholy and A. M. Dabroom, "Adaptive Fuzzy Logic Controllers for DC Drives: A Survey of the State of the art," Journal of Electrical Systems, pp. 116-145, 2006.

[21] A.E. Fitzgerald and Charles kingsley , JR. Alexander KUSKO "Electric machinery" Third Edition United State of America. 1971.

[22] Y. F. Li, and C. C. Lau, " Development of fuzzy algorithms for servo systems," IEEE Control Systems Mag., vo1.9, no,3, pp.65-72, Apr. 1989.

[23] B. S. Zhang, and J .M. Edmunds, "On fuzzy logic controllers," IEE International Conference on Control, pp.961-965, Edinburg, U. K., 1991.

[24] H. F. Soliman, M. M. Mansour,S. A. Kandil, and A. M. Sharaf, "A robust tunable fuzzy logic control scheme for speed regulation of DC series motor drives," Electrical and Computer Engineering, Canadian Conference, Vol. 1, Issue 5-8, pp. 296-299, Sep 1995.

[25] G. Acciani, G. Fornarelli, and A. Giaquinto, "A fuzzy method for global quality index evaluation of solder joints in surface mount technology," IEEE Transation on Industrial Informtics, vol. 7, no. 1, pp.115-124, Feb. 2011.

[26] K. Rajani, Mudi, and R. Pal Nikhil, Member, "A Robust Self-Tuning Scheme for PI- and PD-Type Fuzzy Controllers," IEEE Transactions On Fuzzy Systems, vol. 7, no. 1, pp. 2-16, Feb. 1999.

[27] H. L. Tan, N. A. Rahim, and W. P.Hew, "A dynamic input membership scheme for a fuzzy logic DC motor controller," IEEE International Conference on Fuzzy Systems vol. 1, Issue 25-28, pp. 426-429 May 2003.

[28] Young Im Cho, "Development of a new neuro-fuzzy hybrid system," Industrial Electronics Society, IECON, 30th Annual Conference of IEEE vol. 3, pp. 3184-3189, 2004.

[29] Ping-Zong Lin, Chun-Fei Hsu and Tsu-Tian Lee, "Type -2 Fuzzy Logic Controller design for Buck DC-Dc converters, " Proceedings of the IEEE International Conference on Fuzzy systems, pp. 365- 370, 2005.

[30] Jabri Majed, Chouiref Houda, Jerbi Houssem, Benhadj Braiek and Naceur, "Fuzzy logic parameter estimation of an electrical system," Systems, Signals and Devices, IEEE SSD, 5th International Multi-Conference, pp. 1-6, 2008.

[31] L. A. Zadeh, "Fuzzy sets," Information and Control, Vol. 8, pp. 338-352, 1965.

[32] L. A. Zadeh, "Fuzzy algorithms," Info. and Ctl., Vol. 12, pp. 94-102, 1968.

[33] N. Senthil Kumar, V. Sadasivam and M. Muruganandam, "A Low-cost Four-quadrant Chopper-fed Embedded DC Drive Using Fuzzy Controller," Inter National Journal of Electric Power Components and Systems, Vol. 35, Issue 8, pp. 907-920, August 2007.

[34] M. Muruganandam and M. Madheswaran, "Modeling and Simulation of Modified Fuzzy Logic Controller for Various types of DC motor Drives," IEEE International Conference on Control,Automaion, Communication and Energy conservation, June 2009.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Design of Web-based Smart Home with 3D Virtual Reality Interface

Wenshan Hu, Hong Zhou, Chaoyang Lin, Xianfeng Chen and Zhen Chen
Department of Automation,
Wuhan University,
Wuhan, 430072, China
Wenshan.hu@whu.edu.cn

Yiyan Lu
School of Civil Engineering
Wuhan University,
Wuhan, 430072, China
yylu901@163.com

*Abstract*—**In this paper, the design of the a web-based Smart Home system is introduced. The proposed system provides a web interface through which users are able to check the home status and control the domestic appliance remotely as long as they have a PC system connected to the Internet. In order to give users a more vivid way to access the Smart Home system, a web based 3D interface with virtual reality technology is proposed in this paper. The rooms, appliances and furniture are reconstructed in the web-based interface. Users can "look around" in the virtual home remotely. They are able to check the security alarms, control the appliances in the 3D virtual reality similar as they do in real homes.**

***Keywords-smart home; web-based interface; 3D virtual reality; wireless network***

## I. INTRODUCTION

Recently, smart home has emerged as a hot spot both in academic and industrial communities [1]. Smart Home is the term commonly used to define a residence that integrates technology and services through home networking to enhance power efficiency and improve the quality of living. With the advent of the smart devices including Smart TV and smart pad, the various types of service [2] for smart home are expected to appear in the market.

A lot of smart home features such as home entertainments, surveillance and access control, resource effective management [3], home appliances automation, environmental control, assistive computing and healthcare came into birth so that lead to a promising industrial market in last two decades [4].

Today's Internet is evolving into an "Internet of things," as soon there will be more than one trillion connected devices. By 2013, 1.2 billion connected consumer electronics devices are expected in the more than 800 million homes with broadband connections.

With the rapid development of the Internet, new smart homes are moved out from the home onto the network where users are able to monitor and control the domestic appliances remotely check the status of the home from any place any where as long as they have Internet access.

In recent years, web-based 3D technologies have been widely used in many applications such as online games [5], urban management [6], remote experimentation [7] and GIS [8] *etc*. 3D objects can be imported and rendered in web browsers. If the homes can be reconstructed in web based virtual reality world, the users can access their homes remotely in a more vivid way.

In this paper, the design of a web-based smart home is introduced. It can be accessed remotely using web browsers. Web-based 3D technology is adopted in the design to provide a more vivid remote access interface. They are able to "walk around" in the virtual homes and control the appliances freely as if they are at home.

## II. STRUCTURE OF SMART HOME

### A. General Strucrue

It is a trend that telephone, television and computer networks merging in one optical fiber network with the development communication technology. The proposed Smart Home system is based on the Triple Play communication architecture as shown in Figure 1.

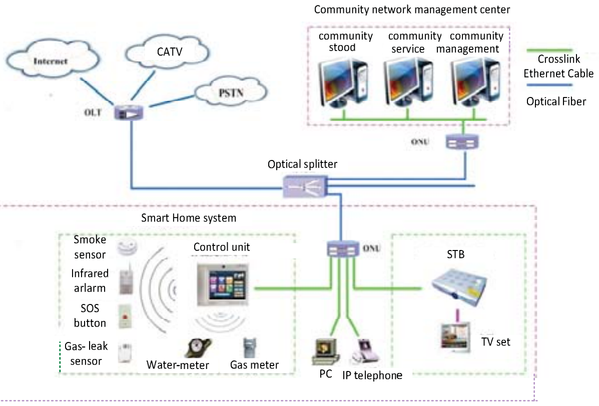


Figure 1. Typical structure of smart home with triple play communication architecture

The proposed Smart Home system uses wireless communication to control various domestic appliances in order to reduce the cost of cabling and increase the flexibilities. Small wireless communication modules are installed on domestic appliances, meters and security sensors so they can be accessed by the control unit through wireless communication.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

The Control Unit is also connected to a central web server located in the community center. The users are able to control the appliances, monitor the real-time images and check the power consumption at home through the web server remotely. Moreover, if any security alarm is trigged, the message is also transmitted to the central web server where a text message is sent to the corresponding user's mobile phone automatically.

### *B. Structure of the Hardware*

The network architecture of smart home system inside the home consists of two parts, as shown in Figure 2. The Control Unit is the central device of the system. It mainly consists of the main board, LCD screen, CMOS camera, wireless module, microphone and speaker. The main board is based on Samsung's S3C6410 ARM11 processor and integrates various functions such as video signal processing, camera control, USB, SD card, LCD screen, Ethernet. It is also equipped with a buzzer, a temperature sensor, a back-up lithium battery holder to ensure that the system time is not lost after power failure and other equipments.

In addition to the control unit, smart home system has many kinds of modules, such as security modules, domestic appliances control modules, smart meter modules. The security modules include window and door magnetic sensor, smoke sensor, gas leak sensor, infrared sensor, SOS button, etc. The domestic appliances control modules are used to control appliances, including light, air conditioner, curtain *etc*. The smart meter modules are used to collect the real time reading of the electricity, water and gas meters. The environmental monitoring modules are used to monitor the environment in house, including temperature and humidity. In addition, the smart socket can measure the real time status of the appliances, including voltage, current and power. These modules are connected to the central Control Unit using wireless communication. The status of the domestic devices is transmitted to the control unit in real time.

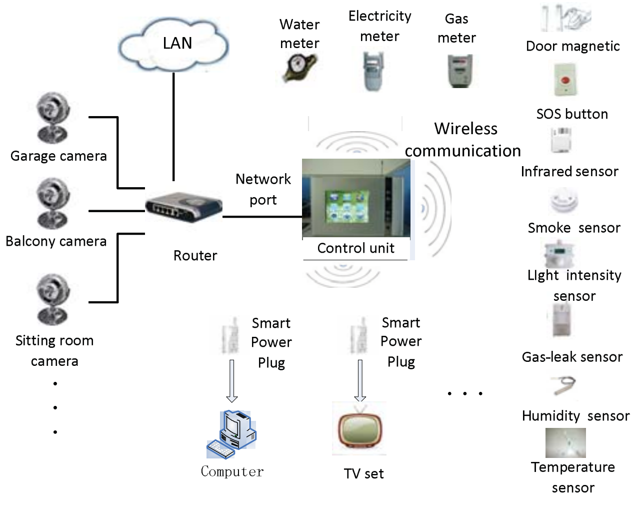
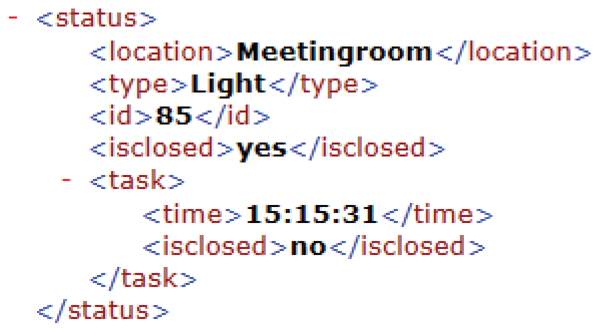


Figure 2. Structure of the Smart Home system

The Smart Home System uses wireless network to establish the communication between the control unit and home devices (such as appliances, security alarms and smart meters *etc*). The wireless module communicates with a variety of sensor modules, domestic appliances control modules, data acquisition modules by a nRF905 single-chip wireless transceiver which operates at 433MHz band. The use of the wireless network has simplified the system structure and reduced the cost of cabling greatly comparing with the wired solutions.

## III. SOFTWARE STRUCTURE OF THE CONTROL UNIT

The operating system running in the Control Unit is Linux and the embedded software is developed using Qt which is a cross-platform application and UI framework. The device table represented in XML (Extensible Markup Language) form is the data core of the Control Unit software. Each device such as domestic appliances, security sensors has a corresponding element in the XML, where its, id, type, locations and status etc are specified. The control Unit can operate on the devices through wireless communication according to information stored in the XML device table. Figure 3 is an example of a light represented in the XML table. The XML elements are synchronized with the real devices. Every time an operation is implemented on a device or the status of a device changes, the corresponding XML elements are modified.



```
- <status>
    <location>Meetingroom</location>
    <type>Light</type>
    <id>85</id>
    <isclosed>yes</isclosed>
  - <task>
      <time>15:15:31</time>
      <isclosed>no</isclosed>
    </task>
  </status>
```

Figure 3. An example of device table

Figure 4 shows the software structure of Control Unit. The network communication module establishes TCP connection channels with central web server. It get the information of the home devices from the XML table and transmit them to the web sever. The wireless communication module keeps connection with home devices. It receives the real time information from the home devices and implements the corresponding modification on the XML table.

If one of the security sensors is trigged, it sends a message to the Control Unit through wireless communication. The corresponding status of the XML is modified. Then the alarm message is transmitted to the central server where a text message is sent to the user's mobile phone.

Another important function of the Control Unit is to control and monitor domestic appliances. If provide a unified user interface to control all the appliances. For example, if a user wants to operate a certain appliance on the Control Unit such as tuning on a light, the Control Unit will send a command to the corresponding appliance through wireless communication and modify the XML document at the mean time.

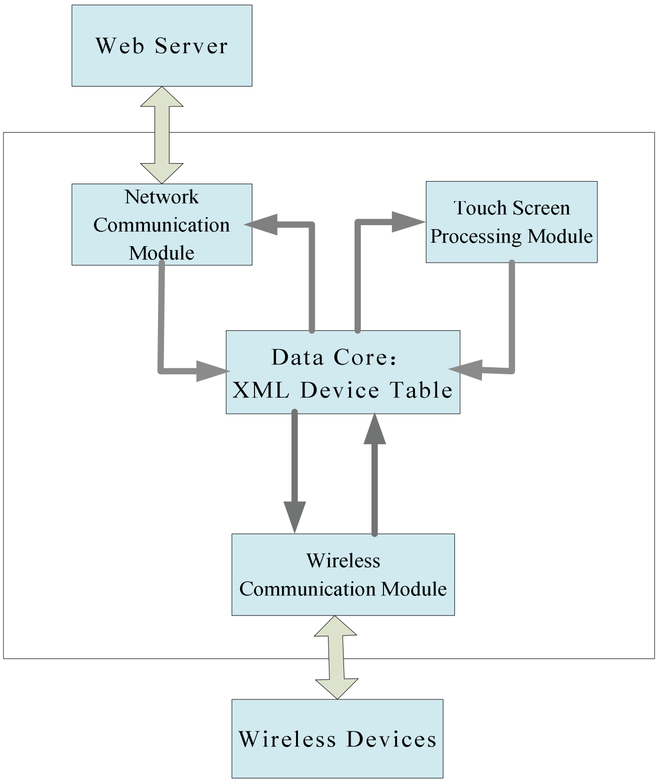

Web Server
Network Communication Module
Touch Screen Processing Module
Data Core: XML Device Table
Wireless Communication Module
Wireless Devices

Figure 4. Software structure of control unit

## IV. WEB-BASED REMOTE ACCESS INTERFACE

The Smart Home system provides a web-based remote interface for users. By a simple click of the mouse on web application, users are able to remotely and conveniently operate and control domestic appliances.

Web-based smart home remote control platform adopts B/S structure with a three-tier structure, as shown in Figure 5. In the overall architecture, Tomcat web server and MySQL database are deployed in the central server located in the community center. Three software modules are designed and implemented in the web server. The Communication Module is responsible for data communication between the wireless devices and Control Unit through the TCP protocol. JSP and Servlets are deployed to dynamically generate web interface for users to remotely manipulate the domestic electronics.

MySQL database is deployed as the middle layer which is the kernel data structure of the web server. All the information such as the user list, device list and security alarm record are stored in the database.

Web services server consists of database server and local servers, it communicates with the Control Unit through TCP connections. A variety of data provided by the Control Unit is sent to the server in XML form. Security alarm status, power consumption information and indoor environment information are packed in XML document and sent to the server.

Web Services server has played the role of a bridge in the whole system. It can not only obtain data form the Control Unit, but also pass the commands (such as tuning on or off a appliance) to from the web client to the Control Unit. These commands are also packed in XML which can be parsed and executed by the Control Unit.

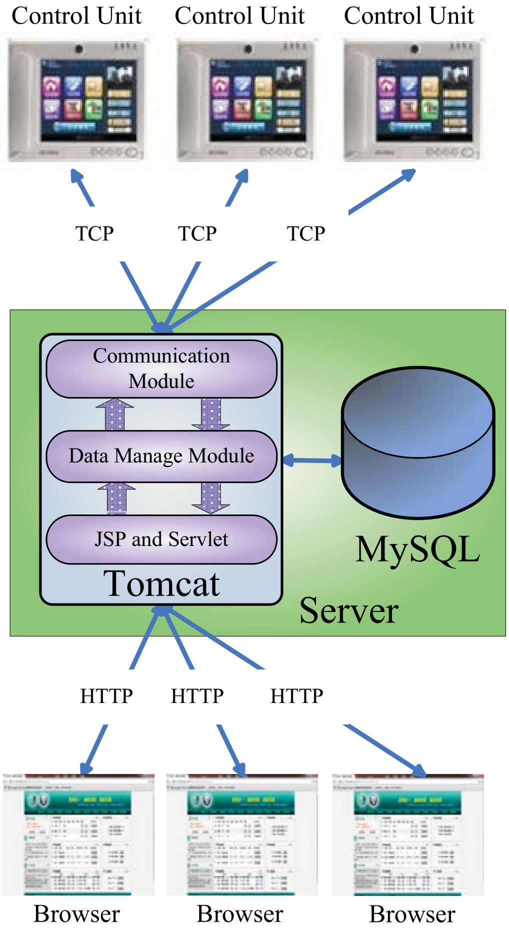
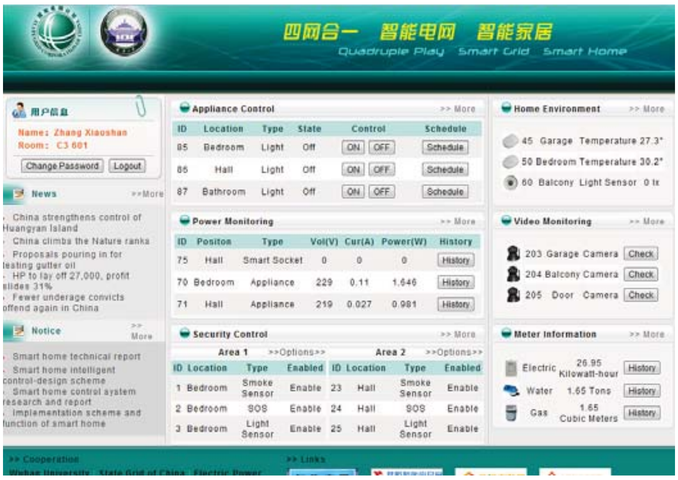


Figure 5. Overall architecture

The web-based Smart Home interface is shown in Figure 6. It can be seen that the main page containing user information, home appliance control, power monitoring, alarm monitoring, indoor environment, video monitoring, meter information *etc*.


四网合一 智能电网 智能家居
Quadruple Play Smart Grid Smart Home

用户信息
Name: Zhang Xiaoshan
Room: C3 601
Change Password | Logout

News >>More
- China strengthens control of Huangyan Island
- China climbs the Nature ranks
- Proposals pouring in for testing gutter oil
- HP to lay off 27,000, profit slides 31%
- Fewer underage convicts offend again in China

Notice >> More
- Smart home technical report
- Smart home intelligent control-design scheme
- Smart home control system research and report
- Implementation scheme and function of smart home

Appliance Control >> More

| ID | Location | Type | State | Control | Schedule |
|---|---|---|---|---|---|
| 85 | Bedroom | Light | Off | ON OFF | Schedule |
| 86 | Hall | Light | Off | ON OFF | Schedule |
| 87 | Bathroom | Light | Off | ON OFF | Schedule |

Power Monitoring >> More

| ID | Positon | Type | Vol(V) | Cur(A) | Power(W) | History |
|---|---|---|---|---|---|---|
| 75 | Hall | Smart Socket | 0 | 0 | 0 | History |
| 70 | Bedroom | Appliance | 229 | 0.11 | 1.646 | History |
| 71 | Hall | Appliance | 219 | 0.027 | 0.981 | History |

Security Control >> More

Area 1 >>Options>> Area 2 >>Options>>

| ID | Location | Type | Enabled | ID | Location | Type | Enabled |
|---|---|---|---|---|---|---|---|
| 1 | Bedroom | Smoke Sensor | Enable | 23 | Hall | Smoke Sensor | Enable |
| 2 | Bedroom | SOS | Enable | 24 | Hall | SOS | Enable |
| 3 | Bedroom | Light Sensor | Enable | 25 | Hall | Light Sensor | Enable |

Home Environment >> More
- 45 Garage Temperature 27.3°
- 50 Bedroom Temperature 30.2°
- 60 Balcony Light Sensor 0 lx

Video Monitoring >> More

| | |
|---|---|
| 203 Garage Camera | Check |
| 204 Balcony Camera | Check |
| 205 Door Camera | Check |

Meter Information >> More

| | | |
|---|---|---|
| Electric | 26.95 Kilowatt-hour | History |
| Water | 1.65 Tons | History |
| Gas | 1.65 Cubic Meters | History |

>> Cooperation
Wuhan University State Grid of China Electric Power
>> Links

Figure 6. Main Page

The web pages for the Smart Home system are different from conventional ones, which are mainly based on information publication. Ajax engine is introduced into the web so that the client browser uses asynchronous mode to communicate with the server. The web-based interface keeps requesting the status of the Smart Home devices from the web server. Therefore, every time there are changes (such as that an alarm is trigged) happened on the web server, if can be reflected on the web based interface immediately.

## V. WEB-BASED 3D VIRTUAL REALITY

The web-based 3D virtual reality technology is adopted for the design of the smart home systems. When users login in their home remotely using their web browser, a web-based 3D interface is also provided for them. Their homes are reconstructed in the web browsers using 3D virtual reality. They can "walk into" their virtual homes and control the appliances remotely similar to the way they do in their real homes.

### *A. 3D Modeling*

Components (appliances, furniture and rooms) of the smart homes must be modelled in 3D before it can be imported into the web-based user interface. There are many commercial software available for the 3D model design, such as 3DS Max, Solid Works and Pro/E *etc.* The original 3D models to be displayed in the web-based interface are firstly designed using the software and then converted into a common format which can be recognised and decoded by Flash 3D engines.

In this case, 3DS Max is selected as the 3D model development tool. The models designed in the 3DS Max are exported into the Wavefront .obj format for the Flash 3D engine. Figure 7 is an example which shows the 3D model of room being designed in the 3DS Max environment.

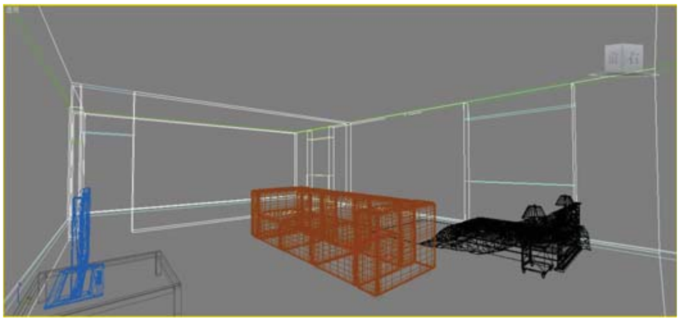


Figure 7. A 3D room designed in 3DS Max

When designing the 3D models, the complexity of these models must be considered carefully. The quality and the complexity of the 3D models have to be balanced. Even though the most powerfully 3D Flash is more than capable of processing tens thousands triangles, too complex models always result in big target files. In the internet environment, the long wait for downloading these files to the web browsers should be avoided as much as possible.

### *B. 3D Rendering*

3D models in can not be displayed in the web browser directly. It must be imported and rendered in Flash Controls. There are several 3D Flash engines available and most of them are open source software. Paperverion3D, Away3D and Sandy3D are examples. However, some early version of 3D engines only support software rendering. Without GPU (Graphic Processing Unit) acceleration, only relatively simple 3D applications can be implemented in the web-based interface, which had significantly limited the growth of the web-based 3D technologies.

The release of Flash 11 had changed the situation. Flash 11.0 starts to support hardware accelerated 3D rendering, which makes it possible to designed more complex 3D applications. Stage 3D which is a set of 3D API in Flash 11 enables advanced 3D capabilities on both PC and mobile platforms. In order to help developers to quickly design 3D web-based applications, Proscenium which is an ActionScript code library built on the top of Stage 3D has also been released. Using Proscenium, models designed in 3D design software can be easily imported and rendered into Flash Controls. The structure of a 3D Flash Controls is depicted in Figure 8.

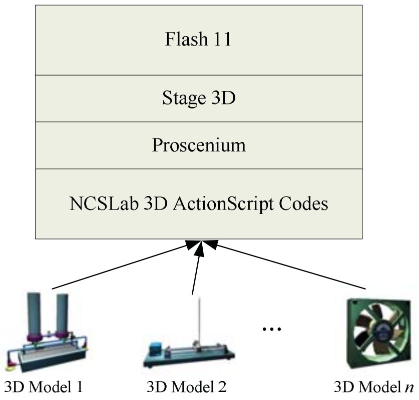


Figure 8. Structure of a 3D Flash Control

Using the resources provided by Stage 3D and Proscenium, 3D models designed in the 3DS Max environment are imported into Flash Controls which are embedded into web browser. Therefore, they can be embedded and displayed in the web-based interface. Apart from the 3D engine, some supporting codes are also designed. These codes are able to communicate with the Web Server and dynamically control the animations of the 3D components.

### *C. Web-based 3D Interface*

Figure 9 shows the structure about how the 3D models are displayed in the web-based interface. The Tomcat Web Servers create the NCSLab web pages dynamically using JSP/Servlet Technology. These pages are downloaded to the web browsers and generate the web-based interface.

On the user's web browser, AJAX scripts embedded in the HTML codes are designed to deploy the 3D Flash Controls in the web-based interface. These codes download and the corresponding 3D Flash Controls and embed them into the web browser. The 3D Flash Controls are also able to obtain real-time data from the web server. The motions of the 3D models can be manipulated according to the real-time experimental data collected from the web server, which gives the synchronization between the animations and real test rigs.

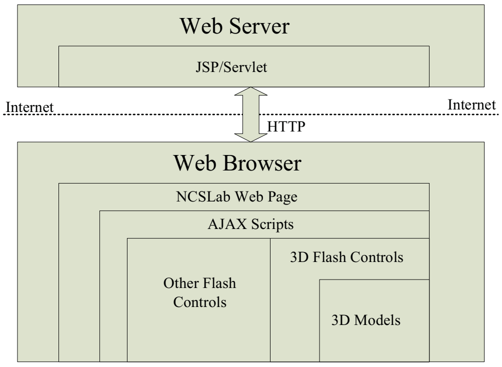


Figure 9. Structure of 3D web-based interface

Even the same 3D Flash Controls running in different circumstance may have different configurations. These configurations are generated dynamically in the XML form by the JSP/Servlet codes running on the web servers. The AJAX scripts pass the XMLs to the 3D Flash Controls where the 3D components are displayed properly according to these configurations.

Figure 10 shows a virtual home reconstructed in the web browser. The real home is displayed in the 3D virtual environment. Users are able to look around in the “virtual home” freely using the mouse the keyboards. They can check the security alarms, control the appliance freely similar to the way they do in real homes.

## VI. Conclusions

In this paper, a web-based smart home system with 3D virtual reality is introduced. In this system, wireless network is used to establish communication between the Control Unit and home devices. Comparing with the wired solution, it is able to simplify the system structure and save the cabling cost greatly. A web server is deployed in the community center. Therefore, users are able to login the web server and access their homes remotely using web-based interface. 3D virtual reality technologies are also used in the system. Rooms, appliances and furniture are reconstructed in the 3D web-based interface. Users can “look around” in the virtual home remotely. They are able to check the security alarms, control the appliances in the 3D virtual reality similar as they do in real homes.

[1] M. Chetty, J. Y. Sung and R. E. Grinter, How Smart Homes Learn: The Evolution of the Networked Home and Household, Ubiquitous Computing, Springer-Verlag Berlin Heidelberg, 2007

[2] R. J. Robles1 and T. H. Kim1, "A Review on Security in Smart Home Development," International Journal of Advanced Science and Technology, vol. 15, pp. 13-22, 2010.

[3] J. Y. Son, J. H. Park, K. D. Moon, Y. H. Lee, "Resource-Aware Smart Home Management System by Constructing Resource Relation Graph," IEEE Transactions on Consumer Electronics, vol. 57, no. 3, pp. 1112-1119, August 2011

[4] H. Hu, D. Yang, L. Fu, H. Xiang, C. Fu, J. Sang, C. Ye and R. Li, "Semantic Web-based policy interaction detection method with rules in

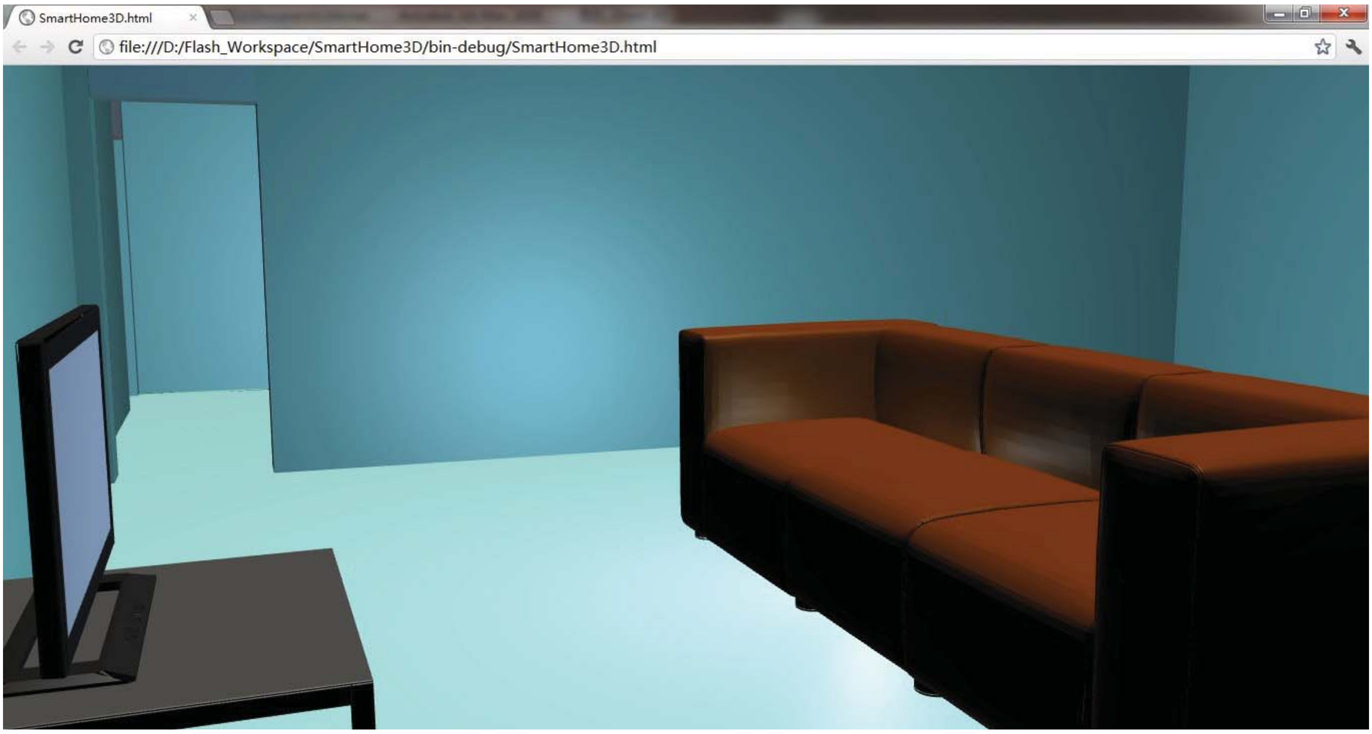

SmartHome3D.html
file:///D:/Flash_Workspace/SmartHome3D/bin-debug/SmartHome3D.html

Figure 10. 3D room Displayed in web-based interface

smart home for detecting interactions among user policies," IET Communications, vol. 5, pp. 2451–2460, 2011

[5] C. Carter, A. E. Rhalibi, M. Merabti, "Development and Deployment of Cross-Platform 3D Web-based Games," *Developments in E-systems Engineering (DESE)*, pp. 149 – 154, London, 2010

[6] F. Lamberti, A. Sanna, and E. A. H. Ramirez, "Web-based 3D visualization for intelligent street lighting," Proceedings of the 16th International Conference on 3D Web, pp. 151-154, New York, 2011

[7] W. Hu, H. Zhou and Q. Deng, "Design of web-based 3D control laboratory," *2nd International Conference on Intelligent Control and Information Processing*, pp. 590-594, Harbin, 2011

[8] M. Over, A. Schilling, S. Neubauer,and A. Zipf, "Generating web-based 3D City Models from OpenStreetMap: The current situation in Germany," *Computers, Environment and Urban Systems*, vol. 34, no. 6, pp. 496-507, Nov, 2010

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# New stability criteria for linear systems with time-varying interval delay

Jian Sun, Jie Chen, *Member, IEEE*, G.P. Liu, *Fellow, IEEE*,

*Abstract*— **This paper is concerned with the problem of stability of systems with time-varying delay in a given interval. A novel Lyapunov-Krasovskii functional is proposed to obtain new stability conditions. Some triple integral terms are introduced in the Lyapunov-Krasovskii functional and the information on the lower bound on the delay are sufficiently used. New delay-dependent stability criteria are derived using integral inequalities and formulated in terms of linear matrix inequality (LMI). Comparing numerical examples show that the proposed criteria yield a larger upper bound on the delay for a given lower bound on the delay than existing results.**

## I. INTRODUCTION

During the past few years, time-delay systems have been an active research area. Much attention has been paid to the stability and stabilization of time-delay systems. In a practical system, time-delay often deteriorates the performance of the system and even causes instability. Especially, in networked control systems, there exist time-delays in both the forward channel and the feedback channel, which poses a negative effect on the stability and performance of the systems and makes the systems difficult to analyze and synthesize [1], [2], [3], [4], [5].

Stability criteria for time-delay systems in the literature can be roughly classified into two categories. One is delay-independent and the other is delay-dependent. Generally speaking, delay-dependent stability conditions are less conservative than delay-independent ones. So, many researchers specialize in developing less conservative stability criteria for time-delay systems and some important results have been obtained [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16]. Based on some model transformations, some stability conditions for time-delay systems have been obtained in [17], [18]. A descriptor system method was proposed in [19], [20], [21] where a time-delay system is presented in the form of a descriptor system. Combining the descriptor system method with Park's inequality [22] or Moon et. al's inequality [23] can yield much less conservative results. In order to further reduce the conservatism of stability criteria, a free-weighting matrices method was proposed by He et. al [24], [25], [26]. Some free-weighting matrices are introduced by Leibniz–Newton formula to estimate the upper bound of the derivative of the Lyapunov-Krasovskii functional. Numerical examples illustrated that this method can yield less conservative results than the descriptor system method. In order to reduce the decision variables in the stability criteria, Jensen's inequality [27] was used to derive stability results for time-delay systems. It has been proved that results obtained by Jensen's inequality are generally equivalent to those obtained by descriptor system method or free-weighting matrices method [28]. For the systems with time-varying delay, [29], [30] reported that some useful terms are ignored when estimating the upper bound of the derivative of the Lyapunov functional, which can introduce significant conservatism. Inspired by this observation, some less conservative results were proposed in [29], [30], [31] by taking into these useful terms account.

In the literature, the time-varying delay is often assumed belong to a given interval, that is,

$$0 < h_1 \leqslant d(t) \leqslant h_2 \tag{1}$$

However, the information on the lower bound of the delay is not sufficiently used in the Lyapunov functional. For example, $\int_{t-h_2}^{t} x^T(s)Qx(s)ds$ and $\int_{t-d(t)}^{t} x^T(s)Rx(s)ds$ are often used as a part of the Lyapunov functional. The integral upper limits of these terms are all $t$ but not $t-h_1$, which may cause some conservatism just as proved in our previous work [32]. Similarly, some double integral terms such as $\int_{-h_2}^{-h_1}\int_{t+\theta}^{t} \dot{x}^{\mathrm{T}}(s)Z\dot{x}(s)dsd\theta$ are often used as a part of the Lyapunov functional. The inner integral upper limit is $t$ but not $t-h_1$ which may also introduce some additional conservatism. Observing this fact, a new Lyapunov functional is proposed in this paper where the information on the lower bound of the delay is sufficiently used, that is, some terms like $\int_{-h_2}^{-h_1}\int_{t+\theta}^{t-h_1} \dot{x}^{\mathrm{T}}(s)Z\dot{x}(s)dsd\theta$ are used in the Lyapunov functional. Furthermore, it has been shown in [32] that introducing some triple-integral terms in Lyaounov functional can significantly reduce the conservatism of the obtained results. In this paper, a triple integral term like $\int_{-h_2}^{-h_1}\int_{\theta}^{-h_1}\int_{t+\lambda}^{t-h_1} \dot{x}^{\mathrm{T}}(s)R\dot{x}(s)dsd\lambda d\theta$ is introduced in the Lyapunov functional. It should be noted that the upper limits of $s$, $\lambda$ and $\theta$ are $t-h_1$, $-h_1$ and $-h_1$, respectively. In this paper, a novel Lyapunov-Krasovskii functional which contains some new triple-integral terms and sufficiently uses the information on the lower bound of the delay is proposed. Some new delay-dependent stability criteria are obtained using some integral inequalities. Numerical examples illustrates

This work was supported by the Beijing Education Committee Cooperation Building Foundation Project XK100070532, National Natural Science Foundation of China under Grant 61104097 and National Science Foundation for Distinguished Young Scholars under the Grants 60925011.

Jian Sun is with School of Automation, Beijing Institute of Technology, Beijing 100081, China, and also with Key Laboratory of Complex Systems Intelligent Control and Decision, Ministry of Education, Beijing Institute of Technology, Beijing 100081, China `sunjian@bit.edu.cn`

Jie Chen is with School of Automation, Beijing Institute of Technology, Beijing 100081, China, and also with Key Laboratory of Complex Systems Intelligent Control and Decision, Ministry of Education, Beijing Institute of Technology, Beijing 100081, China `chenjie@bit.edu.cn`

G.P. Liu is with Faculty of Advanced Technology, University of Glamorgan, Pontypridd CF37 1DL, UK, and also with CTGT Center, Harbin Institute of Technology, Harbin 150001, China `gpliu@glam.ac.uk`

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

that results in this paper are significant improvements over existing ones.

*Notations:* Throughout this paper, the superscripts '-1' and 'T' stand for the inverse and transpose of a matrix, respectively; $\mathbb{R}^n$ denotes an n-dimensional Euclidean space; $\mathbb{R}^{m\times n}$is the set of all $m\times n$ real matrices; $P>0$ means that the matrix $P$ is symmetric positive definite; $I$ is an appropriately dimensional identity matrix.

## II. PROBLEM FORMULATION AND PRELIMINARIES

Consider the following linear system with time-varying interval delay:

$$\begin{aligned} &\dot{x}(t)=Ax(t)+A_1x(t-\tau(t)), \quad t>0 \\ &x(t)=\phi(t), \quad t\in[-\tau_2,\ 0] \end{aligned} \tag{2}$$

where $x(t)\in\mathscr{R}^n$ is the state vector; $A\in\mathscr{R}^{n\times n}$ and $A_1\in\mathscr{R}^{n\times n}$ are constant system matrices with appropriate dimensions; The initial condition $\phi(t)$ is a continuously differentiable vector-valued function; $\tau(t)$ is a time-varying differentiable function and satisfies

$$0<\tau_1\leqslant\tau(t)\leqslant\tau_2 \tag{3}$$

$$\dot{\tau}(t)\leqslant\mu \tag{4}$$

where $0<\tau_1<\tau_2$, and $0\leqslant\mu$ are constants.

Before moving on, the following integral inequalities are introduced.

*Lemma 1:* [27], [32] For any constant matrix $Z>0$ and scalars $\tau_2>\tau_1>0$ such that the following integrations are well defined, then

(1)

$$\begin{aligned} &-\int_{t-\tau_2}^{t-\tau_1} x^{\mathrm{T}}(s)Zx(s)ds \\ &\qquad\leqslant -\tau_{12}^{-1}\int_{t-\tau_2}^{t-\tau_1} x^{\mathrm{T}}(s)dsZ\int_{t-\tau_2}^{t-\tau_1} x(s)ds \end{aligned}$$

(2)

$$\begin{aligned} &-\int_{-\tau_2}^{-\tau_1}\int_{t+\theta}^{t-\tau_1} x^{\mathrm{T}}(s)Zx(s)dsd\theta \\ &\leqslant -2\tau_{12}^{-2}\int_{-\tau_2}^{-\tau_1}\int_{t+\theta}^{t-\tau_1} x^{\mathrm{T}}(s)dsd\theta Z\int_{-\tau_2}^{-\tau_1}\int_{t+\theta}^{t-\tau_1} x(s)dsd\theta \end{aligned}$$

where $\tau_{12}=\tau_2-\tau_1$.

The objective of this paper is to derive less conservative delay-dependent stability conditions for system (2). Using the obtained results, one can obtain a larger maximum upper bound of the delay for a given lower bound of the delay.

## III. MAIN RESULTS

In this section, some less conservative stability criteria are developed. Before presenting the main results, we define $\xi(t)=col\{x(t),x(t-\tau(t)),x(t-\tau_1),x(t-\tau_2),\dot{x}(t-\tau_1),\dot{x}(t-\tau_2),\ \int_{t-\tau_1}^{t}x(s)ds,\ \int_{t-\tau(t)}^{t-\tau_1}x(s)ds,\ \int_{t-\tau_2}^{t-\tau(t)}x(s)ds\}$, and $e_i$ $(i=1,2,\cdots,9)$ are block entry matrices. For example, $e_7^{\mathrm{T}}=[0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ I\ \ 0\ \ 0]$.

*Theorem 1:* Given scalars $0<\tau_1<\tau_2$, and $0\leqslant\mu$, if there exist matrices $P=[P_{ij}]_{5\times5}>0$, $Q=[Q_{ij}]_{2\times2}>0$, $X=[X_{ij}]_{2\times2}>0$, $S>0$, $Z_j>0$, $j=1,\cdots,4$, $R_1>0$, and $R_2>0$ with appropriate dimensions such that the following LMIs hold, then system (2) with a time-varying delay satisfying (3) and (4) is asymptotically stable.

$$\begin{aligned} \Theta_1 =\ &\Phi P\Upsilon^{\mathrm{T}}+\Upsilon P\Phi^{\mathrm{T}}+\Lambda+\Gamma^{\mathrm{T}}Y\Gamma+\Gamma^{\mathrm{T}}Q_{12}^{\mathrm{T}}e_1^{\mathrm{T}} \\ &+e_1Q_{12}\Gamma-\begin{bmatrix} e_4 & e_6\end{bmatrix}X\begin{bmatrix} e_4^{\mathrm{T}} \\ e_6^{\mathrm{T}}\end{bmatrix} \\ &+\begin{bmatrix} e_3 & e_5\end{bmatrix}(X-Q)\begin{bmatrix} e_3^{\mathrm{T}} \\ e_5^{\mathrm{T}}\end{bmatrix} \\ &-(e_1-e_3)Z_1(e_1^{\mathrm{T}}-e_3^{\mathrm{T}}) \\ &-2(e_2-e_4)Z_2(e_2^{\mathrm{T}}-e_4^{\mathrm{T}})-(e_3-e_2)Z_2(e_3^{\mathrm{T}}-e_2^{\mathrm{T}}) \\ &-e_8Z_4e_8^{\mathrm{T}}-2e_9Z_4e_9^{\mathrm{T}}-(\tau_1e_1-e_7)R_1(\tau_1e_1^{\mathrm{T}}-e_7^{\mathrm{T}}) \\ &-(\tau_{12}e_3-e_8-e_9)R_2(\tau_{12}e_3^{\mathrm{T}}-e_8^{\mathrm{T}}-e_9^{\mathrm{T}})<0 \end{aligned} \tag{5}$$

$$\begin{aligned} \Theta_2 =\ &\Phi P\Upsilon^{\mathrm{T}}+\Upsilon P\Phi^{\mathrm{T}}+\Lambda+\Gamma^{\mathrm{T}}Y\Gamma+\Gamma^{\mathrm{T}}Q_{12}^{\mathrm{T}}e_1^{\mathrm{T}} \\ &+e_1Q_{12}\Gamma-\begin{bmatrix} e_4 & e_6\end{bmatrix}X\begin{bmatrix} e_4^{\mathrm{T}} \\ e_6^{\mathrm{T}}\end{bmatrix} \\ &+\begin{bmatrix} e_3 & e_5\end{bmatrix}(X-Q)\begin{bmatrix} e_3^{\mathrm{T}} \\ e_5^{\mathrm{T}}\end{bmatrix} \\ &-(e_1-e_3)Z_1(e_1^{\mathrm{T}}-e_3^{\mathrm{T}}) \\ &-(e_2-e_4)Z_2(e_2^{\mathrm{T}}-e_4^{\mathrm{T}})-2(e_3-e_2)Z_2(e_3^{\mathrm{T}}-e_2^{\mathrm{T}}) \\ &-2e_8Z_4e_8^{\mathrm{T}}-e_9Z_4e_9^{\mathrm{T}}-(\tau_1e_1-e_7)R_1(\tau_1e_1^{\mathrm{T}}-e_7^{\mathrm{T}}) \\ &-(\tau_{12}e_3-e_8-e_9)R_2(\tau_{12}e_3^{\mathrm{T}}-e_8^{\mathrm{T}}-e_9^{\mathrm{T}})<0 \end{aligned} \tag{6}$$

where

$$\begin{aligned} &\Phi=[e_1\ \ e_3\ \ e_4\ \ e_7\ \ e_8+e_9] \\ &\Upsilon=\left[\Gamma^{\mathrm{T}}\ \ e_5\ \ e_6\ \ e_1-e_3\ \ e_3-e_4\right] \\ &\Lambda=\mathrm{diag}\{Q_{11}+\tau_1Z_3,\ -(1-\mu)S,\ +S+\tau_{12}^2Z_4,\ 0 \\ &\qquad \tau_{12}^2Z_2+\frac{\tau_{12}^4}{4}R_2,\ 0,\ -Z_3,\ 0,\ 0\} \\ &\Gamma=[A\ \ A_1\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0] \\ &Y=Q_4+\tau_1^2Z_1+\frac{\tau_1^4}{4}R_1 \end{aligned}$$

*Proof:* Choose a Lyapunov functional as follows:

$$\begin{aligned} V(x_t)=\ &\rho^{\mathrm{T}}(t)P\rho(t)+\int_{t-\tau_1}^{t}\zeta^{\mathrm{T}}(s)Q\zeta(s)ds \\ &+\int_{t-\tau_2}^{t-\tau_1}\zeta^{\mathrm{T}}(s)X\zeta(s)ds \\ &+\int_{t-\tau(t)}^{t-\tau_1}x^{\mathrm{T}}(s)Sx(s)ds \\ &+\int_{-\tau_1}^{0}\int_{t+\theta}^{t}\tau_1\dot{x}^{\mathrm{T}}(s)Z_1\dot{x}(s)dsd\theta \\ &+\int_{-\tau_2}^{-\tau_1}\int_{t+\theta}^{t-\tau_1}\tau_{12}\dot{x}^{\mathrm{T}}(s)Z_2\dot{x}(s)dsd\theta \\ &+\int_{-\tau_1}^{0}\int_{t+\theta}^{t}\tau_1x^{\mathrm{T}}(s)Z_3x(s)dsd\theta \\ &+\int_{-\tau_2}^{-\tau_1}\int_{t+\theta}^{t-\tau_1}\tau_{12}x^{\mathrm{T}}(s)Z_4x(s)dsd\theta \end{aligned}$$

$$+\int_{-\tau_1}^{0}\int_{\theta}^{0}\int_{t+\lambda}^{t}\frac{\tau_1^2}{2}\dot{x}^{\mathrm{T}}(s)R_1\dot{x}(s)dsd\lambda d\theta$$
$$+\int_{-\tau_2}^{-\tau_1}\int_{\theta}^{-\tau_1}\int_{t+\lambda}^{t-\tau_1}\frac{\tau_{12}^2}{2}\dot{x}^{\mathrm{T}}(s)R_2\dot{x}(s)dsd\lambda d\theta \quad (7)$$

where $\rho(t) = col\{x(t), x(t-\tau_1), x(t-\tau_2), \int_{t-\tau_1}^{t} x(s)ds, \int_{t-\tau_2}^{t-\tau_1} x(s)ds\}$, $\zeta(s) = col\{x(s), \dot{x}(s)\}$.
Taking the derivative of $V(x_t)$ along the trajectory of system (2) yields

$$\begin{aligned}
\dot{V}(x_t) = & 2\rho^{\mathrm{T}}(t)P\dot{\rho}(t) \\
& + \zeta^{\mathrm{T}}(t)Q\zeta(t) - \zeta^{\mathrm{T}}(t-\tau_1)Q\zeta(t-\tau_1) \\
& + \zeta^{\mathrm{T}}(t-\tau_1)X\zeta(t-\tau_1) - \zeta^{\mathrm{T}}(t-\tau_2)X\zeta(t-\tau_2) \\
& + x^{\mathrm{T}}(t-\tau_1)Sx(t-\tau_1) \\
& - (1-\mu)x^{\mathrm{T}}(t-\tau(t))Sx(t-\tau(t)) \\
& + \tau_1^2\dot{x}^{\mathrm{T}}(t)Z_1\dot{x}(t) - \tau_1\int_{t-\tau_1}^{t}\dot{x}^{\mathrm{T}}(s)Z_1\dot{x}(s)ds \\
& + \tau_{12}^2\dot{x}^{\mathrm{T}}(t-\tau_1)Z_2\dot{x}(t-\tau_1) \\
& - \tau_{12}\int_{t-\tau_2}^{t-\tau_1}\dot{x}^{\mathrm{T}}(s)Z_2\dot{x}(s)ds \\
& + \tau_1^2 x^{\mathrm{T}}(t)Z_1x(t) - \tau_1\int_{t-\tau_1}^{t}x^{\mathrm{T}}(s)Z_3x(s)ds \\
& + \tau_{12}^2 x^{\mathrm{T}}(t-\tau_1)Z_4x(t-\tau_1) \\
& - \tau_{12}\int_{t-\tau_2}^{t-\tau_1}x^{\mathrm{T}}(s)Z_4x(s)ds \\
& + \frac{\tau_1^4}{4}\dot{x}^{\mathrm{T}}(t)R_1\dot{x}(t) \\
& - \frac{\tau_1^2}{2}\int_{-\tau_1}^{0}\int_{t+\theta}^{t}\dot{x}^{\mathrm{T}}(s)R_1\dot{x}(s)dsd\theta \\
& + \frac{\tau_{12}^4}{4}\dot{x}^{\mathrm{T}}(t-\tau_1)R_2\dot{x}(t-\tau_1) \\
& - \frac{\tau_{12}^2}{2}\int_{-\tau_2}^{-\tau_1}\int_{t+\theta}^{t-\tau_1}\dot{x}^{\mathrm{T}}(s)R_2\dot{x}(s)dsd\theta \quad (8)
\end{aligned}$$

Using Lemma 1, it can be obtained that

$$\begin{aligned}
-\tau_1 & \int_{t-\tau_1}^{t}\dot{x}^{\mathrm{T}}(s)Z_1\dot{x}(s)ds \\
& \leq -\xi^{\mathrm{T}}(t)(e_1-e_3)Z_1(e_1^{\mathrm{T}}-e_3^{\mathrm{T}})\xi(t) \quad (9)
\end{aligned}$$

$$\begin{aligned}
-\tau_1 & \int_{t-\tau_1}^{t}x^{\mathrm{T}}(s)Z_3x(s)ds \\
& \leq -\int_{t-\tau_1}^{t}x^{\mathrm{T}}(s)dsZ_3\int_{t-\tau_1}^{t}x(s)ds \quad (10)
\end{aligned}$$

$$\begin{aligned}
-\frac{\tau_1^2}{2} & \int_{-\tau_1}^{0}\int_{t+\theta}^{t}\dot{x}^{\mathrm{T}}(s)R_1\dot{x}(s)dsd\theta \\
& \leq -\xi^{\mathrm{T}}(t)(\tau_1e_1-e_7)R_1(\tau_1e_1^{\mathrm{T}}-e_7^{\mathrm{T}})\xi(t) \quad (11) \\
-\frac{\tau_{12}^2}{2} & \int_{-\tau_2}^{-\tau_1}\int_{t+\theta}^{t-\tau_1}\dot{x}^{\mathrm{T}}(s)R_2\dot{x}(s)dsd\theta \\
& \leq -\xi^{\mathrm{T}}(t)(\tau_{12}e_3-e_8-e_9)R_2(\tau_{12}e_3^{\mathrm{T}}-e_8^{\mathrm{T}}-e_9^{\mathrm{T}})\xi(t)(12)
\end{aligned}$$

let $\alpha = (\tau(t)-\tau_1)/\tau_{12}$ and use the similar method as in [31]

$$\begin{aligned}
-\tau_{12} & \int_{t-\tau_2}^{t-\tau_1}\dot{x}^{\mathrm{T}}(s)Z_2\dot{x}(s)ds \\
= & -\tau_{12}\int_{t-\tau_2}^{t-\tau(t)}\dot{x}^{\mathrm{T}}(s)Z_2\dot{x}(s)ds \\
& - \tau_{12}\int_{t-\tau(t)}^{t-\tau_1}\dot{x}^{\mathrm{T}}(s)Z_2\dot{x}(s)ds \\
= & -(\tau_2-\tau(t))\int_{t-\tau_2}^{t-\tau(t)}\dot{x}^{\mathrm{T}}(s)Z_2\dot{x}(s)ds \\
& - (\tau(t)-\tau_1)\int_{t-\tau_2}^{t-\tau(t)}\dot{x}^{\mathrm{T}}(s)Z_2\dot{x}(s)ds \\
& - (\tau(t)-\tau_1)\int_{t-\tau(t)}^{t-\tau_1}\dot{x}^{\mathrm{T}}(s)Z_2\dot{x}(s)ds \\
& - (\tau_2-\tau(t))\int_{t-\tau(t)}^{t-\tau_1}\dot{x}^{\mathrm{T}}(s)Z_2\dot{x}(s)ds \\
\leq & -\xi^{\mathrm{T}}(t)(e_2-e_4)Z_2(e_2^{\mathrm{T}}-e_4^{\mathrm{T}})\xi(t) \\
& - \xi^{\mathrm{T}}(t)(e_3-e_2)Z_2(e_3^{\mathrm{T}}-e_2^{\mathrm{T}})\xi(t) \\
& - \alpha\xi^{\mathrm{T}}(t)(e_2-e_4)Z_2(e_2^{\mathrm{T}}-e_4^{\mathrm{T}})\xi(t) \\
& - (1-\alpha)\xi^{\mathrm{T}}(t)(e_3-e_2)Z_2(e_3^{\mathrm{T}}-e_2^{\mathrm{T}})\xi(t) \quad (13)
\end{aligned}$$

Similarly,

$$\begin{aligned}
-\tau_{12} & \int_{t-\tau_2}^{t-\tau_1}x^{\mathrm{T}}(s)Z_4x(s)ds \\
& \leq -\xi^{\mathrm{T}}(t)\left[e_8^{\mathrm{T}}Z_4e_8+e_9^{\mathrm{T}}Z_4e_9\right]\xi(t) \\
& \quad - \alpha\xi^{\mathrm{T}}(t)e_9^{\mathrm{T}}Z_4e_9\xi(t) \\
& \quad - (1-\alpha)\xi^{\mathrm{T}}(t)e_8^{\mathrm{T}}Z_4e_8\xi(t) \quad (14)
\end{aligned}$$

From (8)-(14), one can obtain

$$\dot{V}(x_t) \leqslant \xi^{\mathrm{T}}(t)\left[\alpha\Theta_1+(1-\alpha)\Theta_2\right]\xi(t) \quad (15)$$

Since $0 \leq \alpha \leq 1$, $\alpha\Theta_1+(1-\alpha)\Theta_2$ is a convex combination of $\Theta_1$ and $\Theta_2$. Therefore, $\alpha\Theta_1+(1-\alpha)\Theta_2 < 0$ is equivalent to $\Theta_1 < 0$ and $\Theta_2 < 0$. If (5)-(6) are satisfied, then system (2) is asymptotically stable. ■

*Remark 1:* A new kind of augmented Lyapunov functional is proposed in this paper to develop new delay-interval-dependent stability criteria. Being distinguished from existing Lyapunov functionals, the one in this paper contains some triple-integral terms which has been proved able to reduce the conservatism of the obtained results effectively. Furthermore, the information on the lower bound of the delay is sufficiently used in the Lyapunov functional by including the terms $\int_{-\tau_2}^{-\tau_1}\int_{t+\theta}^{t-\tau_1}\tau_{12}\dot{x}^{\mathrm{T}}(s)Z_2\dot{x}(s)dsd\theta$ and $\int_{-\tau_2}^{-\tau_1}\int_{\theta}^{-\tau_1}\int_{t+\lambda}^{t-\tau_1}\frac{\tau_{12}^2}{2}\dot{x}^{\mathrm{T}}(s)R_2\dot{x}(s)dsd\lambda d\theta$. It can be seen that the integral upper limits of these terms are $t-\tau_1$ or $-\tau_1$. To the best knowledge of the authors', this kind of Lyapunov functional has been never used in the literature. Numerical examples will be given in the next section to show that such a kind of Lyapunov functional can yield less conservative results.

*Remark 2:* Using some integral inequalities and the idea of the convex combination, new delay-interval-dependent stability criteria are obtained without introducing any free-weighting matrices. Therefore, the method proposed in this paper may involve much less variables than the well-known free-weighting matrices method and the descriptor system method.

In some circumstances, the information on the derivative of the delay may not be always available or the delay is not differentiable. For this case, the following corollary can be derived from Theorem 1 by setting $S = 0$.

*Corollary 1:* Given scalars $0 < \tau_1 < \tau_2$, and $0 \leqslant \mu$, if there exist matrices $P = [P_{ij}]_{5\times5} > 0$, $Q = [Q_{ij}]_{2\times2} > 0$, $X = [X_{ij}]_{2\times2} > 0$, $Z_j > 0$, $j = 1, \cdots, 4$, $R_1 > 0$, and $R_2 > 0$ with appropriate dimensions such that the following LMIs hold, then system (2) with a time-varying delay satisfying (3) is asymptotically stable.

$$\begin{aligned}
\hat{\Theta}_1 = {} & \Phi P \Upsilon^{\mathrm{T}} + \Upsilon P \Phi^{\mathrm{T}} + \hat{\Lambda} + \Gamma^{\mathrm{T}} Y \Gamma + \Gamma^{\mathrm{T}} Q_{12}^{\mathrm{T}} e_1^{\mathrm{T}} \\
& + e_1 Q_{12} \Gamma - \begin{bmatrix} e_4 & e_6 \end{bmatrix} X \begin{bmatrix} e_4^{\mathrm{T}} \\ e_6^{\mathrm{T}} \end{bmatrix} \\
& + \begin{bmatrix} e_3 & e_5 \end{bmatrix} (X - Q) \begin{bmatrix} e_3^{\mathrm{T}} \\ e_5^{\mathrm{T}} \end{bmatrix} \\
& - (e_1 - e_3) Z_1 (e_1^{\mathrm{T}} - e_3^{\mathrm{T}}) \\
& - 2(e_2 - e_4) Z_2 (e_2^{\mathrm{T}} - e_4^{\mathrm{T}}) - (e_3 - e_2) Z_2 (e_3^{\mathrm{T}} - e_2^{\mathrm{T}}) \\
& - e_8 Z_4 e_8^{\mathrm{T}} - 2 e_9 Z_4 e_9^{\mathrm{T}} - (\tau_1 e_1 - e_7) R_1 (\tau_1 e_1^{\mathrm{T}} - e_7^{\mathrm{T}}) \\
& - (\tau_{12} e_3 - e_8 - e_9) R_2 (\tau_{12} e_3^{\mathrm{T}} - e_8^{\mathrm{T}} - e_9^{\mathrm{T}}) < 0 \qquad (16)
\end{aligned}$$

$$\begin{aligned}
\hat{\Theta}_2 = {} & \Phi P \Upsilon^{\mathrm{T}} + \Upsilon P \Phi^{\mathrm{T}} + \hat{\Lambda} + \Gamma^{\mathrm{T}} Y \Gamma + \Gamma^{\mathrm{T}} Q_{12}^{\mathrm{T}} e_1^{\mathrm{T}} \\
& + e_1 Q_{12} \Gamma - \begin{bmatrix} e_4 & e_6 \end{bmatrix} X \begin{bmatrix} e_4^{\mathrm{T}} \\ e_6^{\mathrm{T}} \end{bmatrix} \\
& + \begin{bmatrix} e_3 & e_5 \end{bmatrix} (X - Q) \begin{bmatrix} e_3^{\mathrm{T}} \\ e_5^{\mathrm{T}} \end{bmatrix} \\
& - (e_1 - e_3) Z_1 (e_1^{\mathrm{T}} - e_3^{\mathrm{T}}) \\
& - (e_2 - e_4) Z_2 (e_2^{\mathrm{T}} - e_4^{\mathrm{T}}) - 2(e_3 - e_2) Z_2 (e_3^{\mathrm{T}} - e_2^{\mathrm{T}}) \\
& - 2 e_8 Z_4 e_8^{\mathrm{T}} - e_9 Z_4 e_9^{\mathrm{T}} - (\tau_1 e_1 - e_7) R_1 (\tau_1 e_1^{\mathrm{T}} - e_7^{\mathrm{T}}) \\
& - (\tau_{12} e_3 - e_8 - e_9) R_2 (\tau_{12} e_3^{\mathrm{T}} - e_8^{\mathrm{T}} - e_9^{\mathrm{T}}) < 0 \qquad (17)
\end{aligned}$$

where $\Phi$, $\Upsilon$, $Y$ and $\Gamma$ are the same as those defined in Therom 1, and $\hat{\Lambda} = \mathrm{diag}\{Q_{11} + \tau_1 Z_3,\ 0,\ \tau_{12}^2 Z_4,\ 0,\ \tau_{12}^2 Z_2 + \frac{\tau_{12}^4}{4} R_2,\ 0,\ -Z_3,\ 0,\ 0\}$.

*Remark 3:* When there are norm-bounded uncertainties in system (2), Theorem 1 and Corollary 1 can be extended to deal with this case following a similar method as in [9], [23].

## IV. NUMERICAL EXAMPLES

In this section, some numerical examples are given to show the effectiveness of the proposed method, that is, the method in this paper can yield less conservative results than exiting ones.

*Example 1:* Consider the following system [30], [31] with

$$A = \begin{bmatrix} -2 & 0 \\ 0 & -0.9 \end{bmatrix}, \quad A_1 = \begin{bmatrix} -1 & 0 \\ -1 & -1 \end{bmatrix}.$$

For various $\mu$ and unknown $\mu$, the maximum upper bounds of the delay (MUBDs), $\tau_2$, for given lower bound, $\tau_1$, are listed in Table I and II along with those obtained in [31], [32]. It is easy to see from Table I and II that our method can give less conservative results than those obtained in [31], [32].

TABLE I

MUBDs WITH GIVEN $\tau_1$ FOR DIFFERENT $\mu$ FOR EXAMPLE 1

| $\tau_1$ | Methods | $\mu = 0.3$ | $\mu = 0.5$ | $\mu = 0.9$ |
|---|---|---|---|---|
| 2 | [31] | 2.6972 | 2.5048 | 2.5048 |
| | [32] | 3.0129 | 2.5663 | 2.5663 |
| | Theorem 1 | 3.0168 | 2.6116 | 2.6116 |
| 3 | [31] | 3.2591 | 3.2591 | 3.2591 |
| | [32] | 3.3408 | 3.3408 | 3.3408 |
| | Theorem 1 | 3.3932 | 3.3932 | 3.3932 |
| 4 | [31] | 4.0744 | 4.0744 | 4.0744 |
| | [32] | 4.1690 | 4.1690 | 4.1690 |
| | Theorem 1 | 4.2054 | 4.2054 | 4.2054 |
| 5 | [31] | — | — | — |
| | [32] | 5.0275 | 5.0275 | 5.0275 |
| | Theorem 1 | 5.0440 | 5.0440 | 5.0440 |

TABLE II

MUBDs FOR VARIOUS $\tau_1$ AND UNKNOWN $\mu$ FOR EXAMPLE 1

| Methods | $\tau_1$ | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| [31] | $\tau_2$ | 2.5048 | 3.2591 | 4.0744 | — |
| [32] | $\tau_2$ | 2.5663 | 3.3408 | 4.1690 | 5.0275 |
| Corollary 1 | $\tau_2$ | 2.6116 | 3.3932 | 4.2054 | 5.0440 |

*Example 2:* Consider the following system [30], [31] with

$$A = \begin{bmatrix} 0 & 1 \\ -1 & -2 \end{bmatrix}, \quad A_1 = \begin{bmatrix} 0 & 0 \\ -1 & 1 \end{bmatrix}.$$

Given different lower bounds, our objective is to calculate MUBDs which keep the above system asymptotically stable. Table III lists the results for $\mu = 0.6$ and Table IV lists results for unknown $\mu$ comparing those obtained in [31], [32]. It can be seen that the results obtained in this paper are better than those in [31], [32].

*Example 3:* Consider the following system [29] with

$$A = \begin{bmatrix} -0.5 & -2 \\ 1 & -1 \end{bmatrix}, \quad A_1 = \begin{bmatrix} -0.5 & -1 \\ 0 & 0.6 \end{bmatrix}.$$

For various $\mu$, the MUBDs for given lower bound, $\tau_1$, are listed in Table V. In Table V results in [31], [32] are also listed. It is easy to see that our results are less conservative than those in [31], [32].

## V. CONCLUSIONS

In this paper, the problem of the stability of linear systems with time-varying interval delays has been investigated. New delay-interval-dependent criteria have been developed by introducing a new Lyapunov-Krasovskii functional and

TABLE III

MUBDs FOR VARIOUS $\tau_1$ AND $\mu = 0.6$ FOR EXAMPLE 2

| Methods | $\tau_1$ | 0.3 | 0.5 | 0.8 | 1 |
|---|---|---|---|---|---|
| [31] | $\tau_2$ | 1.0715 | 1.2191 | 1.4539 | 1.6169 |
| [32] | $\tau_2$ | 1.0717 | 1.2198 | 1.4558 | 1.6198 |
| Theorem 1 | $\tau_2$ | 1.0948 | 1.2588 | 1.5135 | 1.6867 |

TABLE IV

MUBDs FOR VARIOUS $\tau_1$ AND UNKNOWN $\mu$ FOR EXAMPLE 2

| Methods | $\tau_1$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| [31] | $\tau_2$ | 1.6169 | 2.4798 | 3.3894 | 4.3250 | 5.2773 |
| [32] | $\tau_2$ | 1.6198 | 2.4884 | 3.4030 | 4.3424 | 5.2970 |
| Corollary 1 | $\tau_2$ | 1.6867 | 2.5750 | 3.4878 | 4.4193 | 5.3654 |

TABLE V

MUBDs WITH GIVEN $\tau_1$ FOR DIFFERENT $\mu$ FOR EXAMPLE 3

| $\tau_1$ | Methods | $\mu = 0.2$ | $\mu = 0.5$ | $\mu = 0.7$ |
|---|---|---|---|---|
| 0.3 | [31] | 1.3831 | 1.1000 | 0.9513 |
| | [32] | 1.7022 | 1.3043 | 1.0713 |
| | Theorem 1 | 1.7856 | 1.3261 | 1.1333 |
| 0.5 | [31] | 1.3843 | 1.1000 | 1.0289 |
| | [32] | 1.8580 | 1.3940 | 1.1780 |
| | Theorem 1 | 1.9808 | 1.4216 | 1.2326 |
| 0.7 | [31] | 1.3863 | 1.1117 | 1.1115 |
| | [32] | 2.0148 | 1.4665 | 1.2898 |
| | Theorem 1 | 2.1623 | 1.4933 | 1.3365 |
| 1 | [31] | 1.3918 | 1.2493 | 1.2493 |
| | [32] | 2.2024 | 1.5214 | 1.4743 |
| | Theorem 1 | 2.3897 | 1.5709 | 1.5383 |

using the integral inequality technique and the idea of the convex combination. Due to the new construction of the introduced Lyapunov-Krasovskii functional and the sufficient use of the information on the delay interval, our results are less conservative than the existing ones. Some numerical examples have illustrated that the method proposed in this paper is efficient.

## REFERENCES

[1] G.P. Liu, Y. Xia, and D Rees, "Predictive control of networked systems with random delays," in Proceedings of the IFAC World Congress, Prague, 2005.

[2] G.P. Liu, Y.-Q. Xia, J. Chen, D. Rees, and W.-S. Hu, "Design and stability criteria of networked predictive control systems with random network delay in the feedback channel," IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews, vol. 37, no. 2, pp. 173–184, 2007.

[3] J. Sun, G.P. Liu, J. Chen, D. Rees, and W.-S. Hu, "Networked predictive control for Hammerstein systems," Asian Journal of Control, vol. 13, pp. 265–272, 2011.

[4] D. Yue, Q.-L. Han, C. Peng, "State feedback controller design of networked control systems," IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 51, 640-644, 2004.

[5] D. Yue, Q.-L. Han, J. Lam, "Network-based robust $H_\infty$ control of systems with uncertainty," Automatica, vol. 41, 999-1007, 2005.

[6] H. Gao and C. Wang, "Comments and further results on 'A descriptor system approach to $H_\infty$ control of linear time-delay systems'," IEEE Transactions on Automatic Control, vol. 48, pp. 520-525, 2003.

[7] K. Gu, Q.-L. Han, A.C.J. Luo, and S.-I. Niculescu, "Discretized Lyapunov functional for systems with distributed delay and piecewise constant coefficients," International Journal of Control, vol. 74, pp. 737-744, 2001.

[8] K. Gu, "An improved stability criterion for systems with distributed delays," International Journal of Robust and Nonlinear Control, vol. 13, pp. 819-831, 2003.

[9] Q.-L. Han, "On robust stability of neutral systems with time-varying discrete delay and norm-bounded uncertainty," Automatica, vol. 40, pp. 1087-1092, 2004.

[10] Q.-L. Han, "A descriptor system approach to robust stability of uncertain neutral systems with discrete and distributed delays," Automatica, vol. 40, pp. 1791-1796, 2004.

[11] S. Xu, J. Lam, and Y. Zou, "Further results on delay-dependent robust stability conditions of uncertain neutral systems," International Journal of Robust and Nonlinear Control, vol. 15, pp. 233-246, 2005.

[12] J. Sun, G.P. Liu, and J. Chen, "Delay-dependent stability and stabilization of neutral time-delay systems," International Journal of Robust and Nonlinear Control, vol. 19, pp. 1364–1375, 2009.

[13] J. Sun, G.P. Liu, and J. Chen, "Improved stability criteria for linear systems with time-varying delay," IET Control Theory Appl., vol. 4, No. 4, pp. 683C689, 2010.

[14] C. Peng, and Y.-C. Tian, "Delay-dependent robust stability criteria for uncertain systems with interval time-varying delay," Journal of Computational and Applied Mathematics, vol. 214, pp. 480–494, 2008.

[15] C. Lin, Q.-G. Wang, and H. Lee, "A less conservative robust stability test for linear uncertain time-delay systems," IEEE Transactions on Automatic Control, vol. 51, pp. 87–91, 2006.

[16] S.-I. Niculescu, Delay Effects on Stability: A robust control approach, Lecture Notes in Control and Information Sciences, Springer–Verlag, London, 2001.

[17] S.-I. Niculescu, On delay-dependent stability under model transformations of some neutral linear systems. International Journal of Control, vol. 74, pp. 609–617, 2001.

[18] V. Kharitonov, and D. Melchor-Aguilar, On delay-dependent stability conditions, Systems & Control Letters, vol. 40, pp. 71–76, 2000.

[19] E. Fridman and U. Shaked, "A descriptor system approach to $H_\infty$ control of time-delay systems," IEEE Transactions on Automatic Control, vol. 47, pp. 253-270, 2002.

[20] E. Fridman and U. Shaked, "An improved stabilization method for linear systems with time-delay," IEEE Transactions on Automatic Control, vol. 47, pp. 1931-1937, 2002.

[21] E. Fridman and U. Shaked, "Delay-dependent stability and $H_\infty$ control: constant and time-varying delays," International Journal of Control, vol. 76, pp. 48-60, 2003.

[22] P. Park, "A delay-dependent stability criterion for systems with uncertain time-invariant delays," IEEE Transactions on Automatic Control, vol. 44, pp. 876-877, 1999.

[23] Y.S. Moon, P. Park, W.H. Kwon, and Y.S. Lee, "Delay-dependent robust stabilization of uncertain state-delayed systems," International Journal of Control, vol. 74, pp. 1447-1455, 2001.

[24] Y. He, M. Wu, J.-H. She, and G.P. Liu, "Delay-dependent robust stability criteria for uncertain neutral systems with mixed delays," Systems & Control Letters, vol.51, pp. 57-65, 2004.

[25] Y. He, Q.-G. Wang, C. Lin, and M. Wu, "Augmented Lyapunov functional and delay-dependent stability criteria for neutral systems," International Journal of Robust and Nonlinear Control, vol. 15, pp. 923-933, 2005.

[26] M. Wu, Y. He, J.-H. She, and G.P. Liu, "New delay-dependent stability criteria for robust stability of time-varying delay systems," Automatica, vol. 40, pp. 1435-1439, 2004.

[27] K. Gu, "An integral inequality in the stability problem of time-delay systems," the 39th IEEE Conference on Decision and Control, Sydney, Australia, pp. 2805–2810, 2000.

[28] F. Gouaisbaut and D. Peaucelle, "A note on stability of time delay systems," IFAC Symposium on Robust Control Design, Toulouse, France, 2006.

[29] Y. He, Q.-G. Wang, L. Xie, and C. Lin, "Delay-range-dependent stability for systems with time-varying delay," Automatica, vol. 43, pp. 371–376, 2007.

[30] Y. He, Q.-G. Wang, L. Xie, and C. Lin, "Further improvement of free-weighting matrices technique for systems with time-varying delay," IEEE Transactions on Automatic Control, vol. 52, pp. 293–299, 2007.

[31] H. Shao, "New delay-dependent stability criteria for systems with interval delay," Automatica, vol. 54, pp. 744-749, 2009.

[32] J. Sun, G.P. Liu, and J. Chen, "Improved delay-range-dependent stability criteria for linear systems with time-varying delays," Automatica, vol. 46, pp. 466–470, 2010.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Stability of the Observer-Based Pole Placement for Discrete Time-Varying Non-Lexicographically-Fixed Systems

Yasuhiko Mutoh
Department of Engineering and
Applied Sciences
Sophia University
7-1 Kioicho, Chiyoda-ku, Tokyo, Japan
Email: y_mutou@sophia.ac.jp

Tomohiro Hara
Department of Engineering and
Applied Sciences
Sophia University
7-1 Kioicho, Chiyoda-ku, Tokyo, Japan

***Abstract*—This paper concerns the observer-based pole placement control for MIMO time varying non-lexicographically-fixed discrete systems. If both of the reachability indices and the observability indices are non-lexicographically-fixed, augmented plant equation and augmented observer are needed. Design procedure of this control system is proposed and the stability and the separation principle of the total closed loop system is also shown.**

*Keywords - Pole Placement; Observer; Time Varying System; Descrete System; Non-Lexicographically-Fixed System*

## I. Introduction

It is well known that the pole placement control can be designed for linear time-varying system by using the controllability canonical form as in the time-invariant case [5], [6]. The linear time-varying multivariable system whose controllability indices or observability indices are not constant is called the non-lexicographically-fixed system. Valasec et. al. [7] proposed the pole placement design method for such a system by augmenting the system equation so that the augmented system is lexicographically-fixed. This paper concerns the pole placement and the observer design method for linear time-varying discrete non-lexicographically-fixed system. Using the Valasec's idea, the procedure to extend a discrete non-lexicographically-fixed system to a lexicographically-fixed augmented system will be presented. Then, the simple pole placement technique can be applied to the augmented system without transforming the system into any canonical form [12]. Further, using the property of the anti-causal dual system, it will be shown that the same design method can be used for the augmented observer for non-lexicographically-fixed systems. Finally, as for the time-invariant case, the stability and the separation principle of the total closed loop system are also shown for the case where both of the augmented pole placement controller and the augmented observer are used.

## II. Preliminaries

Consider the following linear time-varying m-input p-output MIMO discrete system.

$$\begin{aligned} x(k+1) &= A(k)x(k) + B(k)u(k) \qquad (1) \\ y(k) &= C(k)x(k) \qquad (2) \end{aligned}$$

where $x \in R^n$, $u \in R^m$, and $y \in R^p$ are the state variable, the input and output, respectively. $A(k) \in R^{n\times n}$, $B(k) \in R^{n\times m}$ and $C(k) \in R^{p\times n}$ are time-varying coefficient matrices.

*Definition 1:* System (1) is called "completely reachable in $n$ steps" if for any $x_1 \in R^n$ there exists a bounded input $u(j)$ $(j = k, \cdots, k+n-1)$ such that $x(k) = 0$ and $x(k+n) = x_1$ for all k.

The reachability matrix, $R(k)$, of this system is defined by

$$\begin{aligned} R(k) &= \left[ \; b_0^1(k) \quad \cdots \quad b_0^m(k) \; \right| \cdots \\ & \qquad \cdots \left| \; b_{n-1}^1(k) \quad \cdots \quad b_{n-1}^m(k) \; \right] \qquad (3) \end{aligned}$$

Here, $b_i^l(k) \in R^n$ is calculated by the following recurrence equations.

$$\begin{aligned} b_0^l(k) &= b^l(k+n-1) \\ b_{i+1}^l(k) &= A(k+n-1)b_i^l(k-1) \qquad (4) \\ & \qquad (i = 0, \cdots, n-2, \quad l = 1, \cdots, m \;) \end{aligned}$$

where, $b^l(k) \in R^n$ is the $l$-th column of $B(k)$.

If the system (1) is completely reachable in $n$ steps, the rank of $R(k)$ is $n$, from which we can define the nonsingular $n \times n$ matrix $\bar{R}(k)$ using the reachability indices $\mu_i (i = 1, 2, \cdots, m)$ as follows.

$$\begin{aligned} \bar{R}(k) &= \left[ \; b_0^1(k), \cdots, b_{\mu_1-1}^1(k) \right| \cdots \\ & \qquad \cdots \left| b_0^m(k), \cdots, b_{\mu_m-1}^m(k) \; \right] \qquad (5) \end{aligned}$$

The reachability indices satisfy that $\sum_{i=1}^m \mu_i = n$ and are assumed that $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_m$ without loss of generallity.

Here, we state the definition of the observability in $n$ steps as a dual concept of the reachability in $n$ steps.

*Definition 2:* The system (1), (2) is said to be completely observable in $n$ steps if for any $k$, $x(k)$ is determined uniquely from $y(k)$, $y(k+1)$, $\cdots$, $y(k+n-1)$.

The following steps are the pole placement control design procedure proposed by the authors in [11] without using a transformation into any canonical form.

**STEP 1** Check the reachability of the system (1) and obtain $\bar{R}(k)$ and $\mu_i$ $(i=1,\cdots,m)$.

**STEP 2**

Calculate the new output signal, $\tilde{y}(k)$, by the following equation, so that the relative degree from $u(k)$ to $\tilde{y}(k)$ is the system degree, $n$.

$$\tilde{y}(k) = \tilde{C}(k)x(k) = W\bar{R}^{-1}(k-n)x(k) \tag{6}$$

where $W$ is the matrix defined by the following.

$$\begin{aligned} W &= diag(w_1, w_2, \cdots, w_m) \\ w_i &= \left[\begin{array}{cccc} 0 & \cdots & 0 & 1 \end{array}\right] \in R^{1\times\mu_i} \end{aligned} \tag{7}$$

**STEP 3** Let $q^i(z)$ be the ideal and stable characteristic polynomial for the closed-loop system of degree $\mu_i$, i.e.,

$$q^i(z) = z^{\mu_i} + \alpha^i_{\mu_i-1}z^{\mu_i-1} + \cdots + \alpha^i_1 z + \alpha^i_0 \tag{8}$$

Here, $z$ is the shift operator. Then, we have the following equation [11].

$$\left[\begin{array}{ccc} q_1(z) & & \\ & \ddots & \\ & & q_m(z) \end{array}\right]\tilde{y}(k) = F(k)x(k) + \Lambda(k)u(k) \tag{9}$$

where $\Lambda(k) \in R^{m\times m}$ is nonsingular. (See Appendix.)

**STEP 4** From (9), the state feedback

$$u(k) = D(k)x(k) = -\Lambda^{-1}(k)F(k)x(k) \tag{10}$$

makes the closed loop system

$$\left[\begin{array}{ccc} q_1(z) & & \\ & \ddots & \\ & & q_m(z) \end{array}\right]\tilde{y}(k) = 0 \tag{11}$$

This implies that the closed loop state equation

$$x(k+1) = \{A(k) + B(k)D(k)\}x(k) \tag{12}$$

is equivalent to the time invariant system with desired poles, i.e., there exists some transformation matrix, $P(k)$, that satisfies the following equation.

$$P(k+1)\{A(k) + B(k)D(k)\}P^{-1}(k) = A^* \tag{13}$$

where

$$\det(zI - A^*) = \prod_{i=1}^{m} q^i(z) \tag{14}$$

Then, if the matrix $P(k)$ is the Lyapunov transformation, the closed loop system is stable and equivalent to some time-invariant system that has the ideal and stable eigen values.

## III. Pole Placement of Non-Lexicographically-Fixed Systems

In the previous section, the reachability indices are supposed to be fixed. Such indices are said to be lexicographically-fixed. However, since the system has time-varying parameters, the reachability indices might be variable as well. Such indices are said to be non-lexicographically-fixed. In this section, we consider the pole placement control designing procedure for a system with non-lexicographically-fixed indices. Valasek et. al. proposed the pole placement design method for non-lexicographically-fixed multivariable continuous systems in [8]. In this paper, we apply this idea to the discrete system together with the new pole placement technique stated in the previous section.

Suppose that the system (1) is completely reachable in $n$ steps, and has non-lexicographically-fixed reachability indices. It is assumed that the maximum value of each reachability index $\mu_i$ is known, i.e.,

$$v_i = \max_k \mu_i(k) \qquad (i=1,\cdots,m) \tag{15}$$

Using $v_i$, we define $n_g$ by

$$n_g = \sum_{i=1}^{m} v_i \tag{16}$$

Define the augmented system by

$$x_g(k+1) = A_g(k)x_g(k) + B_g(k)u(k) \tag{17}$$

$$\left\{\begin{array}{l} x_g(k) = \left[\begin{array}{c} x(k) \\ x_e(k) \end{array}\right] \\ A_g(k) = \left[\begin{array}{cc} A(k) & 0 \\ A_2(k) & A_1(k) \end{array}\right], B_g(k) = \left[\begin{array}{c} B(k) \\ B_e(k) \end{array}\right] \end{array}\right.$$

where $x_g \in R^{n_g}$ and $x_e \in R^{n_g-n}$. $A_1(k) \in R^{(n_g-n)\times(n_g-n)}$, $A_2(k) \in R^{(n_g-n)\times n}$, and $B_e(k) \in R^{(n_g-n)\times m}$ are design parameter matrices so that the above augmented system has lexicographically-fixed reachability indices, $v_i$ $(i=1,\cdots,m)$.

The reachability matrix $R_g(k)$ of this augmented system is

$$\begin{aligned} R_g(k) = [\ & {b_g}^1_0(k) \ \cdots \ {b_g}^m_0(k) \ | \cdots \\ & \cdots \ | \ {b_g}^1_{n_g-1}(k) \ \cdots \ {b_g}^m_{n_g-1}(k) \ ] \end{aligned} \tag{18}$$

where ${b_g}^l_i(k) \in R^{n_g}$ is defined by the following recurrence equations.

$$\begin{aligned} {b_g}^l_0(k) &= {b_g}^l(k+n_g-1) \\ {b_g}^l_{i+1}(k) &= A_g(k+n_g-1){b_g}^l_i(k-1) \\ & \quad (i=0,\cdots,n_g-2, \quad l=1,\cdots,m\ ) \end{aligned} \tag{19}$$

Here, ${b_g}^l(k) \in R^{n_g}$ is the $l$-th column of $B_g(k)$.

For the augmented system to have lexicographically-fixed reachability indices, $v_i$, the following $n_g \times n_g$ matrix $\bar{R}_g(k)$ should be nonsingular for all $k$.

$$\begin{aligned} \bar{R}_g(k) \ = \ [\ & {b_g}^1_0(k), \cdots, {b_g}^1_{v_1-1}(k) | \cdots \\ & \cdots | {b_g}^m_0(k), \cdots, {b_g}^m_{v_m-1}(k)\ ] \end{aligned} \tag{20}$$

On the other hand, $\bar{R}_g(k)$ can be written as

$$\bar{R}_g(k) = \left[ \begin{array}{c} \bar{R}_v(k) \\ \bar{R}_e(k) \end{array} \right] \tag{21}$$

where $\bar{R}_v(k) \in R^{n \times n_g}$ and $\bar{R}_e(k) \in R^{(n_g - n) \times n_g}$ are

$$\begin{aligned} \bar{R}_v(k) &= [\ b_0^1(k), \cdots, b_{v_1-1}^1(k) | \cdots \\ &\qquad \cdots | b_0^m(k), \cdots, b_{v_m-1}^m \ ] \end{aligned} \tag{22}$$

$$\begin{aligned} \bar{R}_e(k) &= [\ {r_e}_0^1(k), \cdots, {r_e}_{v_1-1}^1(k) | \cdots \\ &\qquad \cdots | {r_e}_0^m(k), \cdots, {r_e}_{v_m-1}^m(k) \ ] \end{aligned} \tag{23}$$

Since, from the assumption, the rank of $\bar{R}_v(k)$ is $n$, there exists a matrix, $\bar{R}_e(k)$, such that $\bar{R}_g(k)$ is nonsingular for all $k$. The problem is to find $A_1(k)$, $A_2(k)$, and $B_e(k)$ that give such ${r_e}_i^l(k) \in R^{n_g - n}$.

From (20)-(23), we have

$$ {b_g}_0^l(k) = \left[ \begin{array}{c} b_0^l(k) \\ {r_e}_0^l(k) \end{array} \right] \tag{24}$$

then, using (17) and (24), the recurrence equation (19) can be modified as follows.

$$\begin{aligned} {b_g}_0^l(k) &= \left[ \begin{array}{c} b_0^l(k) \\ {r_e}_0^l(k) \end{array} \right] = \left[ \begin{array}{c} b^l(k + n_g - 1) \\ {b_e}^l(k + n_g - 1) \end{array} \right] \\ {b_g}_{i+1}^l(k) &= \left[ \begin{array}{c} b_{i+1}^l(k) \\ {r_e}_{i+1}^l(k) \end{array} \right] \\ &= \left[ \begin{array}{cc} A(k + n_g - 1) & 0 \\ A_2(k + n_g - 1) & A_1(k + n_g - 1) \end{array} \right] {b_g}_i^l(k - 1) \\ &= \left[ \begin{array}{c} A(k + n_g - 1) b_i^l(k - 1) \\ [\ A_2(k + n_g - 1) \quad A_1(k + n_g - 1)\ ] {b_g}_i^l(k - 1) \end{array} \right] \\ &\qquad (i = 0, 1, \cdots\ , \quad l = 1, \cdots, m\ ) \end{aligned} \tag{25}$$

Here, ${b_e}^l(k)$ is the $l$-th column of $B_e(k)$. From (25), the relation between ${r_e}_i^l(k)$ and $A_1(k)$, $A_2(k)$, and $B_e(k)$ is obtained as follows.

$$\begin{aligned} & B_e(k + n_g - 1) = [\ {r_e}_0^1(k) \quad \cdots, \quad {r_e}_0^m(k)\ ] \\ & [\ A_2(k + n_g - 1) \quad A_1(k + n_g - 1)\ ] \bar{R}_g(k - 1) \\ & \qquad = \bar{R}_{e+}(k) \end{aligned} \tag{26}$$

where $\bar{R}_{e+}(k)$ is defined by

$$\begin{aligned} \bar{R}_{e+}(k) = [\ & {r_e}_1^1(k) \quad \cdots \quad {r_e}_{v_1}^1(k) | \cdots \\ & \cdots \ | {r_e}_1^m(k) \quad \cdots \quad {r_e}_{v_m}^m(k)\ ] \end{aligned} \tag{27}$$

From the above, design parameter matrices such that the augmented plant (17) has lexicographically-fixed reachability indices, $v_i (i = 1, \cdots, m)$, can be calculated as follows. First, determine $\bar{R}_e(k)$ so that $\bar{R}_g(k)$ is nonsingular for all $k$. Then, using arbitrarily determined parameters ${r_e}_{v_1}^1(k), \cdots, {r_e}_{v_m}^m(k)$ in (26) and (27), and then, $A_1(k)$, $A_2(k)$ and $B_e(k)$ are obtained by

$$\begin{aligned} & B_e(k) = [\ {r_e}_0^1(k - n_g + 1) \quad \cdots, \quad {r_e}_0^m(k - n_g + 1)\ ] \\ & [\ A_2(k) \quad A_1(k)\ ] = \bar{R}_{e+}(k - n_g + 1) \bar{R}_g^{-1}(k - n_g) \end{aligned} \tag{28}$$

The state feedback for the pole placement can be obtained as the following form by applying the pole placement design procedure stated in the previous section to this augmented system.

$$u(k) = [D_x(k),\ D_e(k)] \left[ \begin{array}{c} x(k) \\ x_e(k) \end{array} \right] = D_g(k) x_g(k) \tag{29}$$

This implies that there exists the time-varying transformation matrix $P_g(k) \in R^{n_g \times n_g}$ that satisfies

$$P_g(k + 1) \{A_g(k) + B_g(k) D_g(k)\} P_g^{-1}(k) = A_g^* \tag{30}$$

Hence, if the transformation matrix $P_g(k)$ is the Lyapunov transformation, the closed loop system is stable and equivalent to some time-invariant system that has desired and stable constant eigenvalues.

## IV. Observer of Non-Lexicographically-Fixed Systems

In this section, we consider the design of the observer for the system that has non-lexicographically-fixed observability indices. Suppose that the system (1),(2) is completely observable in $n$ steps and has observability indices, $\nu_i (i = 1, \cdots, p)$, which are non-lexicographically-fixed. Further, it is assumed that the following $d_i$ are known.

$$d_i = \max_k \nu_i(k) \qquad (i = 1, \cdots, p) \tag{31}$$

Using these $d_i$, we define $n_s$ by

$$n_s = \sum_{i=1}^{m} d_i \tag{32}$$

If the system has lexicographically-fixed observability indices, its observer can be written as follows.

$$\begin{aligned} \hat{x}(k + 1) &= A(k) \hat{x}(k) + B(k) u(k) \\ &\quad - H(k)(y(k) - C(k) \hat{x}(k)) \end{aligned} \tag{33}$$

where $\hat{x}(k) \in R^n$ is the state estimation of $x(t)$. Then, the problem is to find the observer gain matrix $H(k) \in R^{n \times p}$. But, since the observability indices are non-lexicographically-fixed, we augment the observer system as follows.

$$\begin{aligned} \hat{x}(k + 1) &= A(k) \hat{x}(k) + B(k) u(k) \\ &\quad - H(k)(y(k) - C(k) \hat{x}(k)) \\ &\quad - (A_4(k) + H(k) C_e(k)) \epsilon(k) \end{aligned} \tag{34}$$

$$\begin{aligned} \epsilon(k + 1) &= A_3(k) \epsilon(k) + H_e(k)(y(k) - C(k) \hat{x}(k)) \\ &\quad + H_e(k) C_e(k) \epsilon(k) \end{aligned} \tag{35}$$

Here, $\epsilon(k) \in R^{n_s - n}$ is an auxiliary signal and $A_3(k) \in R^{(n_s - n) \times (n_s - n)}$, $A_4(k) \in R^{n \times (n_s - n)}$, and $C_e(k) \in R^{p \times (n_s - n)}$ are design parameter matrices determined later. Using the state estimation error, $e(k) = x(k) - \hat{x}(k)$ the following state error equation is obtained from (1), (2), (34), and (35).

$$e_s(k + 1) = A_s(k) e_s(k) + H_s(k) C_s(k) e_s(k) \tag{36}$$

where

$$\left\{\begin{array}{l} e_s(k) = \left[\begin{array}{c} e(k) \\ \epsilon(k) \end{array}\right], A_s(k) = \left[\begin{array}{cc} A(k) & A_4(k) \\ 0 & A_3(k) \end{array}\right] \\ H_s(k) = \left[\begin{array}{c} H(k) \\ H_e(k) \end{array}\right], C_s(k) = \left[\begin{array}{cc} C(k) & C_e(k) \end{array}\right] \end{array}\right. \tag{37}$$

From this, the observer design problem is to find $H_s(k)$ so that $A_s(k) + H_s(k)C_s(k)$ is equivalent to some constant matrix which has desired constant eigenvalues.

For this purpose, consider the following anti-causal system as a dual system of the system $(A_s(k),\ C_s(k))$.

$$\begin{array}{l} \xi_s(k-1) = A_s^T(k)\xi_s(k) + C_s^T(k)v(k) \\ A_s^T = \left[\begin{array}{cc} A^T(k) & 0 \\ A_4^T(k) & A_3^T(k) \end{array}\right],\ C_s^T(k) = \left[\begin{array}{c} C^T(k) \\ C_e^T(k) \end{array}\right] \end{array} \tag{38}$$

Since the system $(A(k)$,$C(k))$ is completely observable in $n$ steps and has the observability indices, $\nu_i (i = 1, \cdots, p)$, its dual system is completely reachable in $n$ steps and has reachability indices, $\nu_i$.

The reachability matrix, $R_s(k)$, of the augmented dual system (38) can be written as

$$\begin{array}{r} R_s(k) = \left[\ c_s{}_0^1(k) \quad \cdots \quad c_s{}_0^m(k)\ \middle|\ \cdots \right. \\ \left. \cdots\ \middle|\ c_s{}_{n_s-1}^1(k) \quad \cdots \quad c_s{}_{n_s-1}^m(k)\ \right] \end{array} \tag{39}$$

where $c_s{}_i^l(k) \in R^{n_s}$ is defined by the following recurrence equations.

$$\begin{array}{rcl} c_s{}_0^l(k) & = & c_s{}^l(k - n_s + 1) \\ c_s{}_{i+1}^l(k) & = & A_s^T(k - n_s + 1)c_s{}_i^l(k+1) \\ & & (i = 0, \cdots, n_s - 2, \quad l = 1, \cdots, m\ ) \end{array} \tag{40}$$

Here, $c_s{}^l(k) \in R^{n_s}$ is the $l$-th column of $C_s^T(k)$.

For the augmented system to have lexicographically-fixed reachability indices, $d_i$, the following $n_s \times n_s$ matrix, $\bar{R}_s(k)$, should be nonsingular for all $k$.

$$\begin{array}{rcl} \bar{R}_s(k) & = & [\ c_s{}_0^1(k), \cdots, c_s{}_{d_1-1}^1(k)| \cdots \\ & & \cdots |c_s{}_0^m(k), \cdots, c_s{}_{d_m-1}^m(k)\ ] \end{array} \tag{41}$$

$\bar{R}_s(k)$ can be written as

$$\bar{R}_s(k) = \left[\begin{array}{c} \bar{R}_d(k) \\ \bar{R}_h(k) \end{array}\right] \tag{42}$$

where, $\bar{R}_d(k) \in R^{n \times n_s}$ and $\bar{R}_h(k) \in R^{(n_s - n) \times n_s}$ are defined by

$$\begin{array}{rcl} \bar{R}_d(k) & = & [\ c_0^1(k), \cdots, c_{v_1-1}^1(k)| \cdots \\ & & \cdots |c_0^m(k), \cdots, c_{d_m-1}^m\ ] \end{array} \tag{43}$$

$$\begin{array}{rcl} \bar{R}_h(k) & = & [\ r_h{}_0^1(k), \cdots, r_h{}_{d_1-1}^1(k)| \cdots \\ & & \cdots |r_h{}_0^m(k), \cdots, r_h{}_{d_m-1}^m(k)\ ] \end{array} \tag{44}$$

Since, the anti-causal dual system is reachable in $n$ steps, the rank of $\bar{R}_d(k)$ is $n$, and, hence, there always exists the matrix, $\bar{R}_h(k)$, such that the rank of $\bar{R}_s(k)$ is $n_s$ for all $k$. Thus, as the previous section, $C_e^T(k)$, $A_3^T(k)$, and $A_4^T(k)$ can be obtained by

$$\begin{array}{l} C_e^T(k) = [\ r_h{}_1^0(k + n_s - 1) \quad \cdots \quad r_h{}_m^0(k + n_s - 1)\ ] \\ [\ A_4^T(k) \quad A_3^T(k)\ ] = \bar{R}_{h+}(k + n_s - 1)\bar{R}_s^{-1}(k + n_s). \end{array} \tag{45}$$

Here, $\bar{R}_{h+}(k)$ is defined by

$$\begin{array}{r} \bar{R}_{h+}(k) = [\ r_h{}_1^1(k) \quad \cdots \quad r_h{}_{d_1}^1(k)| \cdots \\ \cdots |r_h{}_1^m(k) \quad \cdots \quad r_h{}_{d_m}^m(k)\ ] \end{array} \tag{46}$$

where $r_h{}_{d_1}^1(k)$,$\cdots$, $r_h{}_{d_m}^m(k)$ are arbitrarily determined parameters.

From the above, the anti-causal augmented dual system, (38), has lexicographically-fixed reachability indices, $d_i$. Thus, using the pole placement technique stated in the section 2, the matrix $H_s^T(k)$ can be obtained so that $A_s^T(k) + C_s^T(k)H_s^T(k)$ is equivalent to some constant matrix $A_o^{*T}$ which has desired constant eigenvalues. i.e., there exists some transformation matrix, $Q_s(k) \in R^{n_s \times n_s}$, such that

$$Q_s(k+1)\{A_s(k) + C_s(k)H_s(k)\}Q_s^{-1}(k) = A_o^* \tag{47}$$

Hence, if $Q_s(k)$ is the Lyapunov transformation, (34) and (35) becomes the augmented observer, and $e(k)$ and $\epsilon(k)$ converge to 0.

## V. Stability of the Total Closed Loop

If the system (1), (2) has both of non-lexicographically-fixed reachability indices and non-lexicographically-fixed observability indices, the augmented plant and the augmented observer are needed for the observer based pole placement. In this section, for such a system, the stability of the total closed loop system and the separation principle are considered.

The augmented plant is (17), and the augmented observer is (34), (35). Then, for the observer based pole placement, the state feedback (29) is modified to

$$u(k) = \left[\begin{array}{cc} D_x(k) & D_e(k) \end{array}\right] \left[\begin{array}{c} \hat{x}(k) \\ \hat{x}_e(k) \end{array}\right] \tag{48}$$

where $\hat{x}(k)$ is the state estimation. In this state feedback, $\hat{x}_e(k)$ is used instead of $x_e(k)$, because, in the second equation of (17), $x(t)$ should be replaced by $\hat{x}(k)$.

Hence, the total closed loop system for this case becomes as follows.

$$\begin{array}{l} \left[\begin{array}{c} x(k+1) \\ \hat{x}_e(k+1) \\ \hat{x}(k+1) \\ \epsilon(k+1) \end{array}\right] = \left[\begin{array}{cc} A & BD_e \\ 0 & A_1 + B_eD_e \\ -HC & BD_e \\ H_eC & 0 \end{array}\right. \\ \left.\begin{array}{cc} BD_x & 0 \\ A_2 + B_eD_x & 0 \\ A + BD_x + HC & -A_4 - HC_e \\ -H_eC & A_3 + H_eC_e \end{array}\right] \left[\begin{array}{c} x(k) \\ \hat{x}_e(k) \\ \hat{x}(k) \\ \epsilon(k) \end{array}\right] \end{array} \tag{49}$$

Using the transformation matrix

$$T = \left[\begin{array}{cc|cc} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ \hline I & 0 & -I & 0 \\ 0 & 0 & 0 & I \end{array}\right] \qquad (50)$$

the total system (49) is transformed into

$$\begin{bmatrix} x(k+1) \\ \hat{x}_e(k+1) \\ e(k+1) \\ \epsilon(k+1) \end{bmatrix} = \begin{bmatrix} A+BD_x & BD_e & -BD_x & 0 \\ A_2+B_eD_x & A_1+B_eD_e & -A_2-B_eD_x & 0 \\ 0 & 0 & A+HC & A_4+HC_e \\ 0 & 0 & H_eC & A_3+H_eC_e \end{bmatrix} \begin{bmatrix} x(k) \\ \hat{x}_e(k) \\ e(k) \\ \epsilon(k) \end{bmatrix}$$

$$= \begin{bmatrix} A_g+B_gD_g & E \\ 0 & A_s+C_sH_s \end{bmatrix} \begin{bmatrix} x(k) \\ \hat{x}_e(k) \\ e(k) \\ \epsilon(k) \end{bmatrix} \qquad (51)$$

where

$$E(k) = \begin{bmatrix} -B(k)D_x(k) & 0 \\ -A_2(k)+B_e(k)D_x(k) & 0 \end{bmatrix}. \qquad (52)$$

In (49) and (51), the symbol "(k)" is omitted because of the small space.

From the above, using the transformation matrix

$$\Phi(k) = \begin{bmatrix} P_g(k) & 0 \\ 0 & Q_s(k) \end{bmatrix} \qquad (53)$$

the following relation is obtained.

$$\Phi(k+1)\begin{bmatrix} A_g(k) & E(k) \\ 0 & A_s(k) \end{bmatrix}\Phi^{-1}(k) = \begin{bmatrix} A^* & P_g(k+1)E(k)Q_s^{-1}(k) \\ 0 & A_o^* \end{bmatrix} \qquad (54)$$

Thus, since the system matrix of (49) is equivalent to the right hand side of (54), if. $P_g(k)$ and $Q_s(k)$ are the Lyapunov transformation matrices, the total closed system is stable and has a property of the separation principle.

## VI. Numerical Example

Consider the system (1), (2) with

$$A(k) = \begin{bmatrix} 2\cos(1.5k) & 0 & 0 \\ 2\sin(1.5(k-1)) & 0 & -2 \\ 2\sin(1.5k) & 2\cos(1.5k) & 0 \end{bmatrix} \qquad (55)$$

$$B(k) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \sin(1.5k) & 0 \end{bmatrix} \qquad (56)$$

$$C(k) = \begin{bmatrix} \frac{1}{2}\cos(1.5(k-1)) & 0 & \frac{1}{2}\cos(1.5(k-1)) \\ 0 & \frac{1}{2} & 0 \end{bmatrix} \qquad (57)$$

This system has non-lexicographically-fixed reachability indices and non-lexicographically-fixed observability indices. We design the observer based pole placement for this system. Because of the small space, we use the following symbols, i.e., $S = \sin(1.5k), S_1 = \sin(1.5(k-1)), C = \cos(1.5k), C_1 = \cos(1.5(k-1))$

The reachability indices of this system is $\mu_1 = 2$, $\mu_2 = 1$ or $\mu_1 = 1$, $\mu_2 = 2$. From this $v_1 = v_2 = 2$. The Augmented plant equation (17) becomes

$$x_g(k+1) = A_g(k)x_g(k) + B_g(k)u(k) \qquad (58)$$

$$\begin{cases} x_g(k) = \begin{bmatrix} x(k) \\ x_e(k) \end{bmatrix} \\ A_g(k) = \begin{bmatrix} A(k) & 0 \\ A_2(k) & A_1(k) \end{bmatrix}, B_g(k) = \begin{bmatrix} B(k) \\ B_e(k) \end{bmatrix} \end{cases}$$

where $x_g(k) \in R^4$ and $x_e(k) \in R^1$, and

$$B_e(k) = \begin{bmatrix} -1 & 0 \end{bmatrix} \qquad (59)$$

$$\begin{bmatrix} A_2(k) & A_1(k) \end{bmatrix} = \begin{bmatrix} 2S^2S_1^2(C_1+1) & -2S & -2C_1S^2S_1 & -2C_1^2S^2 \end{bmatrix}. \qquad (60)$$

On the other hand, the observability indices of this system is also $\nu_1 = 2$, $\nu_2 = 1$ or $\nu_1 = 1$, $\nu_2 = 2$. Then, the augmented observer becomes

$$\begin{aligned} \hat{x}(k+1) &= A(k)\hat{x}(k) + B(k)u(k) \\ &\quad -H(k)(y(k) - C(k)\hat{x}(k)) \\ &\quad -(A_4(k) + H(k)C_e(k))\epsilon(k) \qquad (61) \\ \epsilon(k+1) &= A_3(k)\epsilon(k) + H_e(k)(y(k) - C(k)\hat{x}(k)) \\ &\quad +H_e(k)C_e(k)\epsilon(k). \qquad (62) \end{aligned}$$

Here, $\epsilon(k) \in R^1$ and

$$\begin{aligned} C_e^T(k) &= \begin{bmatrix} 2 & 0 \end{bmatrix} \qquad (63) \\ \begin{bmatrix} A_4^T(k) & A_3^T(k) \end{bmatrix} &= \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}. \qquad (64) \end{aligned}$$

Using the following desired stable characteristic polynomial for both of the pole placement and the observer

$$\begin{aligned} q^1(z) &= \alpha_2^1z^2 + \alpha_1^1z + \alpha_0^1 = z^2 + 0.4z - 0.05 \qquad (65) \\ q^2(z) &= \alpha_2^2z^2 + \alpha_1^2z + \alpha_0^2 = z^2 + 0.4z - 0.05 \qquad (66) \end{aligned}$$

the simulation results are shown in Fig.1 and Fig.2. Fig.1 shows the response of the augmented system, $[x(k), \hat{x}_e(k)] = [x_1(k), x_2(k), x_3(k), \hat{x}_e(k)]$, and Fig.2 shows the response of augmented state estimation error, $[e(k), \epsilon(k)] = [e_1(k), e_2(k), e_3(k), \epsilon(k)]$.

## VII. Conclutions

In this paper, the design procedure of the observer-based pole placement for linear time-varying MIMO systems is proposed. Especially, the system is supposed to have the non-lexicographically-fixed reachability indices and non-lexicographically-fixed observability indices. The total closed loop stability and the separation principle are also established.

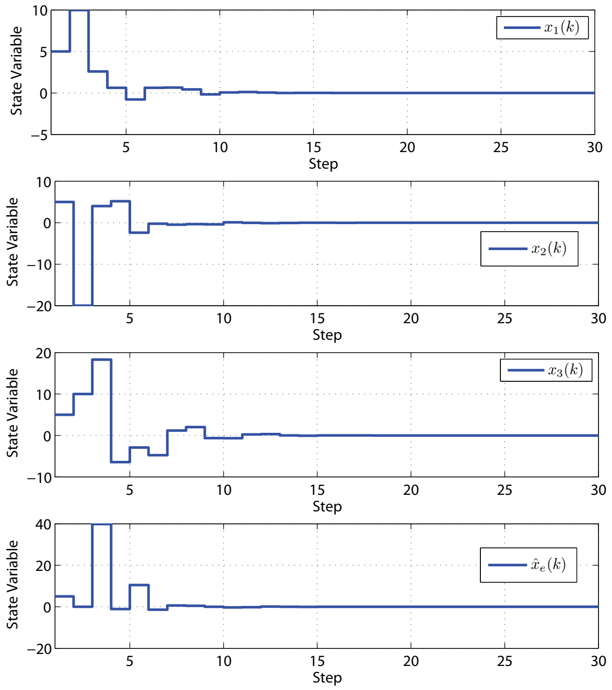


Fig. 1. Response of the Augmented Plant of the Observer-Based Pole Placement Control for the non-lexicographically-fixed System

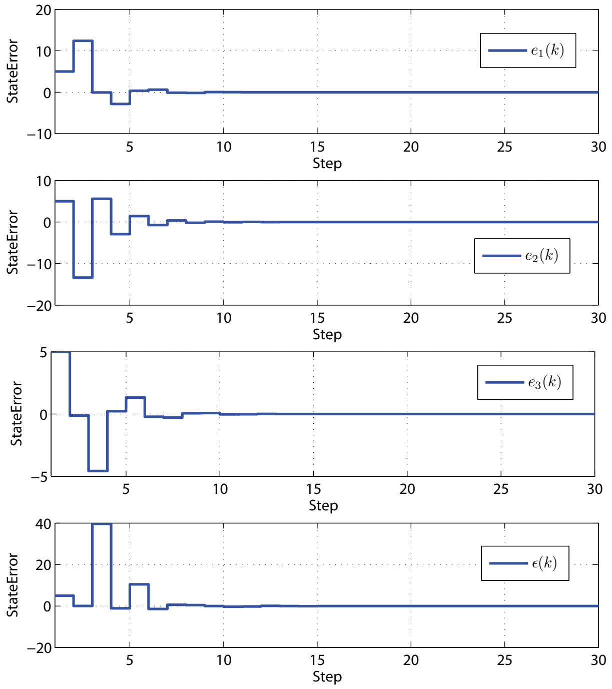


Fig. 2. Response of the Augmented State Estimation Error of the Observer-Based Pole Placement Control for the non-lexicographically-fixed System

## References

[1] W.J.Rugh: Linear System Theory 2nd Edition, prentice hall (1995)

[2] T.Kailath: Linear Systems, prentice hall (1980)

[3] C.Chen: Linear System Theory and Design, Oxford University press (1999)

[4] L.Weiss: ”Controllability, Realization and Stability of Discrete-time Systems”, SIAM J.Control, Vol.10, No.2, pp.231-251 (1972)

[5] C.Nguyen: ”Arbitrary eigenvalue assignments for linear time-varying multivariable control systems”, Internal Journal of Control, Vol.45, No.3, pp.1051-1057 (1987)

[6] M.Valasek and N.Olgac: ”Efficient pole placement technique for linear time-variant SISO systems”, IEE Proc Control Theory Appl, Vol.142, No.5, pp.451-458 (1995)

[7] M.Valasek and N.Olgac: ”Pole placement for linear time-varying non-lexicographically fixed MIMO systems”, Automatica, 35, pp.101-108 (1999)

[8] Richard T. O’Brien, Jr and Pable, A. Iglesias, *On the Poles and Zeros of Linear, Time-Varying Systems.* IEEE Trans Circuit Systemd I Fund Theory and Applications, **48**-5, 565-577, (2001)

[9] H.C.Lee and J.W.Choi: ”Ackermann-like eigenvalue assignment formulae for linear time-varying systems”, IEE Proc Control Theory Appl, Vol.152, No.4, pp.427-434 (2005)

[10] M.G.Matthews and C.Nguyen: ”A New Approach To Control Of Time-Varying Robotics Systems”, Computers and their applications: Proceedings of the ISCA international conference, Long Beach, California U.S.A., March 17-19 (1994)

[11] W.Chai ,N.K.Loh and H.Hu: ”Observer Design for Time-Varying Systems” ,Internal Journal of Systems Science, Vol.22, No.7, pp.1177-1196 (1991)

[12] Y.Mutoh and T.Hara: ”A Simple Design Method of Pole Placement for Linear Time-Varying Discrete Multivariable Systems”, Proceedings of the 30th Chinese Control Conference, Yantai, China, Jul 22-24 (2011)

## Appendix

Let $\tilde{c}_i^T(k)$ be the $i$-th row of $\tilde{C}(k)$. Then, we define $\tilde{c}_i^{lT}(k)$ by the following for $i = 1, \cdots, m$.

$$\begin{aligned} \tilde{c}_i^{0T}(k) &= \tilde{c}_i^T(k) \\ \tilde{c}_i^{(l+1)T}(k) &= \tilde{c}_i^{lT}(k+1)A(k) \end{aligned}$$

Using this, $F(k)$ is calculated by

$$F(k) = \begin{bmatrix} F_1^T(k) \\ \vdots \\ F_m^T(k) \end{bmatrix}$$

where

$$F_i^T(k) = [\alpha_0^i, \alpha_1^i, \cdots, \alpha_{\mu_i-1}^i, 1] \begin{bmatrix} \tilde{c}_i^{0T}(k) \\ \tilde{c}_i^{1T}(k) \\ \vdots \\ \tilde{c}_i^{\mu_i T}(k) \end{bmatrix}.$$

$\Lambda(k)$ is calculated as

$$\Lambda(k) = \begin{bmatrix} 1 & \gamma_{12}(k) & \gamma_{13}(k) & \cdots & \gamma_{1m}(k) \\ 0 & 1 & \gamma_{23}(k) & \cdots & \gamma_{2m}(k) \\ 0 & 0 & 1 & \cdots & \gamma_{3m}(k) \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & 1 \end{bmatrix}$$

where $\gamma_{ij} = \tilde{c}_i^{(\mu_i-1)T} b^j(k) \quad (j \geq i+1)$.

# Periodic Disturbance Rejection of a Class of Nonlinear System

Xiafei Tang and Zhengtao Ding
Control Systems Centre
School of Electrical and
Electronic Engineering
The University of Manchester
Manchester, M13 9PL, UK
Email: zhengtao.ding@manchester.ac.uk

***Abstract*—This study presents a periodic disturbance rejection method for a class of nonlinear systems with the input weighting vector in the proportional nonlinear form. Especially, the periodic disturbance does not match with the system input. A neural network approximator is employed for the estimation of the ideal feedforward control input that tackles the influence brought by disturbances in closed-loop system. Moreover, The adaptive control techniques are applied to deal with nonlinear uncertainties and unknown parameters in the system. The proposed control design ensures the closed-loop convergence of the system, i.e. all states converge to a small set around their equilibrium points. A simulation example is included to support this control approach.**

## I. Introduction

Disturbance rejection of nonlinear systems is a widely discussed issue in the recent decade. Many existing disturbance rejection results are based on the internal model principle. An internal model is constructed to generate a desired feedforward input for the annihilation of disturbance. For the disturbance signal generated by an unknown linear exosystem, the global asymptotic stability of a disturbed nonlinear system is achieved via state feedback with a state observer and an internal model [1]. With an additional filter design, this method is extended to a nonlinear system whose nonlinear uncertainties and disturbance uncertainties are tackled concurrently [2]. Generally, not all disturbance signals are linear in practice. From this point of view, periodic disturbances suppression is investigated. An adaptive feedforward disturbance cancellation scheme is proposed for the rejection of sinusoidal disturbances [3] [4]. In particular, periodic disturbances generated by a nonlinear exosystem are considered. The asymptotic tracking is achieved for an output feedback nonlinear system via incorporating filtered transformation, high gain control and saturation for internal model design [5]. As a special case, the periodic disturbance rejection problem is examined. The nonlinear system disturbed by general periodic disturbances which are half-period alternative is asymptotically suppressed. With known wave profiles of disturbances, an estimate of the feedforward control input is constructed by bringing in new operations of a half-period integration and a delay operator [6]. This method is extended to a nonlinear system with unmatched disturbances in [7].

The aforemetioned publications focus on the reconstruction of feedforward control input via internal model design. In fact, the output of the exosystem is constrained by its structures, i.e. not all periodic disturbances can be generated by the exosystem. In terms of a general periodic disturbance without dynamic limits, new approximation methods are worth being investigated. As a hot topic, Neural Network (NN) is exploited in many control design applications. It is well-known that NN can be used to estimate any nonlinear signals in a compact set. The theory about adaptive NN approximator for the periodically disturbed nonlinear function is established. With different selections of basis functions, two types of three-layer NN approximator are included as FSE-MNN-based approximator and FSE-RBFNN-based approximator [8]. After that, A two-layer NN approximator in [9] is introduced for output feedback stochastic nonlinear stabilisation. Motivated by these NN approximator designs, the two-layer approximator is used to estimate the feedforward control input in [10].

Note that the aforementioned nonlinear system has a constant input vector. Recently, disturbance suppression is established for various types of nonlinear systems with time varying weighting vector. Robust adaptive techniques are used in output feedback control design for a disturbed nonlinear system. It is notified that the system is in output feedback form with time varying input and disturbance vectors. The terminology of flat zone in a neighbourhood of the origin is introduced such that output tracking error converges to the flat zone under that control. It is also specified that, in this case, the element in input vector is coupled with a nonlinear function of output [11]. A novel asymptotic disturbance rejection approach for unknown sinusoidal disturbances is designed for nonlinear systems in output feedback form with an input gain vector whose element is a function of the output [12].

In this paper, a new disturbance rejection method is introduced for a nonlinear system in output feedback form

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

with an input vector with elements being a production of a constant and a nonlinear function of output. This research is still based on the existence of an invariant manifold which has a zero output under an ideal feedforward control input. To extract the zero dynamics of the system, a filter based transformation is proposed as similar as in [5]. As matter of fact that an NN can be used to estimate any nonlinear signals in a prescribed compact set. Since the desired feedforward control input is regarded as an unknown function of bounded disturbance signal, an NN approximator is applied for its estimation. In addition, a new unknown parameter is introduced to compensate the closed-loop influence from nonlinear uncertainties. Adaptive techniques are then applied to estimate unknown parameters online. Finally, backstepping control design based on the estimate of filters is carried out. Lyapunov stability analysis is presented to confirm the closed-loop stability. All states converge to a small area around their equilibrium points. Compared with the existing results, the information of disturbance are not necessarily needed. Furthermore, the disturbance can be any type of periodic signals.

The structure of this paper is organised as follows. The problem description is presented in section II. Section III introduces a filtered transformation and extracts system zero dynamics. The NN approximator design is addressed in section IV. Adaptive backstepping control design procedures are presented in section V. Lyapunov stability analysis is then followed to ensure the convergence of the closed-loop system in section VI. A simulation example is included in section VII to demonstrate a standard control approach. Finally, section VIII concludes this paper.

## II. Problem Statement

Consider SISO nonlinear systems which can be transformed to the output feedback form as

$$\left\{\begin{array}{lcl} \dot{x} & = & Ax+\phi(y)+b\sigma(y)u+dw \\ y & = & C^T x \end{array}\right. \tag{1}$$

with

$$A=\begin{bmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 0 \end{bmatrix}, b=\begin{bmatrix} 0 \\ \vdots \\ b_\rho \\ \vdots \\ b_n \end{bmatrix}, C=\begin{bmatrix} 1 \\ \vdots \\ 0 \end{bmatrix},$$

and $d=[d_1,\ldots,d_n]^T$ where state $x=[x_1,\ldots,x_n]^T$, output $y\in\mathbb{R}$, input $u\in\mathbb{R}$, smooth nonlinear function $\phi(y)=[\phi_1(y),\ldots,\phi_n(y)]^T$ with $\phi(0)=0$, smooth function $\sigma(y):\mathbb{R}\to\mathbb{R}$ with $\sigma(y)\neq 0$, $w\in\mathbb{R}$ is a bounded periodic disturbance signal, $b_i\neq 0$ for $i=\rho,\ldots,n$, i.e. this system is a relative degree $\rho$ system.

The assumption below is necessary for this control design.

***Assumption*** *2.1:* The system is minimum phase, that is, the zeros of polynomial $\mathscr{B}(s)=\sum_{i=\rho}^{n} b_i s^{n-i}$ have negative real parts.

The design objective of disturbance rejection problem is described as: Find out a finite dimensional system

$$\left\{\begin{array}{lcl} \dot{\mu} & = & \nu(\mu,y,u),\mu\in\mathbb{R} \\ u & = & \zeta(\mu,y) \end{array}\right. \tag{2}$$

The closed-loop system is then stable under this controller.

## III. Filtered State Transformation

For system (1) with relative degree $\rho$, a filter is introduced as

$$\left\{\begin{array}{lcl} \dot{\xi}_1 & = & -\lambda_1\xi_1+\xi_2 \\ & \vdots & \\ \dot{\xi}_{\rho-1} & = & -\lambda_{\rho-1}\xi_{\rho-1}+\sigma(y)u \end{array}\right. \tag{3}$$

where $\lambda_i>0$ for $i=1,\ldots,\rho-1$ are the designed parameters. The filtered transformation is implemented by introducing

$$\bar{z}=x-[\bar{f}_1,\ldots,\bar{f}_{\rho-1}]\xi \tag{4}$$

where $\xi\in\mathbb{R}^{\rho-1}$, $\bar{f}_i\in\mathbb{R}^n$ for $i=1,\ldots,\rho-1$. The value of $\bar{f}_i$ is given recursively by

$$\left\{\begin{array}{lcl} \bar{f}_1 & = & b \\ \bar{f}_i & = & [A+\lambda_{i+1}I]\bar{f}_{i+1} \quad \text{for} \quad i=2,\ldots,\rho-1 \end{array}\right. \tag{5}$$

with positive designed $\lambda_\rho$. Then, the system (1) is transformed to

$$\left\{\begin{array}{lcl} \dot{\bar{z}} & = & A\bar{z}+\phi(y)+dw+f\xi_1 \\ y & = & C^T\bar{z} \end{array}\right. \tag{6}$$

where $f=[f_1,\ldots,f_n]^T=[A+\lambda_1 I]\bar{f}_1$. It is noted that $f_1=b_\rho$. Further, it is observed

$$\mathscr{D}(s):=\sum_{i=1}^{n} f_i s^{n-i}=\mathscr{B}(s)\prod_{i=1}^{\rho+1}(s+\lambda_i) \tag{7}$$

With Assumption 2.1, it implies that all zeros of $\mathscr{D}(s)$ are located on the left half plane. $\xi_1$ is considered as the input of system (7), i.e. this system has relative degree 1. For the convenience of analysing zero dynamics, another state $\tilde{z}\in\mathbb{R}^{n-1}$ is brought in as

$$z=\bar{z}_{2:n}-\frac{f_{2:n}}{f_1}y \tag{8}$$

$(\star)_{2:n}$ stands for a new vector or matrix which is constructed by the 2nd to $n$th raw of $\star$.

The system dynamics with new coordinates is given by

$$\left\{\begin{array}{lcl} \dot{z} & = & Dz+\psi_z(y)+d_z w \\ \dot{y} & = & z_1+\psi_y(y)+d_1 w+b_\rho\xi_1 \end{array}\right. \tag{9}$$

with

$$D = \begin{bmatrix} -\frac{f_2}{f_1} & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{f_{n-1}}{f_1} & 0 & \dots & 1 \\ -\frac{f_n}{f_1} & 0 & \dots & 0 \end{bmatrix} \tag{10}$$

where $D$ is the left companion matrix of $f$ and

$$\psi_z(y) = D\frac{f_{2:n}}{f_1}y + \phi_{2:n}(y) - \frac{f_{2:n}}{f_1}\phi_1(y) \tag{11}$$

$$\psi_y(y) = \frac{f_2}{f_1}y + \phi_1(y) \tag{12}$$

$$d_w = d_{2:n} - \frac{f_{2:n}}{b_\rho}d_1 \tag{13}$$

$D$ is Hurwitz from equation (7). Thus, there exist positive matrix $P_z$ and $Q_z$ which satisfy

$$D^T P_z + P_z D = -Q_z \tag{14}$$

With the property of $\phi(y)$, It renders $\psi_z(0) = 0$ and $\psi_y(0) = 0$.

From Isidori's output regulation theory [13], the existence of a controlled invariant manifold is a necessary condition of establishing a solution to nonlinear system output regulation problem. From this point of view, the following assumption is necessary.

***Assumption** 3.1:* There exist an invariant manifold $\pi(w) : \mathbb{R} \to \mathbb{R}^{n-1}$ and a forwarding control input $\mu(w) : \mathbb{R} \to \mathbb{R}$ such that

$$\left\{ \begin{array}{rcl} \dot{\pi} & = & B\pi + d_z w \\ 0 & = & \pi_1 + d_1 w + b_\rho \mu \end{array} \right. \tag{15}$$

Let $\tilde{z}$ denotes the error between $z$ and $\pi$. The final model for control design is obtained as

$$\left\{ \begin{array}{rcl} \dot{\tilde{z}} & = & D\tilde{z} + \psi_z(y) \\ \dot{y} & = & \tilde{z}_1 + \psi_y(y) + b_\rho\left(\xi_1 - \mu\right) \\ \dot{\xi}_1 & = & -\lambda_1\xi_1 + \xi_2 \\ & \vdots & \\ \dot{\xi}_{\rho-1} & = & -\lambda_{\rho-1}\xi_{\rho-1} + \sigma(y)u \end{array} \right. \tag{16}$$

## IV. Neural Networks Disturbance Approximator

In the section, a Neural Networks approximator based on the output $y$ is introduced to approximate the desired input $\mu(\omega)$ on a compact set $\Omega$ [9].

$$\mu = W^T S(y) + \delta \tag{17}$$

where $S(y) = [s_1(y), \dots, s_l(y)]^T : \Omega \to \mathbb{R}^l$ is a known smooth vector function with neural network node number $l > 1$. The basis function is given by Gaussian function as

$$s_i(y) = \exp\left[-\frac{(y-a_i)^2}{h^2}\right], \text{for } i = 1, \dots, l \tag{18}$$

with the centre $a_i \in \Omega$ and the width $h > 0$. The desired weighting vector $W = [W_1, \dots, W_l]^T$ is defined as

$$W := \arg\min_{\hat{W}\in\mathbb{R}^l}\left\{\sup_{y\in\Omega}|\mu - \hat{W}^T S(y)|\right\}$$

$\hat{W}$ is the estimate of $W$, $\delta$ is the NN inherent approximation error satisfying $|\delta| \le \bar{\delta}$, which is the minimum upper bound of $\delta$. It can be decreased by increasing the number of $r$ and $l$. The approximation of $\mu$ can be written as

$$\hat{\mu} = \hat{W}^T S(y). \tag{19}$$

## V. Adaptive backstepping control design

In this control design, some designed parameters are introduced. They satisfy

$$\begin{array}{rcl} c_0 & > & \frac{1}{2} + \frac{b_\rho^2}{4} \\ c_1 & > & -1 \\ c_i & > & 0 \quad \text{for} \quad i = 2, \dots, \rho - 1 \\ h_i & > & 0 \end{array}$$

Define the desired value of $\xi_i$ as $\hat{\xi}_i$ for $i = 1, \dots, \rho - 2$ with error $\tilde{\xi}_i = \xi_i - \hat{\xi}_i$. Therefore, from system (9), we have

$$\dot{y} = \tilde{z}_1 + \psi_y(y) + b_\rho(\hat{\xi}_1 + \tilde{\xi}_1 - \mu) \tag{20}$$

The virtual control is designed as

$$\hat{\xi}_1 = \frac{1}{b_\rho}\left(-c_0 y - \psi_y(y)\right) + \hat{W}^T S(y) \tag{21}$$

Substituting equation (21) into (20), it gives

$$\dot{y} = -c_0 y + \tilde{z}_1 + b_\rho\tilde{\xi}_1 - b_\rho\tilde{W}^T S(y) - b_\rho\delta \tag{22}$$

with the notation $\tilde{W} = W - \hat{W}$.

The backstepping technique is employed to search other virtual controls step by step.

**Step 1**: When $i = 1$,

$$\begin{array}{rcl} \dot{\tilde{\xi}}_1 & = & \dot{\xi}_1 - \dot{\hat{\xi}}_1 \\ & = & -\lambda_1\xi_1 + \xi_2 - \frac{\partial\hat{\xi}_1}{\partial\hat{W}}\dot{\hat{W}} \\ & & -\frac{\partial\hat{\xi}_1}{\partial y}\left(\tilde{z}_1 + \psi_y(y) + b_\rho\left(\xi_1 - \mu\right)\right) \\ & = & -\lambda_1\xi_1 + \tilde{\xi}_2 + \hat{\xi}_2 - \frac{\partial\hat{\xi}_1}{\partial\hat{W}}\dot{\hat{W}} \\ & & -\frac{\partial\hat{\xi}_1}{\partial y}\left(\tilde{z}_1 + \psi_y(y) + b_\rho\left(\xi_1 - \mu\right)\right) \end{array} \tag{23}$$

Design $\hat{\xi}_2$ as

$$\begin{array}{rcl} \hat{\xi}_2 & = & \lambda_1\xi_1 - c_1\tilde{\xi}_1 - h_1\left(\frac{\partial\hat{\xi}_1}{\partial y}\right)^2\tilde{\xi}_1 + \frac{\partial\hat{\xi}_1}{\partial\hat{W}}\dot{\hat{W}} \\ & & +\frac{\partial\hat{\xi}_1}{\partial y}\left(\psi_y(y) + b_\rho\left(\xi_1 - \hat{W}^T S(y)\right)\right) \end{array} \tag{24}$$

**Step i**: Introduce $\xi_\rho$ for the convenience of notation. Similarly, virtual controls are obtained as:

$$\begin{aligned}\hat{\xi}_i \;=\;& -\tilde{\xi}_{i-2}+\lambda_{i-1}\xi_{i-1}-c_{i-1}\tilde{\xi}_{i-1}\\ &+\frac{\partial\hat{\xi}_{i-1}}{\partial y}\left(\psi_y(y)+b_\rho\left(\xi_1-\hat{W}^T S(y)\right)\right)\\ &+\sum_{i=1}^{i-2}\frac{\partial\hat{\xi}_{i-1}}{\partial\xi_i}\dot{\xi}_i-h_{i-1}\left(\frac{\partial\hat{\xi}_{i-1}}{\partial y}\right)^2\tilde{\xi}_{i-1}\\ &+\frac{\partial\hat{\xi}_{i-1}}{\partial\hat{W}}\dot{\hat{W}} \qquad \text{for}\quad i=2,\ldots,\rho \qquad (25)\end{aligned}$$

Finally, the control input $u$ is designed as

$$u=\frac{\hat{\xi}_\rho-\hat{\Theta}r(\tilde{\xi}_{\rho-1},y)}{\sigma(y)} \qquad (26)$$

where $\hat{\Theta}$ is the estimate of $\Theta\in\mathbb{R}$ that is an unknown constant, Smooth nonlinear function $r(\tilde{\xi}_{\rho-1},y)$ is given by

$$r(\tilde{\xi}_{\rho-1},y)=\frac{1}{\tilde{\xi}_{\rho-1}}\left(\|\psi_z(y)\|^2\right) \qquad (27)$$

where $\Theta$ and $r(\tilde{\xi}_{\rho-1},y)$ are introduced for the sake of compensating the closed-loop influence brought by the nonlinear function $\psi_z(y)$. The desired value of $\Theta$ is given by

$$\Theta=\beta\|P_z\|^2 \qquad (28)$$

where $\beta$ is a positive constant satisfying inequality below

$$\beta(1-\lambda_{min})+\frac{1}{2}+\sum_{i=1}^{\rho-1}\frac{1}{4h_i}<0 \qquad (29)$$

where $\lambda_{min}$ denotes the minimum eigenvalue of positive definite matrix $P_z$.

Estimates of unknown parameters $\Theta$ and $W$ are generated by following adaptive laws.

$$\begin{cases}\dot{\hat{\Theta}} \;=\; \gamma_\theta\left(\tilde{\xi}_{\rho-1}r(\tilde{\xi}_{\rho-1},y)-\sigma_\theta\hat{\Theta}\right)\\ \dot{\hat{W}} \;=\; -\Gamma_w\left(b_\rho yS(y)-\displaystyle\sum_{i=1}^{\rho-1}b_\rho\tilde{\xi}_i\frac{\partial\hat{\xi}_i}{\partial y}S(y)+\sigma_w\hat{W}\right)\end{cases} \qquad (30)$$

where $\gamma_\theta\in\mathbb{R}$ and $\Gamma_w\in\mathbb{R}^l$ are positive definite designed adaptive gain matrix, $\sigma_\theta\in\mathbb{R}$ and $\sigma_w\in\mathbb{R}$ are $\sigma$ modification gains which are selected to be small positive constants.

The dynamics of $\tilde{\xi}$ is given by

$$\begin{cases}\dot{\tilde{\xi}}_1 \;=\; \tilde{\xi}_2-c_1\tilde{\xi}_1-h_1\left(\dfrac{\partial\hat{\xi}_1}{\partial y}\right)^2\tilde{\xi}_1-\dfrac{\partial\hat{\xi}_1}{\partial y}\tilde{z}_1\\ \qquad +b_\rho\dfrac{\partial\hat{\xi}_1}{\partial y}\tilde{W}^T S(y)+b_\rho\dfrac{\partial\hat{\xi}_1}{\partial y}\delta\\ \dot{\tilde{\xi}}_i \;=\; -\tilde{\xi}_{i-1}+\tilde{\xi}_{i+1}-c_i\tilde{\xi}_i-h_i\left(\dfrac{\partial\hat{\xi}_i}{\partial y}\right)^2\tilde{\xi}_i\\ \qquad -\dfrac{\partial\hat{\xi}_i}{\partial y}\tilde{z}_1+b_\rho\dfrac{\partial\hat{\xi}_i}{\partial y}\tilde{W}^T S(y)+b_\rho\dfrac{\partial\hat{\xi}_i}{\partial y}\delta\\ \qquad \text{for}\quad i=2,\ldots,\rho-2\\ \dot{\tilde{\xi}}_{\rho-1} \;=\; -\tilde{\xi}_{\rho-2}-c_{\rho-1}\tilde{\xi}_{\rho-1}-\dfrac{\partial\hat{\xi}_{\rho-1}}{\partial y}\tilde{z}_1\\ \qquad -h_{\rho-1}\left(\dfrac{\partial\hat{\xi}_{\rho-1}}{\partial y}\right)^2\tilde{\xi}_{\rho-1}+b_\rho\dfrac{\partial\hat{\xi}_{\rho-1}}{\partial y}\delta\\ \qquad +b_\rho\dfrac{\partial\hat{\xi}_{\rho-1}}{\partial y}\tilde{W}^T S(y)-\hat{\Theta}r(\tilde{\xi}_{\rho-1},y)\end{cases} \qquad (31)$$

## VI. Stability analysis

The stability analysis is based on Lyapunov stabilising theory. The Lyapunov function candidate of this system is chosen as

$$V=V_y+\beta V_z+\sum_{i=1}^{\rho-1}V_{\xi_i}+V_w+V_\theta \qquad (32)$$

with

$$\begin{cases}V_y \;=\; \dfrac{1}{2}y^2\\ V_z \;=\; \tilde{z}^T P_z\tilde{z}\\ \displaystyle\sum_{i=1}^{\rho-1}V_{\xi_i} \;=\; \dfrac{1}{2}\sum_{i=1}^{\rho-1}\tilde{\xi}_i^2\\ V_w \;=\; \dfrac{1}{2}\tilde{W}^T\Gamma_w^{-1}\tilde{W}\\ V_\theta \;=\; \dfrac{1}{2}\gamma_\theta^{-1}\tilde{\Theta}^2\end{cases} \qquad (33)$$

with the notation $\tilde{\Theta}=\Theta-\hat{\Theta}$.

From equation (22), it shows

$$\begin{aligned}\dot{V}_y \;=&\; y\dot{y}\\ =&\; -c_0y^2+y\tilde{z}_1+b_\rho y\tilde{\xi}_1-b_\rho y\tilde{W}^T S(y)-b_\rho y\delta\\ \leq&\; -c_0y^2+\frac{1}{2}y^2+\frac{1}{2}\|\tilde{z}\|^2+\frac{b_\rho^2}{4}y^2+\tilde{\xi}_1^2\\ &\; -b_\rho y\tilde{W}^T S(y)-b_\rho y\delta\\ \leq&\; \left(-c_0+\frac{1}{2}+\frac{b_\rho^2}{4}\right)y^2+\frac{1}{2}\|\tilde{z}\|^2+\tilde{\xi}_1^2\\ &\; -b_\rho y\tilde{W}^T S(y)-b_\rho y\delta \qquad (34)\end{aligned}$$

then from equation (9), it is obtained

$$\begin{aligned}\dot{V}_z &= \dot{\tilde{z}}^T P_z \tilde{z} + \tilde{z}^T P_z \dot{\tilde{z}} \\ &= \tilde{z}^T \left(D^T P_z + P_z D\right) \tilde{z} + 2\tilde{z}^T P_z \psi_z(y) \\ &\le (1-\lambda_{min})\|\tilde{z}\|^2 + \|P_z\|^2 \|\psi_z(y)\|^2 \end{aligned} \tag{35}$$

With dynamics of $\tilde{\xi}_i$ in equation (31), it is derived that

$$\begin{aligned}\sum_{i=1}^{\rho-1} \dot{V}_{\xi_i} &= -\sum_{i=1}^{\rho-1}\left(c_i\tilde{\xi}_i^2 + h_i\left(\frac{\partial\hat{\xi}_i}{\partial y}\right)^2 \tilde{\xi}_i + \frac{\partial\hat{\xi}_i}{\partial y}\tilde{z}_1\right) \\ &\quad + \sum_{i=1}^{\rho-1} b_\rho \tilde{\xi}_i \frac{\partial\hat{\xi}_i}{\partial y}\tilde{W}^T S(y) + \sum_{i=1}^{\rho-1} b_\rho \tilde{\xi}_i \frac{\partial\hat{\xi}_i}{\partial y}\delta \\ &\quad - \tilde{\xi}_{\rho-1}\hat{\Theta} r(\tilde{\xi}_{\rho-1}, y) \\ &\le -\sum_{i=1}^{\rho-1}\left(c_i\tilde{\xi}_i^2 - \frac{1}{4h_i}\|\tilde{z}\|^2\right) + \sum_{i=1}^{\rho-1} b_\rho \tilde{\xi}_i \frac{\partial\hat{\xi}_i}{\partial y}\delta \\ &\quad + \sum_{i=1}^{\rho-1} b_\rho \tilde{\xi}_i \frac{\partial\hat{\xi}_i}{\partial y}\tilde{W}^T S(y) - \tilde{\xi}_{\rho-1}\hat{\Theta} r(\tilde{\xi}_{\rho-1}, y) \end{aligned} \tag{36}$$

Further, from equation (30), we have

$$\begin{aligned}\dot{V}_w &= -\tilde{W}^T \Gamma_w^{-1} \dot{\hat{W}} \\ &= b_\rho y \tilde{W}^T S(y) - \sum_{i=1}^{\rho-1} b_\rho \tilde{\xi}_i \frac{\partial\hat{\xi}_i}{\partial y}\tilde{W}^T S(y) + \sigma_w \tilde{W}^T \hat{W} \end{aligned} \tag{37}$$

$$\begin{aligned}\dot{V}_\theta &= -\gamma_\theta^{-1}\tilde{\Theta}\dot{\hat{\Theta}} \\ &= -\tilde{\xi}_{\rho-1}\tilde{\Theta} r(\tilde{\xi}_{\rho-1}, y) + \sigma_\theta \tilde{\Theta}\hat{\Theta} \end{aligned} \tag{38}$$

Therefore

$$\begin{aligned}\dot{V} &\le \left(-c_0 + \frac{1}{2} + \frac{b_\rho^2}{4}\right) y^2 - (c_1+1)\tilde{\xi}_1^2 - \sum_{i=2}^{\rho-1} c_i \tilde{\xi}_i^2 \\ &\quad + \left(\beta(1-\lambda_{min}) + \frac{1}{2} + \sum_{i=1}^{\rho-1}\frac{1}{4h_i}\right)\|\tilde{z}\|^2 \\ &\quad + \beta\|P_z\|^2\|\psi_z(y)\|^2 - \tilde{\xi}_{\rho-1}\Theta r(\tilde{\xi}_{\rho-1}, y) \\ &\quad + \sigma_w \tilde{W}^T \hat{W} + \sigma_\theta \tilde{\Theta}\hat{\Theta} - b_\rho y \delta + \sum_{i=1}^{\rho-1} b_\rho \tilde{\xi}_i \frac{\partial\hat{\xi}_i}{\partial y}\delta \end{aligned} \tag{39}$$

$$\begin{aligned} &\le 2\left(-c_0 + \frac{1}{2} + \frac{b_\rho^2}{4}\right) V_y - 2(c_1+1)V_{\xi_1} \\ &\quad + \left(\beta(1-\lambda_{min}) + \frac{1}{2} + \sum_{i=1}^{\rho-1}\frac{1}{4h_i}\right)\frac{V_z}{\lambda_p} \\ &\quad - \sum_{i=2}^{\rho-1} 2c_i V_{\xi_i} - \sigma_w \Gamma_w V_w - \sigma_\theta \gamma_\theta V_\theta + \frac{\sigma_w}{2}\|W\|^2 \\ &\quad + \frac{\sigma_\theta}{2}\|\Theta\|^2 - b_\rho y \delta + \sum_{i=1}^{\rho-1} b_\rho \tilde{\xi}_i \frac{\partial\hat{\xi}_i}{\partial y}\delta \end{aligned} \tag{40}$$

Note that the derivative of $V$ is in a form of $\dot{V} = -\zeta V + \eta$ where $\zeta > 0$, and $\eta$ is bounded. The system converges to a small region around its equilibrium points when time $t$ tends to infinity.

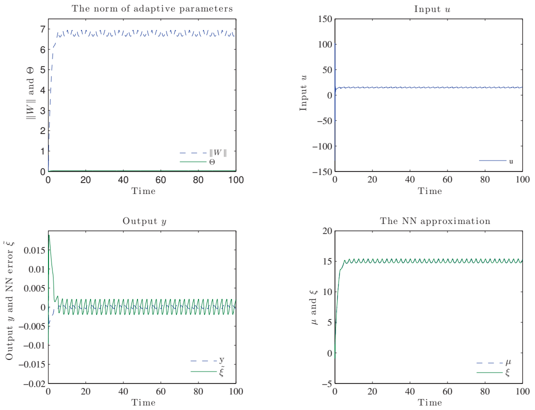


Fig. 1: The simulation result

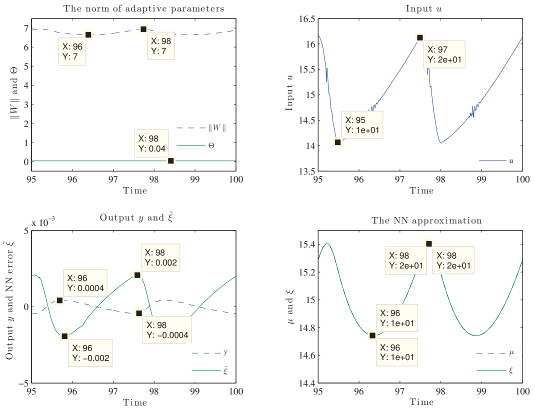


Fig. 2: An enlargement of Figure 1

## VII. Simulation study

Consider a three order disturbed nonlinear system as

$$\begin{cases}\dot{x}_1 &= x_2 + y^2 \\ \dot{x}_2 &= x_3 + y^3 + 2w + (|y| + 0.1)u \\ \dot{x}_3 &= 3w + (|y| + 0.1)u\end{cases}$$

The filter is selected as

$$\dot{\xi} = -\xi + (|y| + 0.1)u$$

with the virtual control given by

$$\hat{\xi} = -3y - 2y - y^2 + W^T S(y)$$

It is easy to find

$$D = \begin{bmatrix} -2 & 1 \\ -1 & 0 \end{bmatrix}$$

Adaptive laws are designed as

$$\begin{cases}\dot{\hat{W}} &= -\Gamma_w(yS(y) - \tilde{\xi}\dfrac{\partial\hat{\xi}}{y}S(y) + \sigma_W W) \\ \dot{\hat{\Theta}} &= \gamma_\theta\left((-3y + y^3 - 2y^2)^2 + (-2y - y^2)^2 - \sigma_\theta\Theta\right)\end{cases}$$

with adaptive gains $\Gamma_w = 100I$, $\gamma_\theta = 100$, and $\sigma$-modification gains $\sigma_W = 0.00001$, $\sigma_\theta = 0.00003$. Following our backstepping design procedure, the control input $u$ is designed as

$$\begin{aligned} u &= \frac{1}{|y|+0.1}\left(\xi - \tilde{\xi} - \left(\frac{\partial \hat{\xi}_1}{\partial y}\right)^2 \tilde{\xi} + \frac{\partial \hat{\xi}}{\partial \hat{W}}\dot{\hat{W}}\right) \\ &\quad + \frac{1}{|y|+0.1}\frac{\partial \hat{\xi}}{\partial y}(2y+y^2+\xi-\hat{W}^T S(y)) \\ &\quad - \frac{1}{|y|+0.1}\frac{\hat{\Theta}}{\tilde{\xi}}\left((-3y+y^3-2y^2)^2+(-2y-y^2)^2\right) \end{aligned}$$

where

$$\begin{aligned} \frac{\partial \hat{\xi}_1}{\partial y} &= -\frac{1}{5+2y} + W^T S'(y) \\ \frac{\partial \hat{\xi}}{\partial \hat{W}} &= S^T(y) \end{aligned}$$

The simulation result is illustrated in Fig. 1 and Fig. 2. Fig. 2 shows the steady state of systems. It can be observed that all states are going to be stable. The first subfigure illustrates the norm of adaptive parameters, the second subfigure presents the system control input which has a little transient attenuation that is caused by the small NN approximation error, the third subfigure shows the convergence of output ($|y| < 0.0001$) and NN approximation error ($|\tilde{\xi}| < 0.002$), to the end, it is also observed that the NN approximator works well from last subfigure.

## VIII. Conclusions

In this paper, disturbance rejection is achieved for a periodically disturbed nonlinear system in output feedback form with time varying input weighting vector. NN is applied to approximate the feedforward control input $\mu$. Compared with the results in [7], it is noticed the control design method is in the same way with different disturbance approximation methods. The proposed NN method estimates the feedforward control input without any pre-known information. While in [7], the wave profile of the disturbance signal is assumed to be half-period alternative. The disturbance period is known for the regeneration of the real amplitude and phase shift. Nevertheless, the method in [7] has a quick and nice time response. Generally, it is not easy to have the period of the disturbance in the real practice. The NN approximation method is a good choice when the information of disturbance signals are not enough.

## References

[1] Z. Ding, "Global stabilization and disturbance suppression of a class of nonlinear systems with uncertain internal model," *Automatica*, vol. 39(3), pp. 471–479, 2003.

[2] ——, "Adaptive tracking with complete compensation of unknown disturbances or nonlinear output feedback systems," *IEE Proc.-Control Theory Appl.*, vol. 149(6), pp. 533–539, 2002.

[3] M. Bodson, A. Sacks, and P. Khosla, "Harmonic generation in adaptive feedforward cancellation schemes," *IEEE Trans. Autom. Control*, vol. 39(9), pp. 1939–1944, 1994.

[4] M. Bodson and S. C. Douglas, "Adaptive algorithms for the rejection of sinusoidal disturbances with unknown frequencies," *Automatica*, vol. 33(10), pp. 2213–2221, 1997.

[5] Z. Ding, "Output regulation of uncertain nonlinear systems with nonlinear exosystems," *IEEE Transaction on Automatic Control*, vol. 51(3), pp. 498–503, 2006.

[6] ——, "Asymptotic rejection of a class of periodic disturbances in nonlinear output-feedback nonlinear systems," *Automatica*, vol. 43(3), pp. 555–561, 2006.

[7] ——, "Asymptotic rejection of unmatched general periodic disturbances in a class of nonlinear systems," *IET Control Theory Appl.*, vol. 2(4), pp. 269–276, 2008.

[8] W. Chen and Y. Tian, "Neural network approximation for periodically disturbed functions and applications to control design," *Neuralcomputing*, vol. 72, pp. 3891–3900, 2009.

[9] W. Chen, "Output-feedback adaptive stochastic nonlinear stabilization using neural networks," *Journal of Systems Engineering and Electronics*, vol. 21(1), pp. 81–87, 2010.

[10] X.Tang and Z. Ding, "Periodic disturbance rejection of nonlinear systems via output feedback with neural network approximation," presented at the Proc. of the 30th Chinese Contrl Conf., Yantai, China, 2011, pp. 705–710.

[11] Z. Ding, "Analysis and design of robust adaptive control for nonlinear output feedback systems under disturbances with unknown bounds," *IEE Proc.-Control Theory Appl.*, vol. 147(6), pp. 655–663, 2000.

[12] ——, "Adaptive disturbance rejection of nonlinear systems in an extended output feedback form," *IET Control Theory Appl.*, vol. 1(1), pp. 298–303, 2007.

[13] A. Isidori, *Nonlinear Control Systems the 3rd Ed.* NewYork, NY: Springer, 1995.

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# A systematic selection of an alternative parameterisation for predictive control

Bilal Khan
Automatic Control and Systems Engineering
The University of Sheffield,UK
Email: b.khan@sheffield.ac.uk

John Anthony Rossiter
Automatic Control and Systems Engineering
The University of Sheffield,UK
Email: j.a.rossiter@sheffield.ac.uk

***Abstract*—Alternative parameterisations have been shown to improve the feasible region for a predictive control law when the number of degrees of freedom is limited. One question yet to be resolved is: which alternative parameterisation is best for a particular problem and what choice of parameter(s) within each parameterisation will lead to an improved feasible region with good performance? This paper tackles this question and demonstrates two systematic approaches to select the best alternative parameterisations. These approaches are based on multiobjective optimisation and a pragmatic selection. Numerical examples demonstrate the efficacy of both methods.**

**Keywords: Alternative parameterisation, Feasibility, multiobjective optimisation.**

## I. Introduction

Model Based Predictive Control (MPC) or receding horizon control (RHC) or embedded optimisation or moving horizon or predictive control [2], [3], [1], is the general name for different computer control algorithms that use past information of the inputs and outputs and a mathematical model of the plant to optimise predicted future behaviour. MPC has been developed widely both in the process industry and control research community and has reached a high degree of maturity in its linear variant. Currently MPC research focuses on stochastic and nonlinear scenarios, robustness and fast optimisation or related computational aspects.

All algorithms to some extent form a trade-off between feasibility, optimality and inexpensive optimisation, but may not have systematic tools for doing this trade-off. The success of earlier industrial heuristic MPC algorithms motivated the research community to develop several algorithms with improved performance and feasibility. There are several successful theoretical approaches but few of them are exploited commercially for real-time implementation. One important issue for real-time implementation is the ability to do a systematic trade-off between feasibility, performance and computational burden when choosing from currently available algorithms. In recent years many authors have focussed on parametric solutions [5] or efficient implementations of online optimisers (e.g. for quadratic programming). However, this paper will follow a different route and consider the underlying structure of the MPC algorithm. Specifically, this paper proposes a systematic selection of an alternative parameterisation to enlarge the feasible region without too much detriment to performance (or optimality) and the computational burden.

In an MPC problem formulation, stability can be ensured using dual mode prediction by including a terminal constraint at the expense of the computational load [16]. In dual mode prediction, different formulations of the degrees of freedom (d.o.f.) or decision variables have been utilized using interpolation techniques [8] to enlarge feasible region without detriment to performance. There are different variants of interpolation, however, these methods are limited to small dimensional systems. Another approach in the literature is a concept of triple mode control [6], [7]; the triple mode strategy introduces an extra mode which may reduce t he online computational burden with good performance and feasibility. However, in this strategy, the challenge is to find a suitable linear time varying (LTV) control law which enlarge feasible region with improved performance. One pragmatic solution to finding this law is tackled in [12] which used Laguerre parameterisations. However, this paper does not pursue Triple mode approaches as the offline complexity and decision making is substantially increased as compared to dual mode approaches.

More recently, alternative parameterisation based on Laguerre [17] and Kautz functions have been proposed to simplify the trade-off without increasing computational burden within dual mode prediction paradigm [9], [10], [11]. The main idea is to form the degrees of freedom in the predictions as a combination of either Laguerre or Kautz functions. These functions have proven to be a very effective for improving the volume of the feasible region with a limited number of d.o.f. These alternative parameterisations are generalised using orthonormal basis functions with Laguerre and Kautz functions as special cases [13].

The key question left to be resolved is, what is the best prediction dynamics to assume for predicted inputs, that is the d.o.f., to allow for a large feasible region without detriment to the closed loop performance and computational burden? In [15], different ways of parameterising the d.o.f. were investigated and a mechanism based on a Monte Carlo approach to define the best parameterisation using optimal sequences for numerous search directions was considered. However, a systematic choice for the underlying dynamics remains an open question. This paper focuses on systematic approaches to identify the best parameterisation dynamics using: (i) a multiobjective optimisation and (ii) a pragmatic choice.

This paper assumes that Laguerre, Kautz and Generalised parameterisation are able to achieve large feasible regions while maintaining local optimality and a relatively low computational complexity [9], [10], [11], [12], [13], [14] and extends the earlier studies in [11], [13] in order to identify a systematic selection of the best parameterisation dynamics. Section II will give the necessary background about an optimal MPC, Laguerre MPC, generalised parameterisation for optimal MPC. Section III discusses two proposed schemes to identify the best parameterisation dynamics based on a muti-objective optimisation and a pragmatic approach. Numerical examples are presented in section IV and paper finishes with conclusions and future work in section V.

## II. Background

This section will summarise the background information related to nominal dual-mode MPC and the use of alternative parameterisations within MPC.

### A. Problem formulation for MPC

Assume a discrete time linear time invariant (LTI) state space model of the form

$$x_{k+1} = Ax_k + Bu_k \tag{1}$$

where $x_k \in R^{n_x}$ and $u_k \in R^{n_u}$ which are the state vectors and the plant input respectively. Assume that the states and inputs at all time instants should fulfill the following constraints:

$$\underline{u} \le u_k \le \overline{u}; \qquad \underline{\Delta u} \le \Delta u_k \le \overline{\Delta u}; \quad \underline{x} \le x_k \le \overline{x} \tag{2}$$

### B. Nominal MPC algorithm

In dual-mode MPC it is assumed that one has total freedom in the choice of the input signal $u_k$ up to horizon $n_c$ subject to constraints. Beyond horizon $n_c$, a terminal control law with an asymptotic stabilising optimal feedback gain $K$ is assumed. The 'predicted' control law [3], [16] takes the form

$$\begin{aligned} u_k &= -Kx_k + c_k, \qquad & k = 0, ..., n_c - 1,\\ u_k &= -Kx_k, & k \ge n_c, \end{aligned} \tag{3}$$

where only first control move is ever implemented, $c_k$ are degrees of freedom (d.o.f.) available for constrained handling. The terminal control law defines an invariant set $\chi_0$ for state vector $x_k$ [8]. This invariant set is also known as maximum admissible set (MAS) which satisfies all polytopic constraints with recursive use of the terminal control law $u_k = -Kx_k \in \chi_0$. The MAS is defined as:

$$\begin{aligned} \chi_0 = \{x_0 \in R^{n_x} : \underline{x} \le x_k \le \overline{x}, \underline{u} \le -Kx_k \le \overline{u},\\ x(k+1) = Ax(k) + Bu(k), \forall k \ge 0\} \end{aligned} \tag{4}$$

In compact form defined as $\chi_0 = \{x_k : Mx_k \le b\}$ for suitable $M$ and $b$. Performance, either predicted or actual, will be assessed by the cost

$$J = \sum_{k=0}^{\infty} x_k^T Q x_k + u_k^T R u_k \tag{5}$$

Substituting the nominal model and predicted control values (3) into (5) and ignoring terms that do not depend upon the d.o.f., one finds from [3] that the optimisation problem in (5) can be reformulated as:

$$\min_{C} \; J_c = C^T S C \; s.t. \; Mx_k + NC \le b; \tag{6}$$

where $C = [c_k^T, \dots, c_{k+n_c-1}^T]$. Details are in the literature [2], [3], [4]. The maximal control admissible set (MCAS) $\chi_c$, the feasible set for optimal control problem in (6) that satisfies all polytopic constraints, is defined as:

$$\chi_c = \{x_k : \exists C, Mx_k + NC \le b\} \tag{7}$$

The optimal MPC (OMPC) algorithm with guarantees of recursive feasibility and convergence, for the nominal case is given by solving QP optimisation (6) at every sampling instance then implementing the first component of $C$, that is $c_k$ in the control law of (3). The algorithm is formulated as:

*Algorithm 2.1:* OMPC [16], [3]

$$c_k^* = arg \min_{c_k} \; J_c \quad s.t. \quad Mx_k + Nc_k \le b;$$

Implement $u_k = -Kx_k + e_1^T c_k^*$ where $e_1^T = [I, 0, \dots, 0]$. When the initial states $x_k \in \chi_0$ then the optimising $c_k^*$ is zero so the terminal control law $u_k = -Kx_k$ is implemented.

*Remark 2.1:* There is a well understood set of potentially conflicting objectives using OMPC e.g. between the desire for good performance and large feasible regions with the equally important desire to keep the number of d.o.f. small. In order for OMPC to obtain a large feasible region and good performance, a large number of d.o.f. or $n_c$ is required.

### C. Generalised parameterisation for Optimal MPC

Alternative parameterisation techniques for the d.o.f. in the future control values have been developed to improve the feasible region in nominal case. Laguerre, Kautz and generalised parameterisations have been proposed in [9], [11], [13] as an effective alternative to the standard basis set. This section will summarise Generalised optimal MPC.

*1) Generalised functions:* The generalised parameterisation [13] is defined using a higher order network such as:

$$\begin{aligned} \mathcal{G}_i(z) = \mathcal{G}_{i-1}(z)\frac{(z^{-1} - a_1)\dots(z^{-1} - a_n)}{(1 - a_1 z^{-1})\dots(1 - a_n z^{-1})};\\ 0 \le a_k < 1, \qquad k = 1 \dots n \end{aligned} \tag{8}$$

With $\mathcal{G}_1(z) = \frac{\sqrt{(1-a_1^2)\dots(1-a_n^2)}}{(1-a_1 z^{-1})\dots(1-a_n z^{-1})}$. The generalised function with $a_k, \forall k = 1, \dots, n$ gives [13]

$$\begin{aligned} &Laguerre: \qquad &\mathcal{G}_i = \mathcal{L}_i, \quad if \;\; a_k = [a]\\ &Kautz: &\mathcal{G}_i = \mathcal{K}_i, \quad if \;\; a_k = [a, b]. \end{aligned} \tag{9}$$

The Laguerre function is a special case of a generalised function and may be computed using the following state-space dynamics model:

$$\mathcal{G}(k+1) = \underbrace{\begin{pmatrix} a & 0 & 0 & 0 & \dots \\ \beta & a & 0 & 0 & \dots \\ -a\beta & \beta & a & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}}_{A_G} \mathcal{G}(k); \tag{10}$$

$$\mathcal{G}(0) = \sqrt{1-a^2}[1, -a, a^2, \dots]^T; \ \beta = 1 - a^2$$

*2) GOMPC: generalised functions and MPC:* The predictions using d.o.f. based on generalised functions [13] are:

$$C = \begin{pmatrix} c_k \\ \vdots \\ c_{k+n_c-1} \\ \vdots \end{pmatrix} = \begin{pmatrix} \mathcal{G}(0)^T \\ \mathcal{G}(1)^T \\ \vdots \end{pmatrix} \rho = H_G \rho \tag{11}$$

where $\rho$ is the $n_G$ dimension decision variable when one uses the first $n_G$ column of $H_G$. The only difference between Laguerre MPC and Kautz or generalised MPC is that of $H_G$ matrix. For further details readers are referred to [11] and [13].

*Algorithm 2.2:* GOMPC

$$\begin{aligned} &\rho^* = arg \ \min_{\rho} \ \rho^T [\sum_{i=0}^{\infty} A_G^i \mathcal{G}(0) S \mathcal{G}(0)^T (A_G^i)^T] \rho \\ &s.t. \ Mx_k + NH_G \ \rho \ \leq \ b \end{aligned} \tag{12}$$

Define $c_k^* = H_G \rho_k^*$ and implement $u_k = -Kx_k + e_i^T c_k^*$. where $e_i^T$ is the $i$-th standard basis vector.

*Remark 2.2:* It is straightforward to show, with conventional arguments, that all algorithms (i.e. LOMPC, KOMPC, GOMPC) using terminal constraints within MPC problem formulation provides recursive feasibility and Lyapunov stability.

## III. Best alternative parameterisation selection

In generalised parameterisation there are two clear choices within the future input predictions. First one can choose the order of the dynamics, that is the number of poles $a_i$ in $G_i(z)$, and second is the parameter selection(s), that is the actual values of $a_i$. This section discusses the systematic selection of the parameterisation dynamics by asking what impact this choice has on feasibility and performance?

### A. The best choice for order of the prediction dynamics

The prediction dynamics for the 3rd order prediction dynamics from (8) (which can easily be extended to $n^{th}$ order) can be defined as:

$$\begin{aligned} \mathcal{G}(k+1) &= \underbrace{\begin{pmatrix} b & 0 & 0 & \dots \\ b & c & 0 & \dots \\ -ab & (1-ac) & a & \dots \\ ab^2 & -b(1-ac) & (1-ac) & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}}_{A_G} \mathcal{G}(k) \\ \mathcal{G}(0) &= \gamma[1, 1, 1, -a, ab, -abc, a^2bc, \dots]^T, \\ \gamma &= \sqrt{(1-a^2)(1-b^2)(1-c^2)}, \end{aligned} \tag{13}$$

The dynamic structure of the prediction in (13) is quite generic with distinct eigenvalues. To fulfill the algebraic relations in (11), $A_G$ used in (13) must be a square matrix with a maximum number of dimensions the same as $n_c$. Moreover a key observation from the prediction matrix structure is that $dim(A_G) = n_c$ is an upper bound on dynamic dimensions. In fact, one could select $dim(A_G) \leq n_c$.

### B. Multiobjective solution to select best parameterisation

The main objective of this paper is to formulate a systematic method for handling the compromise between feasibility, performance and computational burden. This section shows how one can produce trade-off curves between feasible volume $v$, performance loss $\beta$ and computational burden (implicitly linked to $n_c$) of the parameterised MPC problem. Such trade-off curves allow us to formulate a multi-objective optimisation problem as a function of parameterisation parameters $\alpha = [a_1, \dots, a_m] \in (0, 1)^m$:

$$\begin{aligned} J(n_c, n) := \min_{\alpha} \beta, &\max_{\alpha} v \\ s.t. \ \ Mx + NH\rho &\leq b, \\ v &= \frac{vol(P_H)}{vol(P_{opt})}, \\ \beta &= \frac{1}{n} \sum \frac{J_H(x) - J_{opt}(x)}{J_{opt}(x)}, \\ \alpha = [a_1, \dots, a_m], \ 0 < a_i &< 1, \ m \leq n_c, \\ \mathcal{G}(k+1) &= A_G \mathcal{G}(k), \\ H &= \begin{pmatrix} \mathcal{G}^T(0) & \mathcal{G}^T(1) & \dots \end{pmatrix}^T, \end{aligned} \tag{14}$$

where $P_{opt} := \{(x, c) | Mx + Nc \leq b\}$ represents the MCAS with $n_c = 20$ (used as an approximate for the global maximum MCAS) for comparison, $P_H = \{(x, \rho) | Mx + NH\rho \leq b\}$ is the polytope sliced by the matrix $H$, $vol(.)$ is the volume, $J_{opt}(x) = avg. \ \{J_c(x, c) | (x, c) \in P_{opt}\}$ and $J_H(x) = avg. \ \{J_c(x, H\rho) | (x, \rho) \in P_H\}$ for convex function $J_c(x, c)$.

*1) Volume approximation:* Computing the volume of a high dimensional polytope is a complex task, and can, in the worst case be exponential in the size of the data $M$ and $N$ or $NH$. Consequently, this paper approximates the volume. First select a large number of equi-spaced (by solid angle) or random directions in the state space i.e. $x = (x_1, \dots, x_n)$ and then, for each direction, the distance from the origin to the boundary of MCAS is determined by solving a linear programming (LP) and clearly the larger the distance, hereafter denoted as radius, the better the feasibility. Finally radii are normalised against the radii obtained for $P_{opt}$. Although this might be considered somewhat arbitrary, it seems a pragmatic way of indicating relative volumes of different feasible regions compared to the best feasible shape with a large $n_c$. One might argue that a precise volume measurement would, in some sense, be an equally arbitrary comparison.

*2) Performance approximation:* Similarly in case of calculating the performance loss, it is converted from integral to sampled sum i.e.

$$\beta :\approx \frac{1}{n}\sum \frac{J_H(x)}{J_{opt}(x)} - 1 \qquad (15)$$

That is, one computes the predicted performance for all the given state directions and compares each of these to the 'global' optimum. Equation (15) normalises these so that the best achievable performance would correspond to a $\beta$ measure of unity.

By deploying explicit numeric measures of volume and performance, the multi-objective optimisation problem is able to generate trade-off curves between the d.o.f. $n_c$, average radii and average performance loss of the parameterised problem with different parameterisation settings.

### C. Pragmatic selection

In practice it is known from parameterisation insight that the optimal $\rho_k$ is highly nonlinear in terms of its dependence upon the current state $x_k$; more over the solution is to some extent unpredictable. The following gives a simple approach to identify a pragmatic selection of pole location(s) and order selection of parameterisation dynamics.

*1) Order selection:* The generalised functions with higher order orthonormal functions have more flexibility to improve feasibility while retaining good performance with a limited number of d.o.f. [13]. So $dim(A_G) = n_c$ is a sub-optimal choice to select order parameterisation dynamics.

*2) Parameter selection:* It is observed from simulation results that there is a relation between the closed loop poles and good locations of parameterisation dynamics. The pole locations of generalised functions can be selected to be equal to poles of the closed loop system using an optimal gain $K$; this may be sub-optimal but is efficient.

### D. Summary

In summary, there are two choices: (i) the order of parameterisation dynamics $(dim(A_G))$ and (ii) parameter value(s) or pole location(s) $([a_1, \ldots, a_n])$. One can use multi-objective optimisation to find $dim(A_G)$ and $[a_1, \ldots, a_n]$. Another key observation is that the maximum $dim(A_G) \leq n_c$ and closed loop eigenvalues or pole locations with an optimal gain $K$ are a pragmatic choice to tune the parameterisation dynamics.

## IV. Numerical Examples

In this section three numerical examples are presented to illustrate the efficacy of the proposed approach. The aim is to compare two aspects: (i) feasible regions; (ii) performance loss for the generalised parameterisation dynamics with different parameter(s) or pole location(s) selection by varying the d.o.f. or $n_c$. The trade-off curves are shown between average radius and performance loss as a function of the different parameterisation parameter(s) and d.o.f. or $n_c$. The parallel coordinates are also shown to represent an alternative variation of the parameter(s) and the effects on average performance loss and average feasibility gain. The multi-objective optimisation is done using the NSG-II Matlab toolbox.

### A. Examples

Three numerical examples are simulated using weightings $Q = C^TC$, $R = 2$ and are simulated by varying $n_c = [2, \ldots, 6]$ and with different parameter values. Trade-off curves between average radius and average performance loss are plotted by solving the multi-objective optimisation with varying parameter values. Similarly the effect on average performance loss and average radius are shown by varying $n_c$. Conversely, Table 1 shows the average performance loss and average radius using a pragmatic approach to parameter selection.

*1) Example 1:*

$$A = \begin{bmatrix}1.2\end{bmatrix};\ B = \begin{bmatrix}1\end{bmatrix};\ C = \begin{bmatrix}1\end{bmatrix};$$
$$-0.2 \leq u_k \leq 0.2;\ \ -0.2 \leq \Delta u_k \leq 0.2;\ \ -5 \leq Cx_k \leq 5;$$

*2) Example 2:*

$$A = \begin{bmatrix}0.6 & -0.4\\ 1 & 1.4\end{bmatrix};\ B = \begin{bmatrix}0.2\\ 0.05\end{bmatrix};\ C = \begin{bmatrix}1 & -2.2\end{bmatrix}$$
$$-0.8 \leq u_k \leq 1.5;\ \ -0.4 \leq \Delta u_k \leq 0.4;\ \ -5 \leq Cx_k \leq 5;$$

*3) Example 3:*

$$A = \begin{bmatrix}1 & 0.1 & 0.3\\ 0 & 1 & 0.1\\ 0.4 & 0 & 1\end{bmatrix};\ B = \begin{bmatrix}0\\ 1\\ 1\end{bmatrix};$$
$$C = \begin{bmatrix}1 & 0 & 0\end{bmatrix}$$
$$-2 \leq u_k \leq 2;\ \ -1 \leq \Delta u_k \leq 1;\ \ -5 \leq Cx_k \leq 5;$$

### B. Optimal selection based on multiobjective solution

The multi-objective procedure was run for different sample directions, which were chosen uniformly on the unit circle. The parameters $n_c$ representing d.o.f. were chosen in the range $\{2, \ldots, 6\}$. The results for Examples 1, 2, 3 are shown in Figures 1-2, 3-5 and 6-7 respectively. In all plots, colour circles represent the parameter values $\alpha$.

Figures 1, 3 and 6 show the parallel coordinates and trade-off between average performance loss and average radius by varying $n_c$ with different parameter values $\alpha$. Parallel coordinates are used to plot all axes representing different objectives parallel to each other. Only the Laguerre function parameterisation simulation results are shown in the figures, but there are similar results for higher order parameterisations i.e. Kautz and 3rd order functions. These trade-off curves may be used as a selection criteria to choose best function parameterisation $\alpha$.

Figures 2, 4 and 7 show the trade-off curves between average performance loss and average radius for $n_c = 4$ with different parameterisation dynamics. It is observed that $3rd$ order function parameterisation improves both the feasible region without too much detriment to performance. An average performance loss and feasible region gain for example 2 is shown in figure 5 by varying $n_c \in \{2, \ldots, 6\}$ for a Laguerre parameterisation. One can see that best parameter selection may improve the feasible region and performance with a limited number of d.o.f. $n_c$.

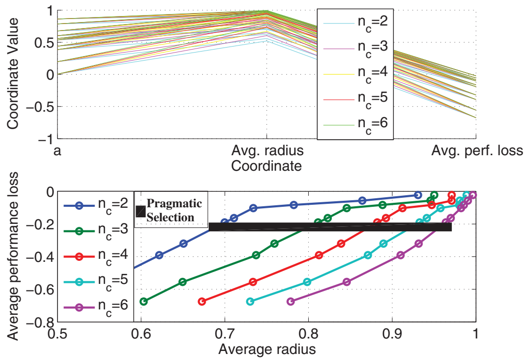


Fig. 1. Parallel coordinates and trade-off curves for Example 1

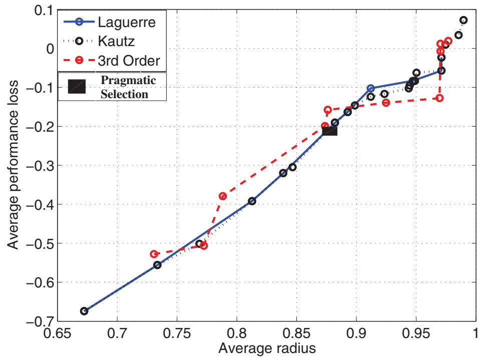


Fig. 2. Trade-off as a function of the parameter $\alpha$ : $n_c = 4$ for Example 1

### C. Sub optimal selection based on closed loop eigenvalues

Table I, II and III show the pragmatic selection of parameterisation dynamics using closed loop eigenvalues with an optimal gain $K$. It is clear from these that pragmatic choices are sub optimal. The selection of parameterisation is based on closed loop eigenvalues so parameterisation order is selected based on example dimensions.

Laguerre parameterisation is the only choice for example 1. With a sub optimal choice of $n_c = 2$ achieves 70 % of feasible set and 22 % of performance loss. Similarly in example 2, there are two parameterisation choices Laguerre and Kautz functions. The Laguerre function achieved 90 % of feasible set with 6 % performance loss. On the otherhand the Kautz function achieved 95 % of feasible set with 1 % performance loss with $n_c = 2$. In example 3, there are three parameterisation choices i.e. Laguerre, Kautz and 3rd order functions. All achieved more than 90 % global feasible set and maximum of 8 % of performance loss.

*Remark 4.1:* It is clear from figures 1-4, 6 and 7 that the pragmatic choice has solution in vicinity of an optimal pareto solution.

*Remark 4.2:* The optimisation structure of an approximate global optimal control is different to the parameterised one, so in few cases performance may even improve.

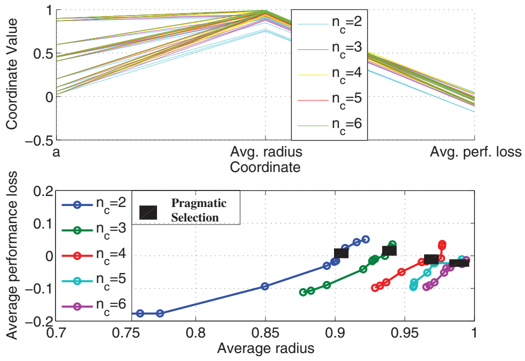


Fig. 3. Parallel coordinates and trade-off curves for Example 2

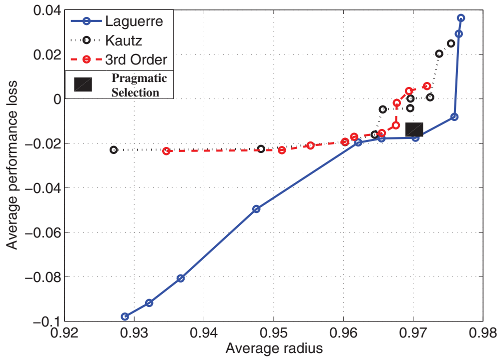


Fig. 4. Trade-off as a function of the parameter $\alpha$ : $n_c = 4$ for Example 2

## V. Conclusion

The main contribution of this paper was to present a systematic approach to selecting the best alternative parameterisation for MPC. Two techniques were discussed based on a multi-objective optimisation and a pragmatic approach using closed loop poles. Examples demonstrate that in many cases nominal closed loop poles are a good sub optimal choice to tune the parameterisation dynamics. However an optimal selection may be made using a multi-objective optimisation technique. In future work there is a need to further investigate in parallel issues such as: will these choices lead to an efficient online optimisation QP structure and/or low complexity multiparametric solution? Finally, it is an interesting future area to consider these choices within a robust scenario.

TABLE I
Pragmatic Selection for $\alpha$

| Example 1 | | | |
|---|---|---|---|
| Parameterisation | $n_c$ | Avg. perf. loss | Avg. radius |
| Laguerre<br>a=0.52 | 2 | 22% | 0.6904 |
| | 3 | 22% | 0.8011 |
| | 4 | 22% | 0.8720 |
| | 5 | 22% | 0.9237 |
| | 6 | 22% | 0.9578 |

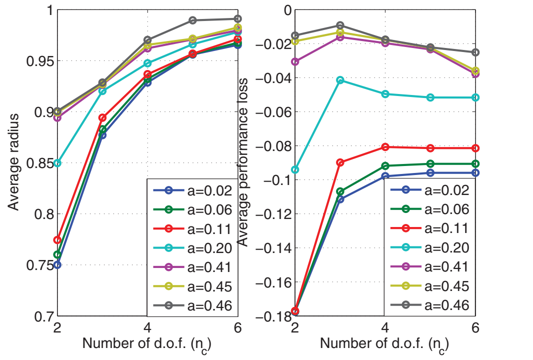


Fig. 5. Average radius and average performance loss for Example 2

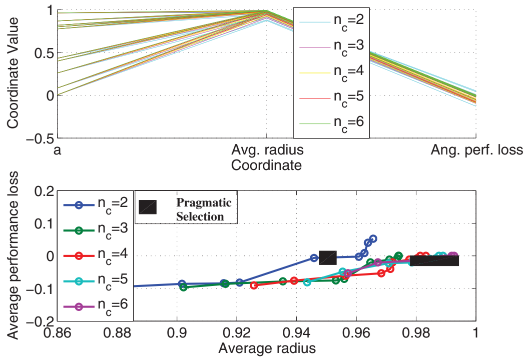


Fig. 6. Parallel coordinates and trade-off curves for Example 3

TABLE II
PRAGMATIC SELECTION FOR $\alpha$

| Example 2 | | | |
|---|---|---|---|
| Parameterisation | $n_c$ | Avg. perf. loss | Avg. radius |
| Laguerre<br>a=0.72 | 2 | 6 % | 0.9098 |
| | 3 | 3 % | 0.9548 |
| | 4 | 2% | 0.9683 |
| | 5 | 2 % | 0.9799 |
| | 6 | 2 % | 0.9896 |
| Kautz<br>(a,b)=0.72±0.37j | 2 | 1 % | 0.9468 |
| | 3 | 0 % | 0.9584 |
| | 4 | 0.5 % | 0.9829 |
| | 5 | 2 % | 0.9885 |
| | 6 | 2 % | 0.9907 |

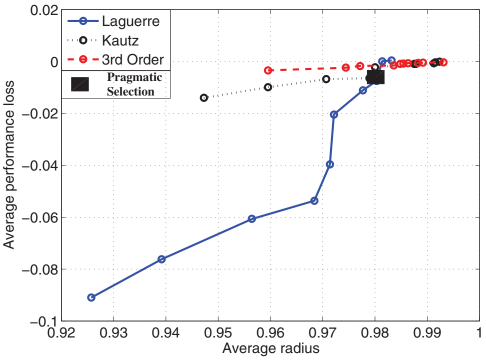


Fig. 7. Trade-off as a function of the parameter $\alpha$ : $n_c = 4$ for Example 3

TABLE III
PRAGMATIC SELECTION FOR $\alpha$

| Example 3 | | | |
|---|---|---|---|
| Parameterisation | $n_c$ | Avg. perf. loss | Avg. radius |
| Laguerre<br>a=0.36 | 2 | 0.2 % | 0.9113 |
| | 3 | 0.2% | 0.9494 |
| | 4 | 0.2% | 0.9658 |
| | 5 | 0.2% | 0.9790 |
| | 6 | 0.2% | 0.9873 |
| a=0.74 | 2 | 8 % | 0.9580 |
| | 3 | 6 % | 0.9731 |
| | 4 | 2 % | 0.9865 |
| | 5 | 0.7 % | 0.9906 |
| | 6 | 0.6 % | 0.9936 |
| Kautz<br>(a,b)=0.74±0.02j | 2 | 8 % | 0.9580 |
| | 3 | 8 % | 0.9600 |
| | 4 | 0.4% | 0.9764 |
| | 5 | 0.4% | 0.9831 |
| | 6 | 0.4% | 0.9954 |
| Generalised<br>(a,b,c)=0.36,0.74±0.02j | 2 | 2 % | 0.9113 |
| | 3 | 0 % | 0.9494 |
| | 4 | 0 % | 0.9689 |
| | 5 | 0 % | 0.9930 |
| | 6 | 0 % | 0.9934 |

## REFERENCES

[1] E. Camacho and C. Bordons, *Model predictive control*, Springer, 2003.

[2] D. Q. Mayne, J. B. Rawlings, C. Rao and P. Scokaret, "Constrained model predictive control", *Automatica*, vol. 36, 2000, pp 789-814.

[3] J. A. Rossiter, *Model-based predictive control, A practical approach*, CRC Press, London; 2003.

[4] E. Gilbert and K. Tan, "Linear sysetms with state and control constraints: The theory and application of maximal output admissible set", *IEEE Trans. on Automatic Control*, vol. 36, 1991, pp 1008-1020.

[5] P. Grieder, Efficient computation of feedback controllers for constrained systems, PhD Thesis, ETH, 2004

[6] L. Imsland and J. A. Rossiter, "Time varying terminal control", *Proc. 16th IFAc World Congress*, Prague, Czech Republic, 2005.

[7] L. Imsland, J. A. Rossiter, B. Pluymers and J. Suykens, "Robust triple mode MPC", *IJC*, vol. 81, 2005, pp 679-689.

[8] J. A. Rossiter, B. Kouvaritakis, M. Bacic, "Interpolation based computationally efficient predictive control", *IJC*, vol. 77, 2004, pp 290-301.

[9] J. A. Rossiter and L. Wang, "Exploiting Laguerre functions to improve the feasibility/performance compromise in MPC", *Proc. CDC.*, Cancun, Mexico, 2008.

[10] J. A. Rossiter, L. Wang and G. Valencia-Palomo, "Efficient algorithms for trading off feasibility and performance in predictive control", *IJC*, vol. 83, 2010, pp 789-797.

[11] B. Khan, J. A. Rossiter and G. Valencia-Palomo, "Exploiting Kautz functions to improve feasibility in MPC", *Proc. 18th IFAC World Congress*, Milan, Itlay, 2011.

[12] B. Khan and J. A. Rossiter, "Triple Mode MPC or Laguerre MPC: a comparison", *Proc. ACC*, San Francisco, California, USA, 2011.

[13] B. Khan and J. A. Rossiter, "Generalised Parameterisation for MPC", *Proc. The 13th IASTED Inter. Conf. On Intelligent Systems and Control (ISC)*, Cambridge, UK, 2011.

[14] B. Khan and J. A. Rossiter, "Computational Efficiency of Laguerre MPC using Active set Methods", *Proc. The 13th IASTED Inter. Conf. On Intelligent Systems and Control (ISC)*, Cambridge, UK, 2011.

[15] G. Valencia-Palomo, J. A. Rossiter, C. N. Colin, R. Gondhalekar and B. Khan, "Alternative parameterisations for predictive control: how and why?", *Proc. ACC*, San Francisco, USA, 2011.

[16] P. Scokaert and J. B. Rawlings, "Constrained linear quadratic regulation", *IEEE Trans. on Automatic Control*, vol. 43, 1998, pp 1163-1168.

[17] L. Wang, Model predictive control design and implementation using MATLAB, Springer, 2009.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Stability analysis and Control design of a Class of Event Based Control Systems

J. Jugo, University of Basque Country, Spain, josu.jugo@ehu.es
M. Eguiraun, ESS-Bilbao Consortium, Spain, meguiraun@essbilbao.org

***Abstract*—Event Based Control (EBC) provides a reduction of mean control rates, which is an important advantage in control systems, specially when network environment gets involved. For this reason, the study of design methodologies for EBC systems that met desired specifications regarding stability and performance issues are a valuable research field. This work presents a control design process applied to a class of EBC systems using LMIs, including stability issues on the light of Asynchronous Dynamical System theory. The application of the proposed methodology is presented by an example, showing good performance in simulation results.**

***Index Terms*—Event Based Control, Linear Matrix Inequality, Asynchronous Dynamical Systems, Control Application**

## I. INTRODUCTION

Event Based Control (EBC) has been considered complex and, in the past, the lack of a well established theory and design methods has limited its use to special cases. However, the interest in EBC control strategies is growing due to its advantages. In particular, its reduction on resource usage, which allows a reduction of the control rate [1], [2], [3], [4]. This characteristic has great importance in embedded systems, where life-time of devices is limited by batteries, or in networked environment, where the reduction on the usage of bandwidth is desirable. For instance, both requirements must be fulfilled in the case of wireless sensor networks.

Following the EBC approach, different alternatives have been developed. In [1], [2], improved system response with a reduction in the control rate is obtained using impulse control signals and observers. In [5], a piecewise control signal based scheme with level-triggered sampling is proposed. More recently in [4], the idea of sporadic control is introduced, where a minimum time difference between sampling times is proposed. A similar control scheme, which is experimentally implemented, presents an improvement of the system response in the presence of delay, jitter and noise, [6]. Moreover, PID controllers have been studied using event based sampling, [7]. The work in [8] presents a modified PID control structure that minimizes different problems related to event based sampling approach.

In addition, different schemes have been analyzed using state-space approach, showing stability properties, [9], and the existence of a lower bound for inter-events time, [10], for instance. This methodology has been extended to NCS systems in [11], including quantization effects. In [12], experimental results have been recently presented, evaluating an event based state-state approach. In [13], similar considerations are obtained for an output-based event-triggered control scheme using LMI techniques.

Therefore, different EBC schemes have been developed and evaluated. The advantages obtained applying such schemes encourage to develop new alternative approaches and specific design methodologies for such systems.

The objective of this work is to describe a design methodology useful for EBC systems, which guarantees stability and some design specifications and its validation under simulation test. The design procedure facilitates the improvement of the dynamics of the resulting EBC system, which will be experimentally implementable. The control scheme switches between different controllers following an event-triggered sampling scheme which is proposed in [6]. The stability of the closed-loop system is determined using a standard LMI methodology, [14], taking into account results derived from Asynchronous Dynamical Systems (ADS) theory. This theory is an interesting resource since it is applicable to Networked Control Systems (NCS), [15], [16].

The paper is organized as follows: first, the particular structure of the EBC scheme under study is described in Section II. In the next section, the analysis of the proposed EBC system is presented, based on the use of the ADS theory. In Section IV, the proposed approach for the design of the EBC controller is described. This process includes the LMI based stability analysis. An example of the proposed methodology is presented in Section V, where its performance is analyzed by means of simulations. Finally, conclusions end the paper.

## II. STATEMENT OF THE PROBLEM

Event Based Control is based on triggering control action only when a condition is met [2]. The main advantage of this control strategy is that the required number of control actions can be considerably reduced while maintaining the system under control.

Figure 1 presents the scheme of a basic EBC system, where a reference signal is explicitly included. The difference from a typical feedback control system is the event-based sampling technique used to perform the feedback.

Let us consider a linear, or linearized, system which can be described by the next equations:

$$\begin{aligned} \dot{x_p}(t) &= f(x,u) = Ax_p(t) + Bu(t) \\ y(t) &= g(x,u) = Cx_p(t) + Du(t) \end{aligned} \qquad (1)$$

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

Here, for the sake of simplicity, linear SISO systems with $D = 0$ are considered. Introducing a zero order hold (ZOH) in the input of the system, the control signal value $u_k$, which is generated by a digital controller, is maintained constant between consecutive control events. The closed-loop correction acts only when the event firing rule is met, e.g. when a selected signal level reaches a limit value defined during design stage. Hence, the continuous control signal, that is, the system's input signal, is

$$u(t) = h(t - t_k)u_k, \quad t \in [t_k, t_{k+1}), k = 0, \ldots, \infty \quad (2)$$

where $u_k$ is the event-triggered control signal and $h(t)$ is the Heaviside step function.

In this work, the event-based sampling of a continuous signal $z(t)$ is defined by the next expressions, with $k \geq 1$

$$\begin{array}{rcl} z_k & = & z(t_k) \\ d^{(k)}(t^*) & = & z(t_{k-1} + t^*) - z(t_{k-1}) \\ t_k & = & t_{k-1} + T_k \qquad (3) \\ T_k & = & \begin{cases} T_{min} & \text{if } \delta^{(k)}(t^*) \geq \delta_{limit} \\ & \quad \text{and } t^* \leq T_{min} \\ t^* & \text{if } \delta^{(k)}(t^*) \geq \delta_{limit} \\ & \text{and } t^* \in (T_{min}, T_{max}) \\ T_{max} & \text{if } \delta^{(k)}(t^*) < \delta_{limit} \\ & \quad \text{and } t^* = T_{max} \end{cases} \end{array}$$

where $t^*$ is the continuous time which is reset at every sampling time $t_k$, $t^* \in [0, T_k]$. The difference between the actual value of the signal and the last sampled value, $d^{(k)}(t^*)$, is a switched continuous function depending on the $k - th$ event. The parameter $\delta^{(k)}$ is the value which fires control action, continuously updated according to the following rule:

$$\delta^{(k)}(t^*) \;=\; K_p^s \left| d^{(k)}(t^*) \right| + K_i^s \int_0^{t^*} \left| d^{(k)}(t) \right| dt \quad (4)$$

This expression is reset at event time, i.e. $\delta^{(k)}(0) = 0$. Thus, the new sampling is triggered at $t_k = t_{k-1} + T_k$ time, when $\delta^{(k)}$ reaches a predefined limit $\delta_{limit}$ or when $t^*$ reaches the limit $T_{max}$. Equation 4 leads to a proportional-integral sampling scheme, where a minimum difference between sampling times $T_{min}$ (sporadic control approach [4]) and a maximum difference between sampling times $T_{max}$ are defined. Hence, $T_k \in [T_{min}, T_{max}]$.

In the same way, an integral action is added to the proportional sampling (or deadband sampling, [1]) to reduce the sticking phenomenon, [17]. Initially, when the reference varies, the proportional sampling is enough to guide the system towards the desired state. However, when it is close to the reference signal, no action will be triggered as long as $\delta^{(k)}(t^*)$ is maintained inside its deadband, so a non null error stands. In such case, the integral part of the event sampling rule will fire the control action after a time interval. This interval depends on the parameter $K_i^s$, defined in the event sampling policy. The choice of values of the parameters $\delta_{limit}$, $K_i^s$, and $K_p^s$ will have a direct impact on the performance of the system.

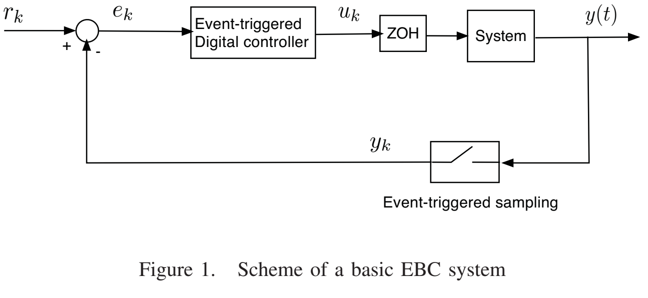


Figure 1. Scheme of a basic EBC system

**Considerations about $T_{max}$.** The existence of a maximum inter-sampling time can be considered that leads to unnecessary control actions. However, the definition of $T_{max}$ helps to minimize the sticking problem, [6]. Note that, in practice, this definition affects only in cases when the open-loop system is stable and $\delta^{(k)}$ is strictly zero, depending on the value of $\delta_{limit}$ and $K_i^s$. It is introduced to prioritize the minimization of the sticking phenomenon, with a possible penalty in the resource optimization which can be adjusted depending on the value of $T_{max}$. In addition, it facilitates the mathematical description proposed below.

For every event at $t_k$ derived from equations (3), the open-loop linear system (1) can be discretized, leading to the next representation,

$$\begin{array}{rcl} x_{k+1}^p & = & \phi_k x_k^p + \Gamma_k u_k \\ y_k & = & C x_k^p \qquad (5) \\ \phi_k & = & \phi(T_k) = e^{AT_k} \\ \Gamma_k & = & \Gamma(\;\; T_k) = \int_0^{Tk} e^{As} ds B \qquad (6) \end{array}$$

Due to the event-based sampling policy (3), $T_k$ can take any value in the interval $[T_{min}, T_{max}]$. Therefore, $\phi_k \in \mathcal{C}([T_{min}, T_{max}])$, that is, it is a continuous function with respect to variable $T_k$ in the interval $[T_{min}, T_{max}]$, that depends on matrix $A$. However, considering the nature of $t_k$, $\phi_k$ is a jump or switched function respect to the index $k$.

Let the structure of the controller be given by the next equations:

$$\begin{array}{rcl} x_{k+1}^c & = & A_k^c x_k^c + B_k^c e_k \\ u_k & = & C_k^c x_k^c + D_k^c e_k \qquad (7) \end{array}$$

where $e_k = r_k - y_k$. Defining the extended state-vector $\bar{x}_k = [x_k^{pT}, x_k^{cT}]^T$, the closed-loop system can be represented using the following equations:

$$\begin{array}{rcl} \bar{x}_{k+1} & = & \begin{bmatrix} \phi_k - \Gamma_k D_k^c C & \Gamma_k C_k^c \\ -B_k^c C & A_k^c \end{bmatrix} \bar{x}_k + \begin{bmatrix} \Gamma_k D_k^c \\ B_k^c \end{bmatrix} r_k \\ & = & \bar{A}_k \bar{x}_k + \bar{B}_k r_k \qquad (8) \end{array}$$

An EBC system following the sampling rule (3) presents the usual advantages shown in the literature for EBC systems, but with an improved behavior respect to the sticking problem, [6]. However, analysis and design techniques for this class of systems are needed, since the control structure could vary at each event.

In the next sections, there is presented an approach valid for determining the stability of the EBC systems using ADS and LMI theory and a design methodology adequate for fulfilling some specifications.

## III. Stability of EBC system under study

Some results from the ADS theory can be applied to the proposed EBC system, under several conditions. Relevant results from [15] are summarized to clarify the discussion. First, the definition of ADS systems:

**Definition 1.** An asynchronous dynamical system (ADS) with rate constraints on events is a tuple

$$\mathcal{A} = (\mathbb{R}_+, \{1, ..., N\}, \mathbb{R}^n, E, R, \mathcal{I}, F) \tag{9}$$

where $\mathbb{R}_+$ is time, $\{1, ..., N\}$ is the discrete state-space, $\mathbb{R}^n$ is the continuous state-space, $E$ is the set of events, $R = \{r_1, \dots, r_M\}$ is the set of event rates, $\mathcal{I} : 1, ..., N \to 2^E$ is the discrete state-event function, and $F$ is the set of continuous dynamical system functions. By definition, $\mathcal{I}(i) = \left\{E_{i_1}, E_{i_2}, \dots, E_{i_{M_i}}\right\}$ is the *i*th discrete state-event set, where $e_j^{(i)} \in E$ for $j = 1, ..., M_i$. An ADS has associated a discrete state $s(t)$ and a continuos state $x(t)$. $s(t) = i$ if and only if the events in $\mathcal{I}(i)$ have occurred and $\dot{x} = f_i(x)$. If the evolution of $x$ is given by a difference equation, then, the last equations is substituted by $x_{k+1} = f_i(x_k)$.

Now, the definition of the stability of such systems:

**Definition 2.** An ADS with continuous state dynamics $x(t)$ is exponentially stable if

$$\lim_{t\to\infty} e^{\alpha t} \left\|x(t)\right\| = 0,$$

for some $\alpha > 0$. If the dynamics is given by a discrete state $x_k$, the condition is

$$\lim_{k\to\infty} \alpha^k \left\|x(t)\right\| = 0$$

Focussing the discussion on the discrete case, the next theorem summarizes the main result from [15]: Consider an ADS system $\mathcal{A}$ with discrete state dynamics $x_k$ $(x_{k+1} = f_i(x_k))$, the ith discrete state-event set $\mathcal{I}(i) = \left\{E_{i_1}, E_{i_2}, \dots, E_{i_{M_i}}\right\}$ and a set of event rates $R = \{r_1, .., r_M\}$ in which $r_i$ satisfying $0 \le r_i \le 1$ is the rate of occurrence of event $E_i \in E$ over time. So, over any time period $[t, t+T]$ for large enough T, $r_iT$ is the total amount of time that $E_i$ has occurred.

In addition, suppose a Lyapunov-type function $V : \mathbb{R}^n \to \mathbb{R}_+$, which is continuously differentiable and

$$\beta_1 \left\|x\right\|^2 \le V(x) \le \beta_2 \left\|x\right\|^2 \tag{10}$$

where $\beta_{1,2} > 0$.

**Theorem 3.** *[15], if there exist scalars $\alpha_1, \alpha_2, ..., \alpha_M$ such that the system fulfills the condition*

$$\alpha_1^{r_1}\alpha_2^{r_2}\dots\alpha_M^{r_M} > \alpha > 1 \tag{11}$$

*and*

$$V(x_{k+1}) - V(x_k) \le (\alpha_{i_1}^{-2}\alpha_{i_2}^{-2}\dots\alpha_{i_{M_i}}^{-2} - 1)V(x_k) \tag{12}$$

*for i = 1,...,N, in which $i_j$ for $j = 1, ..., M_i$ correspond to the definition of $\mathcal{I}(i)$, the decay rate of the ADS is greater than $\alpha$. Then, the ADS system is exponentially stable.*

*Proof:* See, [15]. ■

In order to apply those results to the EBC scheme presented in Section II, some details have to be revised.

*Application of the ADS approach to event-based systems:* Considering the description in Section II and the Definition 1, the EBC system described by equations (8) and event-triggered mechanism (3) is similar to an ADS system. However, in this case, the events are defined by all the possible sampling instants in a continuous interval $[T_{min}, T_{max}]$. That is, $E = \{E_1, E_2, \dots, E_\infty\}$ is an infinite set. In addition, this fact difficulties the implementation of a real system using the sampling mechanism (3). To solve this problem, a finite set of possible sampling times is introduced:

$$\begin{aligned}\mathcal{T}_s \;=\; & [T_{min}, T_{min}+\Delta t_s, T_{min}+2\Delta t_s, \dots, \\ & \dots, T_{min}+n\Delta t_s = T_{max}],\ \Delta t_s = \frac{T_{max}-T_{min}}{n}\end{aligned} \tag{13}$$

with $n \in \mathbb{N}$. Now, consider the next approximation of the event-sampling approach defined in equations (3):

$$\begin{aligned}
z_k &= z(t_k) \\
e^{(k)}(t^*) &= z(t_{k-1}+t^*) - z(t_{k-1}) \\
t_k &= t_{k-1} + T_k \\
T_k^{ideal} &= \begin{cases} T_{min} & \text{if } \delta^{(k)}(t^*) \ge \delta_{limit} \\ & \quad \text{and } t^* \le T_{min} \\ t^* & \text{if } \delta^{(k)}(t^*) \ge \delta_{limit} \\ & \quad \text{and } t^* \in (T_{min}, T_{max}) \\ T_{max} & \text{if } \delta^{(k)}(t^*) < \delta_{limit} \\ & \quad \text{and } t^* = T_{max} \end{cases} \\
T_k &= \left\lfloor \frac{T_k^{ideal}}{\Delta t_s} \right\rfloor \Delta t_s
\end{aligned} \tag{14}$$

with $\delta^{(k)}$ defined by (4), $T_k \in \mathcal{T}_s$ and $T_k^{ideal}$ represents the sampling times derived from equations (3) (which is an ideal case). Note that the equations (3) are obtained from equations (14), when $n \to \infty$. The introduction of this new event-sampling approach is motivated by the implementation which is easier than in the approach given by (3), since can be based in the use of a regular periodic sampling time $\Delta t_s$. This idea is similar to the scheme proposed in [11].

Now, considering the event-sampling scheme (14), the proposed EBC system is an ADS system with discrete dynamics $\bar{x}_k$, being $E = \{E_1, E_2, \dots, E_n\}$. Note that the conclusions derived for this ADS system are valid for any $n$ and, then, the results should be valid for $n \to \infty$. Therefore, the result summarized in Corollary 4 can be concluded.

**Corollary 4.** *The stability of the EBC system described by equations* (8) *and event-triggered mechanism* (14)*, considering the definition presented in Section II, will be guarantied if:*

*- it exists a Lyapunov function $V(x_k) = x_k^T P x_k$ for some symmetric positive matrix $P > 0$ fulfilling*

$$V(x_{k+1}) - V(x_k) \leq (\alpha^{-2} - 1)V(x_k)$$

for some $\alpha > 1$ when $T_k \in [T_1, T_2] \subseteq [T_{min}, T_{max}]$.

*- the rate of event samplings in this interval is sufficiently large.*

*Proof:* By the definition of the EBC system, the finite set of events $E = \{E_1, E_2, \ldots, E_n\}$ and the related set of event rates $R = \{r_1, .., , r_n\}$ are well defined for a large period of time. Hence, there exist the scalars $\alpha_1, \alpha_2, ..., \alpha_n$ related with each n possible matrix $\bar{A}_i$ $i = 0, \ldots, n$ in equation (8). Under the conditions defined in this corollary, the system is exponentially stable for any arbitrary switch in the interval $T_k \in [T_1, T_2]$, [18]. Consider the set of $m$ events $E' \subseteq E$ related with $T_k \in [T_1, T_2]$. By Theorem 3, the set of $m$ systems $\bar{A}_j$ derived from $E'$ fulfills

$$\bar{A}_j^T P \bar{A}_j - P \leq (\alpha^{-2} - 1)P$$

for some $\alpha > 1$. In addition, if the rate of events in interval $T_k \in [T_1, T_2]$ is sufficiently large, the condition (11) in Theorem 3 will be fulfilled and the EBC system is stable. ■

A particular case derived of this Corollary is when the interval considered is $[T_{min}, T_{max}]$. In this case, the second condition is always fulfilled, but the obtaining of a controller fulfilling the first condition can be hard.

*Remark* 5. The event rate in a particular interval is directly related with the selection of $\delta_{limit}$ in each case.

*Remark* 6. The sampling mechanism can be enforced to assure the necessary rate in any interval.

The next section presents a design process valid for the proposed EBC system.

## IV. Controller design and LMI based stability analysis

Consider a continuous controller and its discretization for a sampling rate interval $T_k \in [T_1, T_2]$, under event-triggered sampling (14). This scheme leads to a discrete representation of the controller changing at every event $t_k$. If all the interval $[T_{min}, T_{max}]$ is considered, the n possible system matrices in equation 8for each $T_k \in \mathcal{T}_s$ are:

$$\bar{A}_i = \begin{bmatrix} \phi_i - \Gamma_i D_i^c C & \Gamma_i C_i^c \\ -B_i^c C & A_i^c \end{bmatrix} \quad i = 0, \ldots, n$$

where matrices $(A_i^c, B_i^c, C_i^c, D_i^c)$ are obtained by discretization, following equivalent equations to (6). Let us consider $V_k = x_k^T P x_k$ , for some $P > 0$. The system will be stable if the equations

$$\bar{A}_i^T P \bar{A}_i - P \leq -Q_i$$

are fulfilled for $i = 0, \ldots, n$ and $Q_i > 0$. However, by continuity in the discretization process, if the equations corresponding to $T_{min}$ and $T_{max}$ are fulfilled, the equivalent equation of every sampling time in interval $[T_{min}, T_{max}] \in \mathcal{T}_s$ will be fulfilled.

A more general scheme can be considered. The interval can be subdivided into $N_I$ intervals $[T_l, T_{l+1}] \subset \mathcal{T}_s$ with $l = 0, \ldots, N_I - 1$, covering the full interval $[T_{min}, T_{max}]$, being $T_0 = T_{min}$ and $T_N = T_{max}$. Then, controllers which stabilize the closed-loop for any $T_k \in [T_l, T_{l+1}]$ for each such interval $[T_l, T_{l+1}]$ are enough, if the conditions of Corollary 4 are fulfilled. Therefore, $N$ design problems must be solved, one for each time interval $[T_l, T_{l+1}]$ , which can be expressed as the next LMIs

$$\begin{aligned} \bar{A}_l^T P_l \bar{A}_l - P_l &\leq -Q_l \\ \bar{A}_{l+1}^T P_l \bar{A}_{l+1} - P_l &\leq -Q'_l \quad l = 0, \ldots, N_I - 1 \end{aligned} \tag{15}$$

with $P_l > 0$, $Q_l > 0$ and $Q'_l > 0$. The simplest way to fulfill Corollary 4 is to force a unique $P$ matrix, i.e. $P_l = P > 0$ $l = 0, \ldots, N_I - 1$,. This design problem can be expressed by mean of a set of LMIs to fulfill some specification.

However, the feasibility of the LMI problem considering an unique $P$ matrix can be difficult to satisfy in general. A more open problem is to relax this condition allowing multiple $P_l$ matrices. In such case, in order to achieve the second condition in Corollary 4, a supervisor assuring a minimum rate in each interval $[T_l, T_{l+1}]$, a dwell time, [18], is required. Note that several politics can be followed by this supervisor.

## V. Example of application

In this section, an example of application of the tools described in the previous section is presented. The procedure applied in this example is summarized as follows:

- Considering the plant to control and the basic specifications, the parameters for the event sampling approach (14) are selected: interval $[T_{min}, T_{max}]$, $n$, $\delta_{limit}$, $K_p^s$ and $K_i^s$.
- Selection of the $N_I$ intervals $[T_l, T_{l+1}] \subset \mathcal{T}_s$ covering the full interval $[T_{min}, T_{max}]$.
- Perform a controller design in the continuous domain applicable for each interval. A reasonable design constrain is that the bandwidth of the closed-loop system in each interval is directly related to the sampling times involved, that is, for higher sampling rates, larger bandwidth.
- Solve the LMI feasibility problem derived from equations (15). Note that those problems are solved in the discrete domain and the controller must be conveniently discretized, using a kind of emulation technique.

Resolving satisfactorily the last point, the EBC system can be implemented and performance tests done. Now, details about the example are presented.

### A. Controller Design

The plant used in current analysis consist of a DC motor connected to a rotational-to-translational motion converter. The continuous plant dynamics relating the input voltage and the position has been modeled approximately by the following transfer function:

$$G(s) = \frac{1}{s}\frac{K}{(\tau_m s + 1)} \tag{16}$$

where, $K = 6536\,mm\,vol^{-1}$ and $\tau_m = 1/4.3$. So, a state-space representation is given by the following equations:

$$A_p = \begin{pmatrix} 0 & 1 \\ 0 & -1/\tau_m \end{pmatrix} \; B_p = \begin{pmatrix} 0 \\ K/\tau_m \end{pmatrix} \tag{17}$$

$$C_p = \begin{pmatrix} 1 & 0 \end{pmatrix} \tag{18}$$

A standard PID controller can be enough for controlling this second order system. After design, the standard version must be rewritten as the equivalent representation in the state-space.

For a first example of application, the sampling interval defined by the event-based sampling scheme (14) is split in three zones, $N_I = 3$:

$$\overbrace{[T_{min}, T_1]}^{controller\,1} \bigcup \overbrace{[T_1, T_2]}^{controller\,2} \bigcup \overbrace{[T_2, T_{max}]}^{controller\,3} \tag{19}$$

being $T_{min} = 0.5msec.$ , $T_1 = 1msec.$, $T_2 = 3msec.$, $T_{max} = 50msec.$. In addition, the other parameters needed in (14) are $K_p^s = 0.1$, $\delta_{limit} = 0.01$ and $K_i^s = 100$. The last one has been selected as reference, since an analysis of the effect of this parameter in the behavior of the EBC system has been done. This is presented below.

The EBC sampling scheme leads to a changing variable time between consecutive control actions, depending on the speed of the changes on the error signal. Therefore, three different controllers are proposed for different sampling times, corresponding to the three intervals in (19). For high sampling rates, faster controller is preferred, with a higher bandwidth. On the contrary, for low sampling rates, a control system with a lower bandwidth is enough for leading the system to the desired state.

The overall methodology for the full controller design procedure is presented in Figure (2), starting from the model of the system to be controlled. After obtaining a valid controller for each different region, the LMI feasibility problem is analyzed. The implementation consist on the use of event sampling strategy (14), followed by a decision algorithm which chooses the correct controller for each interval. Depending on the different sampling times $T_k$, appropriate discrete controller is used when applying the event sampling policy.

For each region, the next controller parameters are obtained, requiring in the design process different closed-loop bandwidths:

$$\begin{array}{lll} K_p^{(1)} = 0.0016737 & K_i^{(1)} = 0.0002193 & K_d^{(1)} = 0.0010633 \\ K_p^{(2)} = 0.0072328 & K_i^{(2)} = 0.0011645 & K_d^{(2)} = 0.0095654 \\ K_p^{(3)} = 0.0163627 & K_i^{(3)} = 0.0020823 & K_d^{(3)} = 0.0325810 \end{array} \tag{20}$$

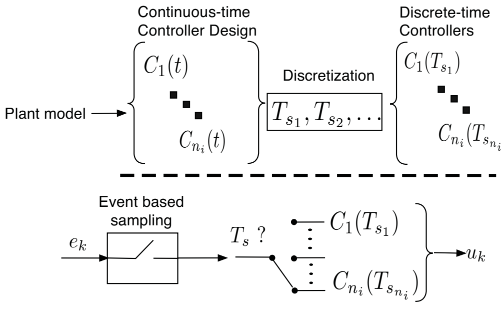


Figure 2. Proposed controller design methodology for the presented Event Based control.

Since the presented event based methodology implies the use of discrete domain controllers, the PID is approximated using the backward approximation of the derivate as

$$G_{pid}(z) = K_p + K_d \frac{(1 - z^{-1})}{T_k(1 - \alpha z^{-1})} + K_i \frac{T_k}{(1 - z^{-1})}$$

which can be represented in the state-space

$$\begin{aligned} A_c &= \begin{pmatrix} 1+\alpha & -\alpha \\ 1 & 0 \end{pmatrix} \quad B_c = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\ C_c &= \begin{pmatrix} K_1^c & K_2^c \end{pmatrix} \quad D_c = K_p + \frac{Kd}{T_k}(1-\alpha) + K_i T_k \\ K_1^c &= -\left( (1+\alpha)K_p + 2(1-\alpha)\frac{Kd}{T_k} + \alpha K_i T_k \right) \\ K_2^c &= \alpha K_p + (1-\alpha) K_i T_k \end{aligned}$$

The parameter $\alpha$ is introduced to solve possible numerical issues. Now, three LMI feasibility problems following (15) can be proposed. Those equations admits some variations.

Indeed, the overall control design can be proposed as a LMI, [19]. In fact, the control representation in the state-space has been selected to facilitate this possibility. In any case, the controllers 20 fulfill the first condition of Corollary 4.

The switching between the different controllers is facilitated for the use of the same structure in all the cases. When an event is received, $T_k$ is known and the appropriated controller gains can be applied. A supervisor acts in parallel in order to prevent continuous switching between different control intervals (second condition in Corollary 4). A minimum dwell time can be established if necessary.

For a first implementation of the proposed control approach, the choice of designing just three different regions is considered enough for comparison purposes. In addition, obtained results show better performance comparing to a single controller solution. The values of parameters $T_{min}$, $T_{max}$ and $T_i$ $(i = 1, 2)$ comes from the fact that the obtained discrete controllers must be stable, but also the performance of the system must be maintained inside reasonable limits.

### B. Simulations results

In order to perform a correct study under a simulation test, the hybrid nature of the overall system must be adequately described. In this case, the tool used is Ptolemy II, a Java based software system valid for the use of heterogeneous mixtures of models of computation, [20]. This software allows the definition of components for the event based scheme proposed in this paper, including switching between different controllers, in addition to the continuous plant.

Using this simulation framework, the performance of the system when using a multiple EBC controller scheme has been compared to a single EBC controller solution. Both solutions use the same parameter values when applying (14).

The results of the simulations show the evolution of the *Mean Sampling Time*, the *Mean Error*, defined as $\langle |r-y| \rangle$, and the *Event Ratio* (Event ratio is defined as the mean number of produced control events for a given time interval). As long as event occurrences increases, this value will also increase.

While most parameters are maintained constant, the integral part of the event triggering level has been varied in the range $K_i^s \in [50, 200]$ (Figures 3, 4 and 5).

As can be expected, increasing $K_i^s$, which rises the effect of the integral part of the event sampling, the number of event occurrences increases. Thus, more control actions are generated, reducing the system error.

The multi-controller case presents higher mean sampling time and lower error and event ratio. In addition, as Figures 6 and 7 show, the multi-controller scheme presents an improved system behavior, for same event sampling parameters ($K_p^s = 0.1$, $\delta_{limit} = 0.01$ and $K_i^s = 100$). The oscillations are reduced by switching to a more suitable controller for higher sampling times. Therefore, the use of multiple controllers allows a better definition of the closed-loop dynamics, adding more degree of freedom.

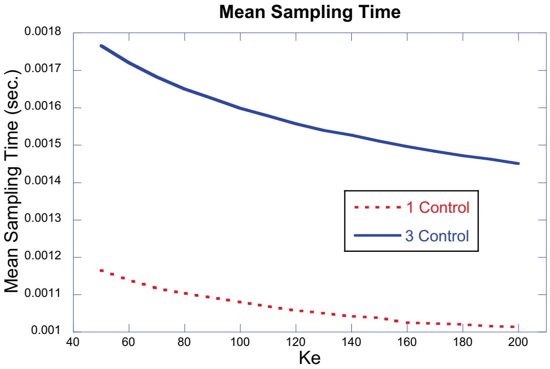


Figure 3. Mean Sampling Time vs parameter $K_i^s$, for both control schemes

A remarkable fact is that, in both cases, those results are qualitatively similar to the ones presented in previous works (e.g. [4]). They show a reduction in the needed resources, specially mean sampling time, without a notorious system performance reduction comparing to a pure periodic case.

Increasing the number of controllers, i. e., increasing the number of ranges in (19), the performance of the system is improved, but it also requires a more complex implementation. For the system under control, presented results are considered good enough and the addition of extra regions does not give significant advantages.

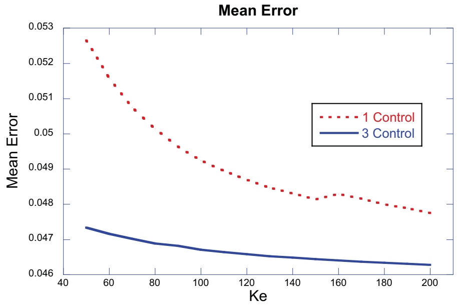


Figure 4. Mean Error vs parameter $K_i^s$, for both control schemes

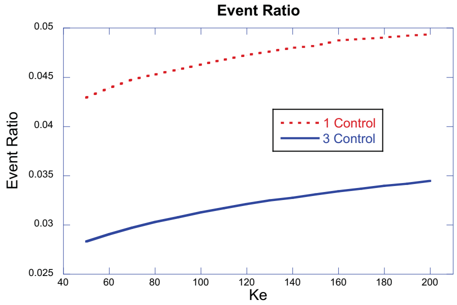


Figure 5. Event Ratio vs parameter $K_i^s$, for both control schemes

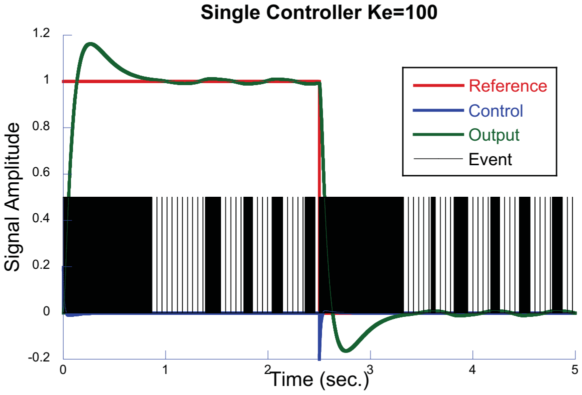


Figure 6. System output and event occurrences for single controller

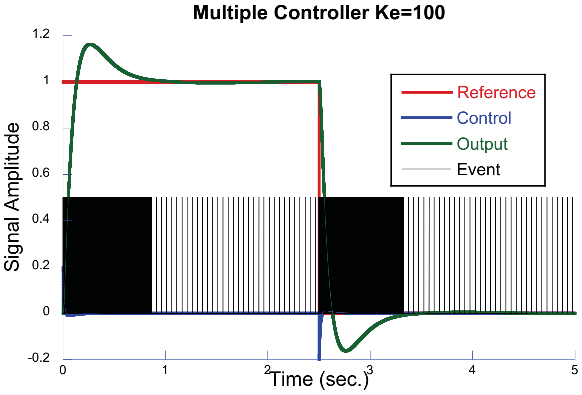


Figure 7. System output and event occurrences for multiple controller

## VI. CONCLUSIONS

In this work, a design methodology for a class of event based controllers has been presented. The stability of the closed-loop system is guaranteed by applying ADS theory and solving several LMI problems Moreover, the proposed scheme allows the use of different controllers depending on the event rate, leading to a better closed-loop dynamics definition. So, desired dynamics specifications can be met more efficiently.

The principal advantage of the resulting closed-loop system is that the EBC control structure lets a reduced control rate. This characteristic makes the proposed scheme very adequate for networked environments, for saving energy and for reducing actuators' stress. In fact, this methodology has been chosen since LMI and ADS theory have been successfully applied to NCS systems in the literature. Therefore, the extension of the approach for EBC systems in a networked environment will follow a natural way.

In addition, the simulations in the example show that an improved system performance can be obtained by designing different controllers for several sampling time intervals, as opposed to use a single controller independent of the sampling period of each control action. The reduction of the number of control actions, comparing with a periodic sampling scheme, is still guaranteed. However, the improvement of the dynamics obtained with multiple EBC controllers requires, in general, an increased event rate, comparing with the single EBC controller case.

Future work will consist in the extension of the presented control strategy to a networked environment.

## ACKNOWLEDGEMENTS

The authors are very grateful to the Basque Government for the financial support to our research group and to ESS Bilbao Consortium for supporting this work.

## REFERENCES

[1] P. Otanez, J. Moyne, and D. Tilbury, "Using deadbands to reduce communication in networked control systems," in *American Control Conference, 2002. Proceedings of the 2002*, vol. 4, pp. 3015 – 3020 vol.4, 2002.

[2] K. J. Astrom, *Analysis and design of nonlinear control systems*, ch. Event based control. Springer Verlag, 2007.

[3] X. Wang and M. Lemmon, "Event design in event-triggered feedback control systems," in *Proceedings of IEEE Conference on Decision and Control 2008*, vol. 1, December 2008.

[4] T. Henningsson, E. Johannesson, and A. Cervin, "Sporadic event-based control of first-order linear stochastic systems," *Automatica*, vol. 44, no. 11, pp. 2890 – 2895, 2008.

[5] M. Rabi and J. S. Baras, "Level-triggered control of a scalar linear system," in *Proceedings of 15th IEEE Mediterranean Conference on Control and Automation*, vol. 1, (Athens, Greece), June 27-29 2007.

[6] J. Jugo and M. Eguiraun, "Experimental implementation of a networked input-output sporadic control system," in *Proceedings of the 2010 IEEE International Conference on Control Applications*, pp. 1779 – 1784, September 2010.

[7] S. Durand and N. marchand, "Further results on event-based pid controller," in *Proceedings of European Control Conference 2009*, vol. 1, (Budapest), 2009.

[8] V. Vasyutynskyy and K. Kabitzsch, "Deadband sampling in pid control," in *Industrial Informatics, 2007 5th IEEE International Conference on*, vol. 1, pp. 45 –50, 2007.

[9] J. H. S. W. P. M. H. Heemels and P. P. J. van den Bosch, "Analysis of event-driven controllers for linear systems," *Int. Journal of Control*, vol. 81, no. 4, pp. 571–590, 2008.

[10] P. Tabuada, "Event-triggered real-time scheduling of stabilizing control tasks," *IEEE Transactions on Automatic Control*, vol. 52, no. 9, pp. 1680–1685, 2008.

[11] D. Y. Songlin Hu, "Event-triggered control design of linear networked systems with quantizations," *ISA Transactions*, vol. 51, pp. 153–162, January 2012.

[12] D. Lehmann and J. Lunze, "Extension and experimental evaluation of an event-based state-feedback approach," *Control Engineering Practice*, vol. 19, no. 2, pp. 101 – 112, 2011.

[13] M. Donkers and W. Heemels, "Output-based event-triggered control with guaranteed $l_\infty$-gain and improved event-triggering," in *Decision and Control, 2010., Proceedings of the 49th IEEE Conference on*, pp. 3246–51, abendua 2010.

[14] S. Boyd, L. Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*. SIAM Society for Industrial and Applied Mathematics, 1994.

[15] A. Hassibi, S. Boyd, and J. How, "Control of asynchronous dynamical systems with rate constraints on events," in *Decision and Control, 1999. Proceedings of the 38th IEEE Conference on*, vol. 2, pp. 1345 –1351 vol.2, 1999.

[16] W. Zhang, M. S. Branicky, and S. M. Phillips, "Stability of networked control systems," *IEEE Control Systems Magazine*, vol. 21, pp. 84–99, 2001.

[17] M. Á. G. José Sánchez and S. Dormido, "On the application of different event-based sampling strategies to the control of a simple industrial process," *Sensors*, vol. 9, pp. 6795–6818, 2009.

[18] D. Liberzon, *Switching in Systems and Control*. Birkhauser, 2003.

[19] P. G. Carsten Scherer and M. Chilali, "Multiobjective output-feedback control via lmi optimization," *IEEE Transactions on Automatic Control*, vol. 42, pp. 896–911, July 1997.

[20] *Ptolemy II, http://ptolemy.eecs.berkeley.edu/ptolemyII/index.htm.*

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Sensitivity of Optimal Operation of an Activated Sludge Process Model

Antonio Araujo*, Simone Gallani*, Michela Mulas† and Sigurd Skogestad‡
*Dept. of Chemical Engineering
Federal University of Campina Grande, 58429-140 Campina Grande, Paraiba, Brazil
Email: antonio@deq.ufcg.edu.br and simonegallani@hotmail.com
†Dept. of Civil and Environmental Engineering
Aalto University, P. O. Box 15200, FI-00076 Aalto, Finland
Email: michela.mulas@aalto.fi
‡Dept. of Chemical Engineering
Norwegian University of Science and Technology, N-7491 Trondheim, Norway
Email: skoge@ntnu.no

***Abstract*—This paper describes a systematic sensitivity analysis of optimal operation conducted on an activated sludge process model based on the test-bed Benchmark Simulation Model No. 1 (BSM1). The objective is to search for an operational structure that leads to optimal economic operation, while promptly rejecting disturbances at lower layers in the control hierarchy avoiding thus violation of the more important regulation constraints on effluent discharge. We start by optimizing a steady-state nonlinear model of the process. The resulting active constraints must be chosen as economic controlled variables. These are the effluent ammonia from the bioreactors and the final effluent total suspended solids at their respective upper limits, as well as the internal recycle flow rate at its lower bound. The remaining degrees of freedom need to be fulfilled, and we use several local (linear) sensitivity methods to find a set of unconstrained controlled variables that minimizes the loss between actual and optimal operation; particularly we choose to control linear combinations of readily available measurements so to minimize the effect of disturbances and implementation errors.**

## I. INTRODUCTION

Much to the authors' surprise, optimization of wastewater treatment plants has not received much attention in the WWTP research community given the small number of contributions found in the literature. Only few articles discuss the subject from a heuristic economic point of view[1], [2], [3] to formal optimization using an explicit mathematical model of the process[4], [5], [6] for optimal design and operation. However, none of the publications define an optimal operation policy from a systematic prism. Araujo *et al.*[7] applied a systematic procedure for control structure design of an activated sludge process in which optimization for various operational conditions were carried out on a mathematical model of the process.

In this communication a systematic sensitivity analysis of optimal operation of an activated sludge process model based on the Benchmark Simulation Model No. 1 (BSM1) [8] is conducted. It must be clear that all analysis, and hence all conclusions, from this work are based on the underlying mathematical model of the process, and should not be considered as definite guidelines for actual plant operation since the mathematical model may not be able to reproduce many operational situations. However, the results can be used in practice as general rules-of-thumb to be tested in actual wastewater treatment plants of the kind discussed here.

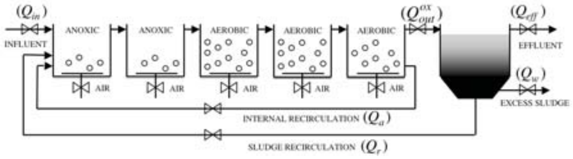


Fig. 1. Schematic representation of the BSM1 activated sludge process.

## II. PROCESS DESCRIPTION

The BSM1 [8] represents a fully defined protocol that characterizes the process including the plant layout, influent loads, modeling and test procedures as well as evaluation criteria. Figure 1 shows a schematic of the process consisting of a bioreaction section divided into five compartments, which can be anoxic or aerobic, and a secondary settling device. In order to maintain the microbiological population, sludge from the settler is re-circulated into the reaction section (returned activated sludge, $Q_r$). Also, part of the mixed liquor leaving the last reactor can be recycled to the inlet of the bioreactor (internal recycle, $Q_a$) to enhance nitrogen removal. Moreover, excess sludge at a rate $Q_w$ is continuously withdrawn from the settler underflow.

From a modeling point of view, the original BSM1 is based on two widespread accepted process models: the celebrated Activated Sludge Model No.1 (ASM1) [9] used to model the biological process, and a non-reactive Takacs one dimensional layer model for the settling process [10], [11]. The full model equations as well as kinetic and stoichiometric parameters are given within the benchmark description [8]. In addition, influent data are provided in terms of flow rates and ASM1 state variables over a period of 14 days with 15 minutes sampling time.

Each reactor is modeled as a perfectly-mixed, constant-volume tank within which complex biological reactions give

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

rise to component mass balance equations, generating a system of (coupled) ordinary differential equations. The ASM1 is a well establish and reliable model widely used among WWT modelers, and further discussion on its known capabilities of reproducing with fidelity the behavior of the reaction section of an activated sludge process can be found in the vast wastewater literature.

However, the same cannot be said about the settler model, though, for the reason that these units display very complex mechanisms that are not still fully understood [12]. Nevertheless, much progress has been made towards building a physically sound model for the secondary clarifier based on the theory of partial differential equation applied to conservation law with discontinuous fluxes [13], [14], [15], [16]. While these more meaningfully grounded mathematical models satisfying fundamental physical properties [17] still have not found widespread application in the WWT field, it is a commonplace to resort to approximate models of the settler, and the one due to Takacs [10], [11] is the most widely used representation of the secondary settler in published studies and commercial software environments. Some authors [18], [19], [20], however, pointed out many setbacks related to this model, among which is the fact that the number of discretization layers is not in agreement with numerical convergence and without distinguishing model formulation and numerical solution, but instead it is used solely as a model parameter in order to match experimental observations [21]. Numerical simulations have showed [17] the failure of Takacs' model to represent the complex behavior of secondary settlers under certain conditions, and this has led researchers to switch to more reliable physically meaningful sedimentation models. One such development is described by Diehl [13] who formulated and analyzed dynamically the settler model based on the one-dimensional scalar mass conservation law (1)

$$\frac{\partial X(z,t)}{\partial t} + \frac{\partial}{\partial z}(F(X(z,t),z)) = s(t)\delta(z) \tag{1}$$

where $X$ is the flocculated solids concentration, $\delta$ is the Dirac measure, $s$ is a source, and $F$ is a flux function, which is discontinuous at three points in the space coordinate $z$, namely at the inlet and the two outlets. Details are fully given in the cited references. We here are interested in the sensitivity of the static optimum of the settler coupled with the biological reaction section, and the steady-state solutions of the above equation provide the basis for our analysis.

A simple analysis of the model described in [15], shows that the steady-state model of the settler that must hold for optimization purposes is given by (2)

$$\begin{aligned} X_e &= \frac{s - f(X_M)}{q_e} \\ X_u &= \frac{f(X_M)}{q_u} \\ X_M &= M(q_u) \\ X_f &\in (X_m, X_M) \end{aligned} \tag{2}$$

where $X_e$ and $X_u$ are the solids concentration in the effluent and in the wastage streams, respectively; $s$ is the feed flux given by $s = q_f X_f$, where $q_f$ is the feed flux and $X_f$ is the solids concentration in the feed; $f(X)$ is a flux function given by $f(X) = Xv_s(X) + q_u X$, where $v_s(X)$ is the settling velocity law given by the double exponential equation [10]; $X_M$ is a minimizer of $f(X)$; $X_m$ is a value strictly less than $X_M$ satisfying $f(X_m) = f(X_M)$; $q_e$ and $q_u$ are the effluent and wastage fluxes, respectively; $M$ is a function that computes the local minimizer of $f(X_M)$. In addition, we can also calculate the steady-state concentration of suspended solids in the clarification ($X_{cl}$) and thickening zones ($X_{th}$) as in (3) [15]

$$\begin{aligned} g(X_{cl}) + s &= f(X_M) \\ X_{th} &= X_M \end{aligned} \tag{3}$$

where $g(X_{cl}) = X_{cl}v_s(X_{cl}) - q_e X_{cl}$.

Note that, although in this communication a nonreactive settler is considered, we here follow [14] and treat the dissolved oxygen in the settler in a special way. We assume that the oxygen is consumed within the settler and, consequently, the oxygen concentration at the settler's outlets is set to zero, which is indeed a realist assumption. This results in a more conservative computation of the oxygen demand in the reaction section.

## III. Systematic sensitivity analysis methodology

The methodology is mainly based on the first 4 steps, known as "top-down analysis", of the more general procedure described in [22], where economic variable selection is the key issue. The analysis conducted is of local nature, i.e., we use linearized models of the process to develop the methodology.

In this communication we use optimal measurement combinations [23] for unconstrained variable selection, i.e. the ones left after choosing the active constraints as "primary" economic variables. The basic idea is to select combinations $c$ of the measurements $y$ such that $c = Hy$, where $H$ is a (static) selection matrix. To determine $H$, two approaches are developed based on a linearized model of the process and a second-order Taylor series expansion of the cost function used for optimization; two sources of uncertainty are assumed which are represented by (1) external disturbances ($d$) and (2) implementation (measurement) errors ($n$). The first of two approaches combines these uncertainties in one single scaled vector to minimize the worst case economic loss ($L$), defined as the difference between actual operation (with a given control structure in place) and operation under optimal control. The second approach is to first minimize the loss with respect to external disturbances, and then, if there are still available measurements, minimize the loss with respect to implementation (measurement) errors.

The steps to be followed are:

**1. Define operational objectives**: We first quantify the operational objectives in terms of a scalar cost function (here denoted $J$) as given in (4) that should be minimized

$$J = \textit{cost of feed + cost of utilities (energy)} \\ -\textit{revenue from valuable products} \tag{4}$$

Constraints can then be added to the process as inequality equations of the form $g \leq 0$.

**2. Determine the steady-state optimal operation**: Using a steady-state model of the process, identify degrees of freedom and expected disturbances, and perform optimizations to assess sensitivity for the expected disturbances.

Usually, the economics of the plant are primarily determined by the (pseudo) steady-state behavior [24], so the steady-state degrees of freedom ($u_0$) are usually the same as the economic degrees of freedom. Which variables to include in the set $u_0$ is immaterial, as long as they make up an independent set. The important disturbances ($d$) and their expected range for future operation must then be identified. These are generally related to feed rate and feed composition, as well as external variables such as temperature and pressure of the surroundings. We should also include as disturbances possible changes in specifications and active constraints (such as product specifications or capacity constraints) and changes in parameters (such as equilibrium constants, rate constants and efficiencies). Finally, we need to include as disturbances the expected changes in prices of products, feeds and energy.

In order to achieve near optimal operation without the need to re-optimize the process when disturbances occur, one needs to minimize the loss in (5)

$$L = J_0(c, d) - J_0(c^{opt}(d), d) \geq 0 \tag{5}$$

where $J_0(c, d)$ is the value of the cost for a chosen set of constant setpoint variables $c$ that fulfill all remaining degrees of freedom and $J_0(c^{opt}(d), d)$ is the value of the cost after re-optimization. Clearly, the loss in (5) depends on the objective function as well as on the measurements through $c$, since $c$ is a function of the available $y$. We then need to learn about the sensitivity to disturbances not only of the cost function, but also of the measurements.

At last, the steady-state optimization problem can be formulated as in (6)

$$\begin{aligned} &\min_{u_0} J_0(x, u_0, d) \\ &\text{subject to} \\ &\text{Model equations: } f(x, u_0, d) = 0 \\ &\text{Operational constraints: } g(x, u_0, d) \leq 0 \end{aligned} \tag{6}$$

where $x$ are internal variables (states). In $f(x, u_0, d) = 0$ possible operational equality constraints (like a given feed flow) is also included. The main objective is to determine the optimal nominal operating condition to be used in the variable selection step.

**3. Select "economic" (primary) controlled variables**: In this step, the issue is the implementation of the optimal operation point found in the previous step in a robust and, most importantly, simple manner. We need to identify as many economic controlled variables ($c$) as there are economic degrees of freedom ($u_0$). For economic optimal operation, active constraints must be selected [25], which in turn consumes part of the degrees of freedom ($u'$). For the remaining degrees of freedom $u$ (with $n_u = n_{u_0} - n_{u'}$), we select variables for which close-to-optimal operation is achieved with constant nominal setpoints, even when there are disturbances [26]. Because our considerations in this communication are of local nature, we assume that the set of active constraints does not change with changing disturbances, and we consider the problem in reduced space in terms of the remaining unconstrained degrees of freedom $u$, which can be expressed as in (7) [27]

$$\begin{aligned} &\min_{u} J_0(x, u, d) \\ &\text{subject to} \\ &\text{Model equations: } f(x, u, d) = 0 \\ &\text{Active constraints: } g_{active}(x, u, d) = 0 \end{aligned} \tag{7}$$

where we consider as active constraints a subset $g_{active}(x, u, d)$ of $g(x, u_0, d)$ for which optimal values are always at bounds for all disturbances. By eliminating the states using the equality constraints in (7), the unconstrained optimization problem can be expressed simply as in (8)

$$\min_{u} J(u, d) \tag{8}$$

Ensuring active constraint operation consumes part of the degrees of freedom for optimization. The remaining degrees of freedom need to be fulfilled, and we select variables such that when kept at optimal setpoints leads to near optimal economic operation despite of disturbances, i.e., the deviation (loss $L$ in (5)) from re-optimization as a function of disturbances should be small. The optimal setpoints of $c$ are then determined from the optimization at the nominal operating point. This is the celebrated "self-optimizing" control technology [26]. It can be shown that the loss in (5) can be expressed in its worst case form ($L_{wc}$) as in (9) [23]

$$L_{wc} = \max_{\left\| \begin{bmatrix} d' \\ n^{y\prime} \end{bmatrix} \right\|_2 \leq 1} L = \frac{1}{2}\bar{\sigma}^2(M) \tag{9}$$

where $d'$ and $n^{y\prime}$ are the scaled disturbance and measurement error variables related by $d = W_d d'$ and $n^y = W_{n^y} n^{y\prime}$ ($W_d$ and $W_{n^y}$ are scaling matrices), and $M = -M_n H \tilde{F}$, where $\tilde{F} = [F W_d W_{n^y}]$ being $F = \frac{\partial y^{opt}}{\partial d}$ the optimal measurement ($y^{opt}$) sensitivity with respect to the disturbances and $M_n$ any nonsingular $n_u \times n_u$ matrix. In other words, we need to find $H$ that minimizes $\bar{\sigma}(M)$, i.e., $H = arg \min_H \bar{\sigma}(M)$.There are basically two approaches to solve for this minimization problem. The first one solves the minimization at once by combining disturbances and measurement errors, and an explicit formula for $H$ is given by (10) [23]

$$H^T = (\tilde{F}\tilde{F}^T)^{-1}G^y(G^{yT}(\tilde{F}\tilde{F}^T)^{-1}G^y)^{-1}J_{uu}^{1/2} \quad (10)$$

where $G^y$ is the static model of the process from the unconstrained inputs $u$ to the measurements and $J_{uu} = \left(\frac{\partial^2 J}{\partial u^2}\right)_{u^{opt}}$ is the Hessian of $J$ with respect to $u$ evaluated at $u^{opt}$ ($u^{opt}$ is the optimal value of the manipulated variables).

The second approach, called the extended nullspace method [23], solves the problem by first minimizing the loss with respect to disturbances, and then, if there are still enough measurements left, minimize the loss with respect to measurement errors. It can be shown that the explicit expression for $H$ in this case is given in (11)

$$H = M_n^{-1}\tilde{J}(W_{n^y}^{-1}\tilde{G^y})^{\dagger}W_{n^y}^{-1} \quad (11)$$

where $\tilde{J} = [J_{uu}^{1/2} \quad J_{uu}^{1/2}J_{uu}^{-1}J_{ud}]$ ($J_{ud} = \frac{\partial^2 J}{\partial u \partial d}$) and $\tilde{G^y} = [G^y \quad G_d^y]$ ($G_d^y$ is the static model from disturbances $d$ to $y$).

There are four cases where (11) can be applied:

3a. "Just-enough" measurements, i.e., $n_y = n_u + n_d$. Here, the expression for $H$ becomes (12)

$$H = M_n^{-1}\tilde{J}(\tilde{G^y})^{-1} \quad (12)$$

which is the same as having $H$ in the left null space of $F$, i.e., $H \in N(F^T)$.

3b. Extra measurements (select just enough measurements), i.e., $n_y > n_u + n_d$, and we want to select a subset of the measurements $y$ such that $n_y = n_u + n_d$. The solution is to find such a subset that maximizes $\underline{\sigma}(\tilde{G^y})$ using, e.g. existing efficient branch-and-bound algorithms [28]. The resulting $\tilde{G^y}$ is then used to compute $H$ in (12).

3c. Extra measurements (use all available measurements), i.e., $n_y > n_u + n_d$. $H$ is calculated using (11), where † denotes the left inverse, calculated as $A^{\dagger} = (A^TA)^{-1}A^T$ for any given matrix $A$.

3d. "Too few" measurements, i.e., $n_y < n_u + n_d$. In this case, the optimal $H$ in (11) is not affected by the noise weight and therefore becomes

$$H = M_n^{-1}\tilde{J}(\tilde{G^y})^{\dagger} \quad (13)$$

where † denotes the right inverse, calculated as $A^{\dagger} = A^T(AA^T)^{-1}$.

The above procedure boils down to selecting suitable candidate measurements, i.e. identify $n_y$ vis-a-vis $n_u + n_d$, and find that linear combination (matrix $H$) of all, or a given subset, which results in the smallest loss among all possible solutions. One big hurdle to be surmounted is the numerical calculation of $J_{uu}$ and $J_{ud}$. For some ill-posed problems, it may become an intractable task, and one solution is to compute $F$ numerically instead, since $F = \frac{dy^{opt}}{dd}$. Particularly, the extended nullspace general formula (11) can, after some matrix algebra, be reformulated as in (14)

$$H = M_s(G^y)^{\dagger}[G^y \quad (G_d^y - F)](W_{n^y}^{-1}\tilde{G^y})^{\dagger}W_{n^y}^{-1} \quad (14)$$

where $M_s = (J_{uu}^{-1/2}M_n)$ can be any non-singular $n_u \times n_u$ matrix. In this case, we could select $M_n = J_{uu}^{1/2}$ so that (10) and (11) are independent of Hessian information.

## IV. Sensitivity analysis results

### A. Step 1. Operational objectives

The operational costs in a wastewater treatment plant depend on the wastewater system itself and can be divided into manpower, energy, maintenance, chemicals usage, chemical sludge treatment, and disposal costs. However, in this work, the objective is to reduce the cost of energy and sludge disposal as much as possible. Therefore, the following costs are considered:

- Required pumping energy ($E_P$ expressed in $kWh/d$);
- Required aeration energy ($E_A$ expressed in $kWh/d$);
- Required mixing energy when the aeration is too low ($E_M$ expressed in $kWh/d$);
- Sludge disposal ($C_D$ expressed in \$ $/d$).

The mathematical expressions for all these quantities can be found in [8], and by assuming a constant energy price of $k_E = 0.09$ \$ $/kWh$ and a sludge disposal price of $k_D = 80$ \$ $/ton$, the total cost in \$$/d$ can be calculated as:

$$Cost = k_E(E_P + E_A + E_M) + k_D C_D \quad [\$/d] \quad (15)$$

The overall cost function in (15) is then minimized subject to environment regulations for the effluent and some constraints related to process operability, as listed in Table I.

TABLE I
CONSTRAINTS TO THE PROCESS.

| Constraint | Unit | Status |
|---|---|---|
| $COD^{(eff)} \leq 100$ | $gCOD/m^3$ | Regulation constraint |
| $TSS^{(eff)} \leq 30$ | $gSS/m^3$ | Regulation constraint |
| $TN^{(eff)} \leq 18$ | $gN/m^3$ | Regulation constraint |
| $BOD_5^{(eff)} \leq 10$ | $gBOD/m^3$ | Regulation constraint |
| $S_{NH}^{(eff)} \leq 4$ | $gN/m^3$ | Regulation constraint |
| $Q_w \leq 1845$ | $m^3/d$ | Manipulation constraint |
| $Q_r \leq 36892$ | $m^3/d$ | Manipulation constraint |
| $Q_a \leq 92230$ | $m^3/d$ | Manipulation constraint |
| $K_La^{(1-5)} \leq 360$ | $d^{-1}$ | Manipulation constraint |

### B. Step 2. Steady-state optimal operation

There are 8 manipulated variables (last four entries in Table I), which correspond to 8 steady-state degrees of freedom ($u$). The liquid levels in the reactor tanks are assumed constant at maximum capacity due to the overflow layout considered for the plant.

Compared to other process industries, a wastewater treatment plant is subject to very distinct operation modes because of daily, weekly and seasonal variation in the incoming wastewater. In this paper we consider the influent load data as given by the IWA Task Group in the benchmark website. The data are presented in terms of ASM1 state variables and influent flow rates. In general, these data reflect expected diurnal trend variations in weekdays which are typical for normal load behavior at a municipality treatment facility. Table II summarizes the given disturbances in terms of influent flow rate and load. The average composition and flow rate and the average values for the process inputs are taken from the various weather data.

TABLE II
WEATHER PROFILES EVENTS.

| | $Q_0$ | $COD^{(in)}$ | $TSS^{(in)}$ | $TN^{(in)}$ | $T$ |
|---|---|---|---|---|---|
| | $[m^3/d]$ | $[gCOD/m^3]$ | $[gSS/m^3]$ | $[gN/m^3]$ | $[^oC]$ |
| $e_0$ | 18446 | 381 | 211 | 54 | 15 |
| $e_1$ | 21320 | 333 | 183 | 48 | 15 |
| $e_2$ | 40817 | 204 | 116 | 28 | 15 |
| $e_3$ | 19746 | 353 | 195 | 50 | 15 |
| $e_4$ | 34286 | 281 | 101 | 37 | 15 |
| $e_5$ | 20850 | 347 | 199 | 41 | 15 |
| $e_{5,min}$ | 20850 | 347 | 199 | 41 | 9 |
| $e_{5,max}$ | 20850 | 347 | 199 | 41 | 21 |

### C. Step 3. Variable selection

The results of the optimization shows that three constraints are active, namely, $TSS^{(eff)}$ (upper limit), $S_{NH}^{(eff)}$ (upper limit), and $Q_a$ (lower limit). As expected, $TSS^{(eff)}$ is at its maximum to make $Q_w$ small. In general, the reason why free ammonia ($S_{NH}^{(eff)}$) is active at its upper bound is that, as nitrification is an oxygen demanding process and because the transfer efficiency of oxygen from gas to liquid is relatively low so that only a small amount of oxygen supplied is used by the microorganisms, the aeration demand ($E_A$), which is the major cost contributor in a wastewater treatment plant, is high. One interesting fact is that the process is optimally operated aerobically, that is to say, with no anaerobic zone. The possible reason is due to the attempt to minimize the high aeration costs and to the fact that the effluent total nitrogen and ammonia constraints are quite easily reached for the given influent loads.

As those 3 active constraints must be implemented to ensure optimal operation [25], we are left with 5 degrees of freedom, and we use the local methods described in step 3 of the procedure to decide for the best (optimal) set of unconstrained self-optimizing control variables to fulfill the available degrees of freedom. We consider $n_y = 28$ measurements (the list is not shown here for the sake of compactness), $n_u = 5$ manipulated variables, and $n_d = 5$ disturbances, and clearly with $n_y > n_u + n_d$ one can expect to substantially reduce the loss for disturbances and measurement errors. As there are as many measurements as there are manipulations and disturbances, one can compute various $H$ matrices and their respective local losses. The methods considered in this communication are

1. The combined disturbances and measurements errors using all available measurements, where $H$ is computed by (10). In this case, $H_1$ is a $5 \times 28$ combination matrix.
2. The extended nullspace using all measurements, with $H$ computed by (14). In this case, $H_2$ is also a $5 \times 28$ combination matrix.
3. The extended nullspace using just enough measurements, where $\tilde{G^y}$ in (14) is found by a branch and bound algorithm [28]. In this case, $H_3$ is a $5 \times 10$ combination matrix.

For the sake of compactness only the resulting local losses calculated using (9) are reported, and they are $L_{wc}^{H_1} = 0.1184$, $L_{wc}^{H_2} = 0.1291$, and $L_{wc}^{H_3} = 0.0761$, and one should expect to have actual (nonlinear) losses of the same magnitude for any of the calculated $H$ matrices. In the last case, where the loss is expected to be the smallest, the variables chosen by the branch and bound algorithm that maximized the minimum singular value of $\tilde{G^y}$ were $S_O^{(3)}$, $S_O^{(4)}$, $S_{NO}^{(4)}$, $MLSS$, $K_La^{(1)}$, $K_La^{(2)}$, $K_La^{(3)}$, $K_La^{(4)}$, $COD^{(in)}$ and $T^{(in)}$.

## V. Discussion

The nominal optimization results showed that it is economically optimal to keep effluent suspended solids and ammonia concentrations at their respective upper bounds, and that no internal recirculation of sludge should be used, at least under the steady-state assumption. This fact is surprising but quite realistic. Indeed, the main purpose of the internal recirculation is to provide enough nitrate to enhance denitrification in the bioreactor anoxic zones and from an economical point of view this can be efficiently achieved by the return sludge (Qr) only which brings back sufficient nitrate for denitrification reducing the pumping costs due to $Q_a$. However, when operating the process dynamically, one may consider using $Q_a$ to control some internal variable so as to improve the disturbance rejection capability of the process.

If these variables are controlled at their respective optimal setpoints (active constraint control), a choice had to be made on the selection of the remaining 5 degrees of freedom, and we use the sensitivity analysis based on a plantwide procedure to decide on which 5 variables to fix/control at their respective nominal optimum values. The exact local (linear) method and the extended nullspace method based on the concept of self-optimizing control were used to systematically select those variables such that the cumbersome combinatorial curse of choosing and testing 5 out of 28 possible variable combinations is avoided. The resulting combination matrices $H$ were easily computed using elementary matrix algebra, as described by the formulas (10), (11), and (12). The only burden with those calculations lies on the computation of the optimal matrices $J_{uu}$, $J_{ud}$, and $F$. Since accuracy of second order information found numerically is known to be difficult to guarantee, in addition to assuring positive definiteness of $J_{uu}$, calculation of $F$ might become more attractive, and a replacement formula for (11) was derived as in (14). $M_n$ in this equation can be freely selected, as long as it is a non-singular matrix, and we chose $M_n = J_{uu}^{1/2}$ so to avoid the need to compute $J_{uu}$. Moreover, since the solution for $H$ in (10) is not unique [23], we can also find a non-singular $n_u \times n_u$ $D$ matrix such that $H_{new} = DH$ is another yet solution, and we can select $D$ as a function of $J_{uu}^{1/2}$; in this paper we assumed $D = J_{uu}^{-1/2}$.

The above derivations are local since we assume a linear process and a second-order objective function in the inputs and the disturbances. Thus, the proposed controlled variables are only globally optimal for the case with a linear model and a quadratic objective. In this article, for a final validation, the actual losses are checked using the nonlinear model of the process. Table III shows that the losses are about the same order of magnitude for a given disturbance. Note also that feasibility is not always guaranteed for all alternatives, and indeed only the alternative where $H$ was computed using the extended nullspace method with "just-enough" measurements is feasible for all disturbance spectrum.

TABLE III
NONLINEAR LOSS CALCULATION FOR VARIOUS DISTURBANCES.

| | $e_1$ | $e_2$ | $e_3$ | $e_4$ | $e_5$ | $e_{5,min}$ | $e_{5,max}$ |
|---|---|---|---|---|---|---|---|
| $J^{opt}$ | 426.78 | 490.09 | 420.56 | 599.36 | 419.83 | 491.28 | 357.96 |
| $J^{H_1}$ | 427.08 | 507.18 | 420.62 | 602.53 | 420.36 | 494.37 | Inf |
| $\%L^{H_1}$ | 0.07 | 3.49 | 0.014 | 0.53 | 0.13 | 0.63 | Inf |
| $J^{H_2}$ | 426.97 | 495.86 | 420.59 | 608.95 | 420.37 | 492.82 | 358.71 |
| $\%L^{H_2}$ | 0.04 | 1.18 | 0.01 | 1.60 | 0.13 | 0.31 | 0.21 |
| $J^{H_3}$ | 427.07 | Inf | 420.60 | Inf | 419.94 | 507.74 | 359.40 |
| $\%L^{H_3}$ | 0.07 | Inf | 0.01 | Inf | 0.03 | 3.35 | 0.40 |

## VI. CONCLUSION

This paper discussed the application of a sensitivity procedure for optimal operation of a wastewater treatment plant. For the given modified mathematical model of the process, where the settler is modeled based upon the static one-dimension scalar mass conservation law with discontinues fluxes theory, keeping the active constraints ($S_{NH}^{(eff)}$, $TSS^{(eff)}$, and $Q_a$) at their optimal values and using linear combinations of the measurements as the five remaining unconstrained degrees of freedom can guarantee near-optimal operation with minimum loss when operating at the nominal optimal mode despite the severe disturbances that affect the process.

## REFERENCES

[1] A. Stare, D. Vrecko, S. Hvala, and S. Strmcnik, "Comparison of control strategies for nitrogen removal in an activated sludge process in terms of operating costs: a simulation study," *Water Research*, vol. 41, pp. 2004–2014, 2007.

[2] P. Ingildsen, G. Olsson, and Y. Z., "A hedging point strategy - balancing effluent quality, economy and robustness in the control of wastewater treatment plants," *Wat. Sci. Tech.*, vol. 45, no. 4-5, pp. 317–324, 2002.

[3] P. Samuelsson, B. Halvarsson, and B. Carlsson, "Cost-efficient operation of a denitrifying activated sludge process," *Water Research*, vol. 41, pp. 2325–2332, 2007.

[4] E. Ayesa, B. Goya, A. Larrea, L. Larrea, and A. Rivas, "Selection of operational strategies in activated sludge processes based on optimization algorithms," *Wat. Sci. Tech.*, vol. 37, no. 2, pp. 327–334, 1998.

[5] A. Rivas, I. Irizar, and E. Ayesa, "Model-based optimisation of wastewater treatment plants design," *Environ. Model. Softw.*, vol. 23, pp. 435–450, 2008.

[6] B. Chachuat, N. Roche, and M. A. Latifi, "Dynamic optimisation of small size wastewater treatment plants including nitrification and denitrification processes," *Comp. Chem. Engn.*, vol. 25, pp. 585–593, 2001.

[7] A. C. B. Araujo, S. Gallani, M. Mulas, and G. Olsson, "Systematic approach to the design of operation and control policies in activated sludge systems," *Industrial and Engineering Chemistry Research*, vol. 50, no. 14, pp. 8542–8557, 2011.

[8] J. Alex, L. Benedetti, J. Copp, K. V. Gernaey, U. Jeppsson, I. Nopens, M. N. Pons, L. Rieger, C. Rosen, J. P. Steyer, P. Vanrolleghem, and S. Winkler, "Benchmark simulation model no. 1 (bsm1)," Dept. of Industrial Electrical Engineering and Automation - Lund University, Sweden, Technical Report, 2008.

[9] M. Henze, L. C. P. Grady, W. Gujer, G. V. R. Maris, and T. Matsuo, "Activated sludge model no. 1 (ASM1)," IAWQ, London, UK, Scientific and Technical Report no. 1, 1987.

[10] I. Takacs, G. G. Patry, and D. Nolasco, "A dynamic model of the clarification-thickening process," *Water Research*, vol. 29, no. 10, pp. 1263–1271, 1991.

[11] Z. Z. Vitasovic, "Continuous settler operation: A dynamic model," in *Dynamic Modeling and Expert Systems in Wastewater Engineering*, Lewis, Chelsea, Michigan, USA, 1989, pp. 59–81.

[12] B. G. Plosz, I. Nopens, J. DeClerq, L. Benedetti, and P. A. Vanrolleghem, "Shall we upgrade one-dimensional secondary settler models used in wwtp simulators? an assessment of model structure uncertainty and its propagation," *Water Science and Technology*, vol. 63, no. 8, pp. 1726–1738, 2011.

[13] S. Diehl, "A conservation law with point source and discontinuous flux function modelling continuous sedimentation," *SIAM Journal on Applied Mathematics*, vol. 56, no. 2, pp. 388–419, 1996.

[14] S. Diehl and U. Jeppsson, "A model of the settler coupled to the biological reactor," *Water Research*, vol. 32, no. 2, pp. 331–342, 1998.

[15] S. Diehl, "Operating charts for continuous sedimentation I - Control of steady states," *Journal of Engineering Mathematics*, vol. 41, pp. 117–144, 2001.

[16] S. Diehl, "The solids-flux theory - Confirmation and extension by using partial differential equations," *Water Research*, vol. 42, no. 20, pp. 4976–4988, 2008.

[17] R. Burger, S. Diehl, and I. Nopens, "A consistent modelling methodology for secondary settling tanks in wastewater treatment," *Water Research*, vol. 45, pp. 2247–2260, 2011.

[18] U. Jeppsson and S. Diehl, "An evaluation of a dynamic model of the secondary clarifier," *Water Science and Technology*, vol. 34, no. 5/6, pp. 19–26, 1996.

[19] D. Queinnec and D. Dochain, "Modelling and simulation of the steady-sate of secondary settlers in wastewater treatment plants," *Water Science and Technology*, vol. 43, no. 7, pp. 39–46, 2001.

[20] L. B. Verdickt and J. F. Van Impe, "Simulation analysis of a one-dimensional sedimentation model," in *Preprints of the 15th triennial IFAC World Congress (CDROM)*, Barcelona, Spain, 2002, p. 6.

[21] R. David, J. L. Vasel, and A. Vande Wouwer, "Settler dynamic modeling and matlab simulation of the activated sludge process," *Chemical Engineering Journal*, vol. 146, pp. 174–183, 2009.

[22] S. Skogestad, "Control structure design for complete chemical plants," *Computers and Chemical Engineering*, vol. 28, pp. 219–234, 2004.

[23] V. Alstad, S. Skogestad, and E. S. Hori, "Optimal measurement combinations as controlled variables," *Journal of Process Control*, pp. 138–148, 2009.

[24] M. Morari, G. Stephanopoulos, and Y. Arkun, "Studies in the synthesis of control structures for chemical processes, part I: formulation of the problem, process decomposition and the classification of the control task, analysis of the optimizing control structures," *AIChE Journal*, vol. 26, no. 2, pp. 220–232, 1980.

[25] A. Maarleveld and J. E. Rijnsdorp, "Constraint control on distillation columns," *Automatica*, vol. 6, pp. 51–58, 1970.

[26] S. Skogestad, "Plantwide control: The search for the self-optimizing control structure," *Journal of Process Control*, vol. 10, pp. 487–507, 2000.

[27] I. J. Halvorsen, S. Skogestad, J. C. Morud, and V. Alstad, "Optimal selection of controlled variables," *Ind. Eng. Chem. Res.*, vol. 42, pp. 3273–3284, 2003.

[28] V. Kariwala, Y. Cao, and Janardhanan, "Local self-optimizing control with average loss minimization," *Industrial and Engineering Chemistry Research*, vol. 47, pp. 150–1158, 2008.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Trajectory tracking of batch product quality using intermittent measurements and moving window estimation

Jian Wan, Ognjen Marjanovic and Barry Lennox
Control Systems Center, School of Electrical and Electronic Engineering
The University of Manchester, Manchester M13 9PL, UK
Email: {jian.wan, ognjen.marjanovic, barry.lennox}@manchester.ac.uk

***Abstract*—In order to meet tight product quality specifications for batch/fed-batch processes, it is vital to monitor and control batch product quality throughout the batch duration. The ideal strategy is to control batch product quality through trajectory tracking of a desirable batch product quality evolution during the batch run. However, due to the lack of in-situ sensors for continuous measurements of batch product quality, the measurement of batch product quality is usually implemented by laboratory assay of samples and thus these measurements are generally intermittent. Therefore direct trajectory tracking of batch product quality is not feasible for such scenarios with intermittent measurements. This paper proposes an approach to use intermittent measurements to realize trajectory tracking control of batch product quality through moving window estimation. The first step of the approach is to identify a partial least squares (PLS) model using intermittent measurements to relate process variable trajectories and batch product quality. Then the identified PLS model is further applied to predict product quality trajectory during the batch run so as to realize trajectory tracking of a desirable product quality evolution. An example from fed-batch fermentation for penicillin production is used to illustrate the principle and the effectiveness of the proposed approach.**

## I. Introduction

Batch processes are widely used in industry for manufacturing low-volume and high-value added products such as specialty chemicals, polymers and pharmaceuticals [1]. The popularity of batch operation in industry has two main reasons [2]: one is that batch processes are easier to set up and operate with the possibility of continuous improvements from earlier batch runs; the other is that batch operation is more efficient than continuous operation for frequent product changes and the production of small quantities with little or no hardware modification at all, which is especially attractive for starting commercial productions of novel materials to recover research and development costs before competing products affect prices.

The ultimate task for batch processes is to ensure consistent and desirable batch end-product quality for each batch run. This is not easy to fulfil in practice as batch processes are usually complex physical-chemical processes with time-varying and nonlinear dynamics. Furthermore, there still lack reliable in-situ sensors to monitor batch product quality during the batch run. Batch-to-batch variations resulting from changes to raw material properties and operating conditions also render robust control of batch end-product quality even more challenging. Many process monitoring and control schemes have been proposed in the literature to confront the issues encountered in batch operations [3], [4], [5], [6], [7]. Initial studies for the control of batch processes were based on mechanistic process models and traditional control methods [8], [9], [10], [11], [12], [13]. However, the identification of an accurate mechanistic model for a batch process is often difficult and time-consuming. Therefore multivariate statistical process control methods, which are based on process history data to develop empirical models, become a popular technique for modern process monitoring and control [14], [1], [15], [4], [16], [7], [17]. Among them, multi-way principal components analysis (PCA) and multi-way projection to latent structures (PLS), which are the extensions of PCA and PLS, enabling them handle three-dimensional matrices, are most widely used [18], [19], [20], [21], [7].

Using latent variable models such as multi-way PCA and PLS, currently there are two control approaches for batch processes: batch end-product quality control and trajectory tracking. The typical batch end-product quality control approach is addressed in [19], [22], where a PLS model relates process variable trajectories to batch end-product quality. Manipulated variable trajectories (MVTs) are determined such that they minimize the difference between the predicted and the target batch end-product quality. Since there is no measurement of batch product quality during the batch running period, the effectiveness of the approach relies heavily on the accuracy of the identified PLS model. The typical trajectory tracking approach is addressed in [21], [7], where a PCA model is identified to model the process dynamics of all process variable trajectories and MVTs are deduced from feeding future process variable trajectories with the target trajectories. Due to the difficulty of on-line measurements of batch product quality during the batch run, the target trajectories are usually some key process variable trajectories such as temperature set-points rather than the target batch product quality trajectories and thus the assumption is that the batch end-product quality can be guaranteed if these key process variable trajectories follow their pre-determined set-points. However, such an assumption is not always true as the batch end-product quality

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

can deteriorate in the case of disturbances even if the predetermined process variable trajectories are perfectly tracked.

Although batch product quality cannot typically be measured continuously along with other process variables due to the lack of in-situ sensors for quality measurement, it can often be measured intermittently through laboratory assay of samples taken during the batch run. Making use of the intermittent measurement data, a PLS model can be identified according to the method proposed in [5], where a series of created pseudo batches are synchronized to their batch endpoints and a PLS model is identified upon the synchronized pseudo batch data. The identified PLS model can further be applied to predict future batch product quality trajectories using moving window estimation. Thus it is possible to realize trajectory tracking of a pre-determined product quality trajectory directly rather than trajectory tracking of other process variable trajectories for a new batch run.

This paper proposes a practical approach to realize trajectory tracking control of batch product quality using intermittent measurements and moving window estimation. The paper is organized as follows: Section 2 details the methodology of the proposed trajectory tracking control of batch product quality; a case study of penicillin fermentation is described in Section 3; some conclusions and remarks are provided in Section 4.

## II. Trajectory Tracking Control of Batch Product Quality

In order to realize trajectory tracking control of batch product quality using intermittent measurements and moving window estimation, three steps are to be fulfilled: the first step is to identify a PLS model using intermittent quality measurements and a PCA model without any data from intermittent quality measurements; the second step is to predict future quality trajectories using the identified PCA&PLS models and the strategy of moving window estimation; and the third step is to compute the MVTs and implement them in a receding horizon manner. These three steps are to be described in detail in the following subsections.

### A. Model identification

For PLS, process variables are divided into two groups: the predictor values such as measured process variable trajectories and manipulated variable trajectories; the response values such as measured batch product quality variables. The method for identifying the PLS model using intermittent measurements is similar to the approach proposed in [5], where pseudo batches are created at those measurement points and they are further aligned toward their end-points for identifying the PLS model based on a selected modeling window.

Taking two batches shown in Figure 1 as an example, each batch has three intermittent measurements for product quality during the batch run. Therefore a total of six pseudo batches are created and they are aligned toward their end-points as shown in Figure 1. Then a modeling window is selected to identify the PLS model with the intermittent measurements as the response values and all other process variable measurements including the manipulated variables as the predictor values. The predictor and response values are generally three-dimensional matrices of size $I \times J \times K$, where $I$ is the number of pseudo batches for which data are available, $J$ is the number of variables that are measured and $K$ is the number of samples collected during the time period of the modeling window. These three dimensional matrices of data can be unfolded in a batch-wise way to model differences among batches [7]. The batch-wisely unfolded data are further mean-centered and scaled to be unit variance and performing PLS on the obtained data results in a latent variable model of the form:

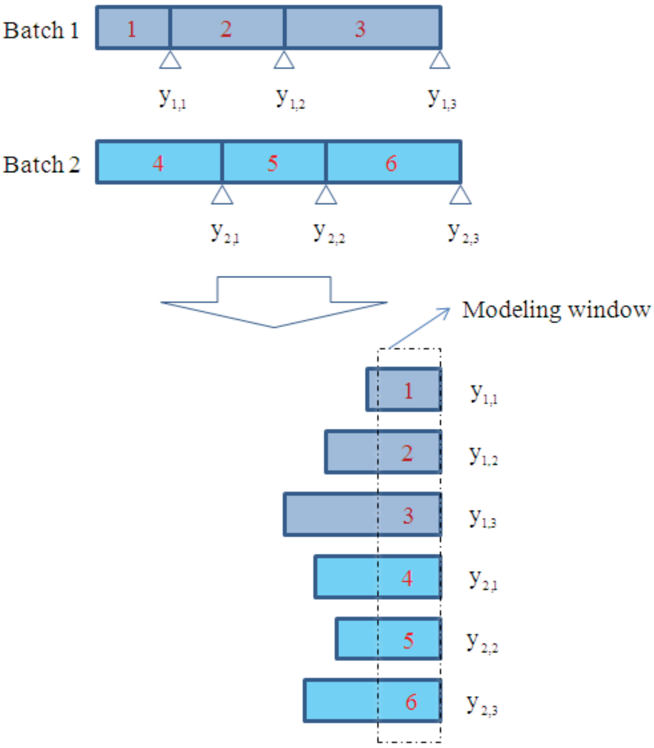


Fig. 1. Model building using intermittent measurements

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}, \tag{1}$$
$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F}, \tag{2}$$

where $\mathbf{X}$ is a matrix of $I \times J_x K_x$ for the predictor variables, $\mathbf{Y}$ is a matrix of $I \times J_y K_y$ for the response variables, $\mathbf{P}$ of $J_x K_x \times A$ and $\mathbf{Q}$ of $J_y K_y \times A$ are the loading matrices, respectively. Here $A$ is the number of latent variables. The scores $\mathbf{T}$ and $\mathbf{U}$ are related by a diagonal matrix $\mathbf{B}$ of proper dimensions with $\mathbf{U} = \mathbf{TB}$. $\mathbf{T} = \mathbf{XW}(\mathbf{P}^T\mathbf{W})^{-1}$, where $\mathbf{W}$ is the weight matrix. Finally, $\mathbf{E}$ and $\mathbf{F}$ are residual matrices. In practice, the PLS model is often expressed as a predictive model relating the predictor variables and the response variables directly [23]:

$$\mathbf{Y} = \mathbf{XW}(\mathbf{P}^T\mathbf{W})^{-1}\mathbf{B}\mathbf{Q}^T + \mathbf{F}^*, \tag{3}$$

where $\mathbf{F}^*$ is a residual matrix.

Without considering the intermittent measurements for product quality variables, multi-way PCA can be identified instead to model the correlation structure for all predictor variables:

$$\mathbf{X} = \mathbf{T}_c\mathbf{P}_c^T + \mathbf{E}_c. \tag{4}$$

### B. Moving window estimation

Once the PCA and PLS models have been identified from past batch data, they can be further used to predict future batch product quality for a new batch run through the strategy of moving window estimation. The principle of moving window estimation can be described in Figure 2, where the length of the modeling window for the identified PCA&PLS models is assumed to be $l$ and the current time instant is $\theta_i$. The first step for moving window estimation is to place the modeling window to cover the measured $l-1$ samples of predictor variables, which are $\mathbf{u}_{\theta_i-l+1\to\theta_i-1}$ for the manipulated variable $\mathbf{x}_{\theta_i-l+2\to\theta_i}$ and for all other measured process variables. Assume that the future manipulated variable trajectory is known in advance, then the future process variables $\mathbf{x}_{\theta_i+1}$ can be estimated using the identified PCA model and missing data algorithms. Several missing data imputation methods have been proposed in the literature [24], [25]. The common idea of them is to make use of the underlying data pattern to deduce the missing part from the known part. Taking the missing data algorithm called Projection to the Plane as an example, the predictor variables $\mathbf{x}$ are grouped into two parts $\mathbf{x}^T = [\mathbf{x}^{*^T}\mathbf{x}^{\sharp^T}]$, where $\mathbf{x}^{*^T} = [\mathbf{x}^T_{\theta_i-l+2\to\theta_i}\ \ \mathbf{u}^T_{\theta_i-l+1\to\theta_i}]$ contains the known data and $\mathbf{x}^\sharp = \mathbf{x}_{\theta_i+1}$ contains the missing data. The loading matrix $\mathbf{P}$ from the identified PCA model can also be grouped into two corresponding parts $\mathbf{P}_c^*$ and $\mathbf{P}_c^\sharp$. Then the missing data can be deduced from the optimal score vector $\hat{\tau}$, which is obtained from minimizing the following objective function [21]:

$$J = \frac{1}{2}(\mathbf{x}^* - \mathbf{P}_c^*\tau)^T(\mathbf{x}^* - \mathbf{P}_c^*\tau). \tag{5}$$

The optimal solution for Eq. (5) is $\hat{\tau} = (\mathbf{P}_c^{*^T}\mathbf{P}_c^*)^{-1}\mathbf{P}_c^{*^T}\mathbf{x}^*$ and thus the missing data $\mathbf{x}^\sharp$ can be deduced from it straightforwardly:

$$\mathbf{x}^\sharp = \mathbf{P}_c^\sharp\hat{\tau}. \tag{6}$$

Therefore the estimation of the future process variable $\mathbf{x}_{\theta_i+1}$, denoted as $\hat{\mathbf{x}}_{\theta_i+1}$, can be expressed as a function of the measured process variable trajectories $\mathbf{x}_{\theta_i-l+2\to\theta_i}$ and the manipulated variable trajectories $\mathbf{u}_{\theta_i-l+1\to\theta_i}$:

$$\hat{\mathbf{x}}_{\theta_i+1} = \mathbf{P}_c^\sharp(\mathbf{P}_c^{*^T}\mathbf{P}_c^*)^{-1}\mathbf{P}_c^{*^T}[\mathbf{x}^T_{\theta_i-l+2\to\theta_i}\ \ \mathbf{u}^T_{\theta_i-l+1\to\theta_i}]^T. \tag{7}$$

Using the estimated $\hat{\mathbf{x}}_{\theta_i+1}$ and the identified PLS model, the product quality at the time instant $\theta_i+1$, denoted as $\hat{\mathbf{y}}_{\theta_i+1}$, can be predicted as follows according to Eq. (3):

$$\hat{\mathbf{y}}_{\theta_i+1} = [\mathbf{x}^T_{\theta_i-l+2\to\theta_i}\ \ \hat{\mathbf{x}}^T_{\theta_i+1}\ \ \mathbf{u}^T_{\theta_i-l+1\to\theta_i}]\mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}\mathbf{B}\mathbf{Q}^T. \tag{8}$$

After the process variable values and the product quality value at the time instant $\theta_i+1$ have been estimated using the identified PCA&PLS models, the modeling window is to be moved forward as shown in Figure 2. Then the process variable values and the product quality value at the time instant $\theta_i+2$ can be deduced in a similar way, where the value of $\mathbf{x}_{\theta_i+1}$ is assumed to be known in advance as the formerly estimated value $\hat{\mathbf{x}}_{\theta_i+1}$. The whole estimation process is repeated sequentially up to the end of the control horizon $\theta_c$. So the estimated process variable trajectories $\hat{\mathbf{x}}_{\theta_i+1\to\theta_c}$ and the estimated product quality trajectories $\hat{\mathbf{y}}_{\theta_i+1\to\theta_c}$ can both be expressed as a function of the future manipulated variable trajectories $\mathbf{u}^T_{\theta_i\to\theta_c-1}$.

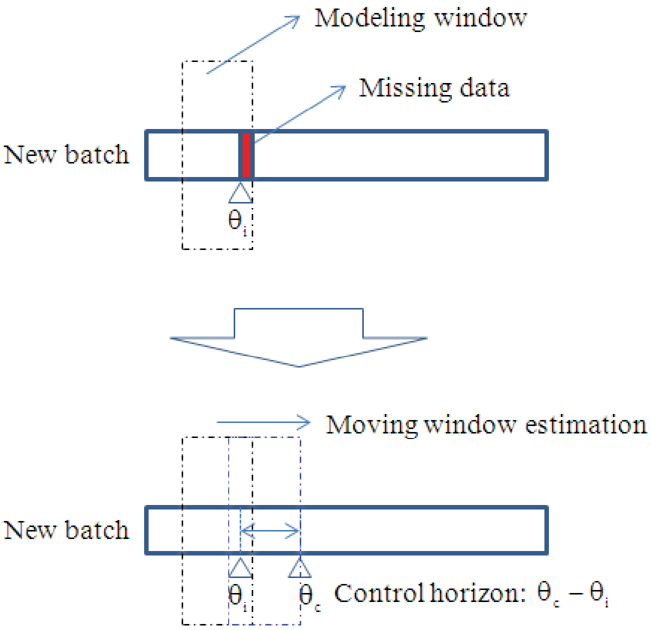


Fig. 2. Moving window estimation

### C. Trajectory tracking control

The proposed trajectory tracking control is performed in a typical receding horizon manner: the future manipulated variable trajectories with the horizon of $c$ are optimized to minimize the difference between the predicted future quality trajectory and the target future quality trajectory at each control decision point; the optimized future manipulated variable trajectories are implemented into the process up to the next control decision point and the whole process is repeated until the operation ends. Assume that the control decision point is at the current time instant $\theta_i$ and the target future quality trajectory is $\bar{\mathbf{y}}_{\theta_i+1\to\theta_i+c}$, the predicted future quality trajectory $\hat{\mathbf{y}}_{\theta_i+1\to\theta_i+c}$ can be obtained using moving window estimation. According to the moving window estimation procedure described in Eqs. (7-8), the predicted future quality trajectory $\hat{\mathbf{y}}_{\theta_i+1\to\theta_i+c}$ can be illustrated as a function of the future manipulated variable trajectory $\mathbf{u}_{\theta_i\to\theta_i+c-1}$, i.e.,

$$\hat{\mathbf{y}}_{\theta_i+1\to\theta_i+c} = \mathbf{f}(\mathbf{x}_{\theta_i-l+c+1\to\theta_i}, \mathbf{u}_{\theta_i\to\theta_i+c-1}). \tag{9}$$

The corresponding optimization for the optimal future manipulated variable trajectory $\tilde{\mathbf{u}}_{\theta_i \to \theta_i + c - 1}$ can then be formulated as follows:

$$\begin{aligned} \min_{\mathbf{u}_{\theta_i \to \theta_i + c - 1}} \quad & (\hat{\mathbf{y}}_{\theta_i + 1 \to \theta_i + c} - \bar{\mathbf{y}}_{\theta_i + 1 \to \theta_i + c})^T \mathbf{Q}_1 \\ & (\hat{\mathbf{y}}_{\theta_i + 1 \to \theta_i + c} - \bar{\mathbf{y}}_{\theta_i + 1 \to \theta_i + c}) + \\ & \Delta \mathbf{u}^T_{\theta_i \to \theta_i + c - 1} \mathbf{Q}_2 \Delta \mathbf{u}_{\theta_i \to \theta_i + c - 1}, \\ s.t. \quad & \mathbf{U}_{lb} \leq \mathbf{u}_{\theta_i \to \theta_i + c - 1} \leq \mathbf{U}_{ub} \end{aligned} \tag{10}$$

where $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are the weight matrices for trajectory errors and control rates, respectively; $\mathbf{U}_{lb}$ and $\mathbf{U}_{ub}$ are the lower and upper bound for $\mathbf{u}_{\theta_i \to \theta_i + c - 1}$, respectively.

## III. CASE STUDY

In order to assess and validate the proposed approach for trajectory tracking control of product quality, a benchmark simulation for fed-batch fermentation of penicillin is used. The simulator, called Pensim, is based upon a series of detailed mechanistic models that describe the fermentation process [26]. The following process variables are collected hourly during the fermentation process: aeration rate, agitator power, substrate feed temperature, substrate concentration, dissolved oxygen concentration, culture volume, carbon dioxide concentration, pH, fermenter temperature, generated heat and substrate feed rate. The substrate feed rate is the manipulated process variable and the batch product quality is the biomass concentration that is measurable intermittently through laboratory assay during the batch run. Ten batch data are collected for model building and each batch has a duration time of 200 hours. It is further assumed that the biomass concentration is measured four times at a time interval of 50 hours during the batch running for each batch and therefore a total of 40 pseudo batches are created according to Figure 1. The validation of the proposed approach is performed in two steps: the first step is focused on building the PCA&PLS models and validating their accuracy for the strategy of moving window estimation; the second step is focused on realizing trajectory tracking control of batch product quality using the identified PCA&PLS models and the strategy of moving window estimation. These two steps are detailed in the following two subsections.

### *A. Building the PCA&PLS models and validating their accuracy using moving window estimation*

The PCA&PLS models are identified using the created 40 pseudo bath data, where the length for the modeling window is selected to be 50 hours and the number of latent variables for the identified PCA&PLS models is selected to be 12 through cross-validation [27]. Taking a new batch run for an example, the control decision point is set at 50th hour and the future substrate feed rate is also assumed to be known. Then the future process variables such as the future carbon dioxide concentration can be estimated recursively using the identified PCA model according to Eq. (7). The estimated future carbon dioxide concentration from 50th hour to 90th hour is shown in Figure 3, where the estimated values are compared with their actual values. It can be seen that the estimated values are close to their actual values and thus the identified PCA model can be used to estimate the future process variable trajectories using the strategy of moving window estimation.

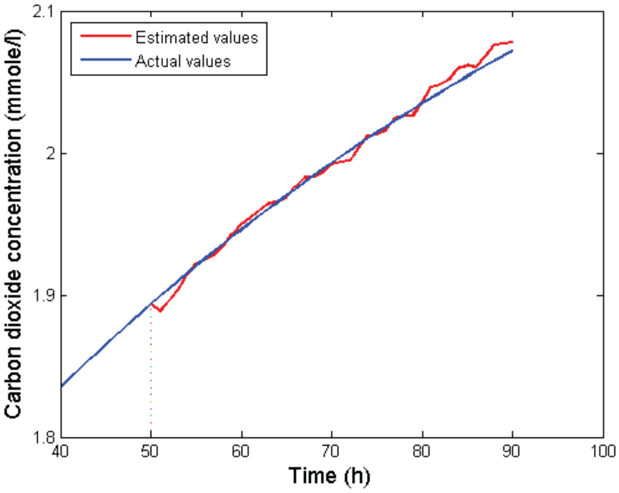


Fig. 3. Moving window estimation of the future carbon dioxide concentration

Using the same known future substrate feed rate and the estimated future process variable trajectories, the future biomass concentration can also be predicted using the identified PLS model and the strategy of moving window estimation according to Eq. (8). The predicted future biomass concentration from 50th hour to 90th hour is shown in Figure 4, where the predicted values are compared with their actual values. It can be seen that the identified PLS model can also successfully apply to predict the future product quality trajectory. These results shown in Figures 3 and 4 have demonstrated the accuracy of the identified PCA&PLS models as well as the effectiveness of moving window estimation and therefore they can be used for the following trajectory tracking control of batch product quality.

### *B. Trajectory tracking control using the identified PCA&PLS models and moving window estimation*

For the same new batch run, the task of the proposed controller is to track a pre-determined product quality trajectory during the batch run. Taking a pre-determined biomass concentration trajectory under nominal conditions as an example, the control result of tracking the predetermined trajectory is shown in Figure 5, where the control horizon is selected to be 10 hours and the controller is switched on at 50th hours. It can be seen that the target product quality trajectory has been tracked approximately. However, there are some oscillations for the controlled quality trajectory. This is due to the oscillatory trajectory of the computed substrate feed rates, as shown in Figure 6. Additional measures such as adding extra constraints on control input sequences are to be taken to reduce such oscillatory behavior for a smoother trajectory tracking in the

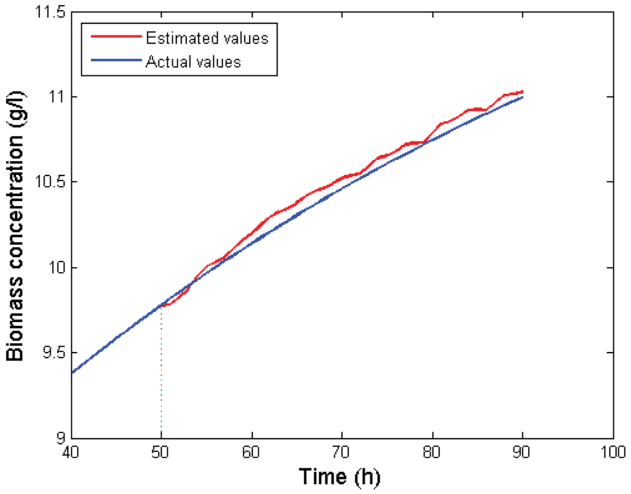


Fig. 4. Moving window prediction of the future biomass concentration

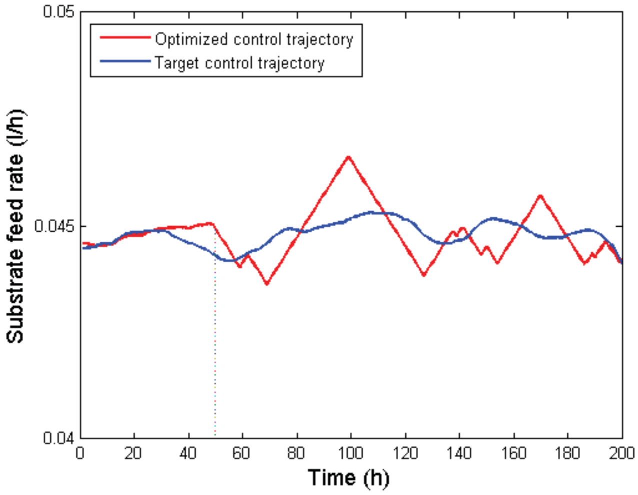


Fig. 6. The computed manipulated variable trajectory from the controller

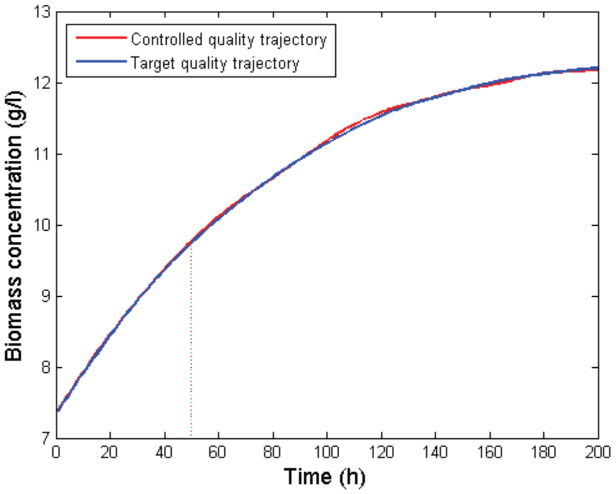


Fig. 5. Tracking an ideal product quality trajectory using the controller

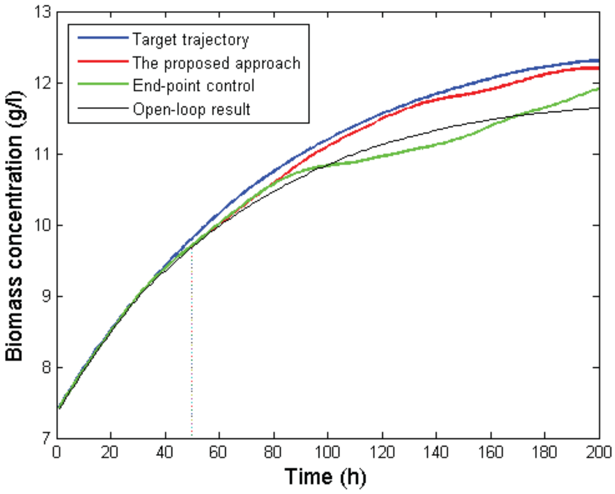


Fig. 7. Tracking product quality trajectory subject to un-modeled disturbances

future work [22].

The proposed control approach for trajectory tracking of batch product quality is further compared to the traditional end-point control approach of batch product quality [19]. In order to demonstrate the benefits of considering intermediate product quality trajectory rather than just the batch end-product quality, the fed-batch process is assumed to be subject to un-modeled disturbances. The added disturbance is chosen to be a step change in the concentration of the substrate feed from its nominal value of 600 $g/l$ to 570 $g/l$ occurring at the 30th hour. As shown in Figure 7, the traditional batch end-product quality control approach lacks the capability to detect the occurred disturbance and thus generates a much lower batch end-product quality than the proposed control approach. The proposed control approach manages to track the target product quality trajectory approximately despite the occurrence of the added disturbance and thus it can generate a much better batch end-product quality as well.

## IV. Conclusions

Due to the lack of in-situ sensors for continuous measurements of batch product quality, it is hard to realize trajectory tracking control of product quality directly. Making use of intermittent measurements for batch product quality, this paper proposes a practical approach for tracking a desirable quality trajectory during the batch runs. The proposed approach is based on the strategy of moving window estimation for online prediction of future product quality trajectory. The benchmark simulation results have demonstrated the accuracy of the identified PCA&PLS models using intermittent measurements and

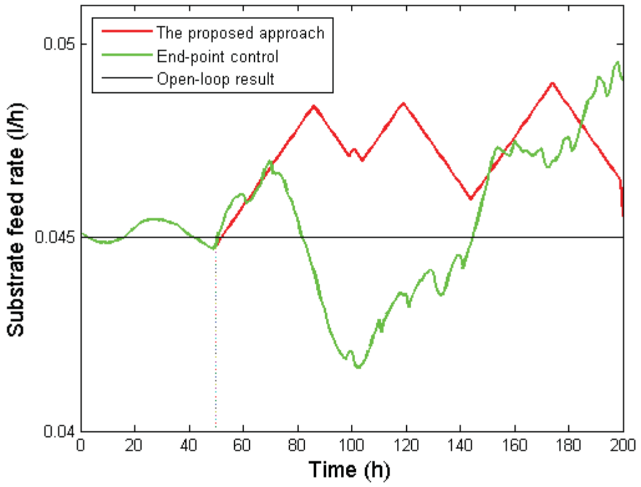


Fig. 8. The computed manipulated variable trajectories subject to un-modeled disturbances

the effectiveness of the proposed trajectory tracking scheme for batch product quality control, especially in the case of un-modeled disturbances. The obtained manipulated variable trajectory tends to be oscillatory and measures are to be taken to reduce such oscillatory behavior in the future work.

## Acknowledgment

The authors would like to acknowledge The Process Modeling, Monitoring, and Control Research Group at Illinois Institute of Technology who generously provided the source code for their Pensim simulator. The project is funded by EPSRC with the grant number EP/G022445/1.

## References

[1] S. A. Russell, P. Kesavan, J. H. Lee, and B. A. Ogunnaike. Recursive data-based prediction and control of batch product quality. *AIChE Journal*, 44:2442–2458, 1998.

[2] A. Çinar, S. J. Parulekar, C. Ündey, and G. Birol. *Batch Fermentation: Modeling, Monitoring, and Control*. Marcel Dekker, Inc., 2003.

[3] E. B. Martin and A. J. Morris. An overview of multivariate statistical process control in continuous and batch process performance monitoring. *Transactions of the Institute of Measurement & Control*, 18:51–60, 1996.

[4] T. Kourti. Application of latent variable methods to process control and multivariate statistical process control in industry. *International Journal of Adaptive Control and Signal Processing*, 19:213–246, 2005.

[5] O. Marjanovic, B. Lennox, D. Sandoz, K. Smith, and M. Crofts. Real-time monitoring of an industrial batch process. *Computers and Chemical Engineering*, 30:1476–1481, 2006.

[6] J. Chen and K. C. Lin. Integrated batch-to-batch control and within-batch online control for batch processes using two-step MPLS-based model structures. *Industrial & Engineering Chemistry Research*, 47:8693–8703, 2008.

[7] M. Golshan, J. F. MacGregor, M. J. Bruwer, and P. Mhaskar. Latent variable model predictive control (LV-MPC) for trajectory tracking in batch processes. *Journal of Process Control*, 20:538–550, 2010.

[8] A. Jutan and A. Uppal. Combined feedforward-feedback servo control scheme for an exothermic batch reactor. *Industrial & Engineering Chemistry Process Design and Development*, 23(3):597–602, 1984.

[9] P. L. Lee and G. R. Sullivan. Generic model control (GMC). *Computers & Chemical Engineering*, 12(6):573 – 580, 1988.

[10] C. Kravaris and M. Soroush. Synthesis of multivariable nonlinear controllers by input/output linearization. *AIChE Journal*, 36(2):249–264, 1990.

[11] D. Ruppen, D. Bonvin, and D. W. T. Rippin. Implementation of adaptive optimal operation for a semi-batch reaction system. *Computers and Chemical Engineering*, 22:185–199, 1997.

[12] N. Aziz, M. A. Hussain, and I. M. Mujtaba. Performance of different types of controllers in tracking optimal temperature profiles in batch reactors. *Computers and Chemical Engineering*, 24:1069–1075, 2000.

[13] M. Joly and J. M. Pinto. Optimal control of product quality for batch nylon-6,6 autoclaves. *Chemical Engineering Journal*, 97:87–101, 2004.

[14] A. Y. D. Tsen, S. S. Jang, D. S. H. Wong, and B. Joseph. Predictive control of quality in batch polymerization using hybrid ANN models. *AIChE Journal*, 42(2):455–465, 1996.

[15] J. Flores-Cerrillo. Quality control for batch processes using multivariate latent variable methods. *Ph.D thesis, McMaster University, Hamilton, ON, Canada*, 2003.

[16] I. M. Mujtaba, N. Aziz, and M. A. Hussain. Neural network based modelling and control in batch reactor. *Chemical Engineering Research and Design*, 84(8):635–644, 2006.

[17] D. Laurí, J. A. Rossiter, J. Sanchis, and M. Martnez. Data-driven latent-variable model-based predictive control for continuous processes. *Journal of Process Control*, 20(10):1207–1219, 2010.

[18] P. Nomikos and J. F. MacGregor. Monitoring batch processes using multiway principal component analysis. *AIChE Journal*, 40:1361–1375, 1994.

[19] J. Flores-Cerrillo and J. F. MacGregor. Control of batch product quality by trajectory manipulation using latent variable models. *Journal of Process Control*, 14:539–553, 2004.

[20] H. Zhang and B. Lennox. Integrated condition monitoring and control of fed-batch fermentation processes. *Journal of Process Control*, 14:41–50, 2004.

[21] J. Flores-Cerrillo and J. F. MacGregor. Latent variable MPC for trajectory tracking in batch processes. *Journal of Process Control*, 15:651–663, 2005.

[22] J. Wan, O. Marjanovic, and B. Lennox. Disturbance rejection for the control of batch end-product quality using latent variable models. *Journal of Process Control*, 22(3):643–652, 2012.

[23] A. Ferrer, D. Aguado, S. Vidal-Puig, J. M. Prats, and M. Zarzo. PLS: A versatile tool for industrial process improvement and optimization. *Applied Stochastic Models in Business and Industry*, 24(6):551–567, 2008.

[24] P. R. C. Nelson, P. A. Taylor, and J. F. MacGregor. Missing data methods in PCA and PLS: Score calculations with incomplete observations. *Chemometrics and Intelligent Laboratory Systems*, 35:45–65, 1996.

[25] F. Arteaga and A. Ferrer. Dealing with missing data in MSPC: several methods, different interpretations, some examples. *Journal of Chemometrics*, 16:408–418, 2002.

[26] G. Birol, C. Ündey, and A. Çinar. A modular simulation package for fed-batch fermentation: penicillin production. *Computers and Chemical Engineering*, 26:1553–1565, 2002.

[27] G. Diana and C. Tommasi. Cross-validation methods in principal component analysis: a comparison. *Statistical Methods and Applications*, 11:71–82, 2002.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Batch-to-batch Iterative Learning Control Using Linearised Models with Adaptive Model Updating

J. Jewaratnam, J. Zhang and J. Morris
School of Chemical Engineering and Advanced Materials
Newcastle University
Newcastle Upon Tyne NE1 7RU, UK
jegalakshimi.lingeson@ncl.ac.uk; jie.zhang@ncl.ac.uk

A. Hussain
Department of Chemical Engineering
Engineering Faculty, University of Malaya
Kuala Lumpur, Malaysia

*Abstract*— **This paper presents batch-to-batch iterative learning control (ILC) of a fed-batch fermentation process using batch-wise linearised models identified from process operation data. The newly obtained process operation data after each batch is added to the historical data base and an updated linearised model is adaptively identified. In an effort to adapt to the current process environment, the updated model is identified from a moving window of the most recent historical batches. The new model is used to compute control policy for the next batch. The control policy at different batch stages are generally correlated as the overall control policy is obtained to maximize the amount of product at the end of a batch. To address the colinearity issue, partial least square (PLS) is used in estimating the linearised model parameters. The proposed strategy is applied to a simulated fed-batch fermentation process and the performance is evaluated. The effect of window sizes was studied. Simulation results show that the proposed approach improves the batch-to-batch ILC performance.**

***Keywords-batch-to-batch control; iterative learning control; partial least square; fermentation process***

## I. INTRODUCTION

Fed-batch fermentation is vital in manufacturing high value added pharmaceutical and biochemical products. Fed-batch process is an evolution from batch reactor. The only difference between batch and fed-batch reactor is the feeding technique. In a batch reactor the substrate is only fed at the beginning of the process while in a fed-batch reactor it is fed over a few intervals with varying feed rates if necessary depending on the cell growth curve. The objective of using fed-batch in a fermentation process is to feed the substrate at the same rate that the organism utilizes it. That way, fed-batch fermentation increases input-yield ratio compared to the batch fermentation.

A core issue in fed-batch fermentation is the inability to sustain end-product quality specification due to the presence of model plant mismatches and unknown disturbances. Although every single batch repeats with exactly the same nominal initial parameters, somehow the end-product concentration varies due to unknown process variations. At present, the optimal control policy (feed rate) to obtain desired product quality in fed-batch fermentation is calculated off-line. The off-line calculated control policy may not be optimal when implemented to the real process due to model-plant mismatches and presence of unknown disturbances.

The repetitive nature of batch processes allows information from previous batches being used in modifying the control policy of the next batch in the framework of iterative learning control (ILC). ILC exploits every possibility to incorporate past control information into construction of present control action through memory based learning. The basic idea of ILC is to update the control trajectory for a new batch run using information from previous batch runs so that the output trajectory converges asymptotically to the desired reference trajectory [1],[2],[3]. The concept of ILC was first developed in the robotics industry to render a high precision in performing repetitive action of a given task [4], [5], [6]. In the recent years, ILC have been actively studied for application in injection moulding, batch reactor, chemical batch process, extrusion and batch distillation [7], [8], [9].

In a batch reactor, final product quality is usually controlled through controlling measurable variables such as pH, temperature and feed rate. The optimal trajectory of the measured variable is set and every batch run for the same process follows the same fixed trajectory [10]. This strategy fails when process disturbance in non-measured variables such as feedback condition, raw material properties, impurities and catalyst activities is present and affects the product quality [1]. This means a consistent input trajectory will not ensure product quality especially when unknown disturbances are present. This setback brings about the idea to update and re-optimize the input trajectory after every run to achieve desired product quality iteratively [10]. In a recent development, batch to batch ILC based on linearized perturbation model identified using multiple linear regressions (MLR) is reported [10]. In that work [10], the perturbation model is obtained using deviations of process input and output from their nominal trajectories and is updated after every batch by using the immediate previous batch as the nominal batch. This way, the unexpected process and parametric disturbances is expected to be captured and removed to render a more precise model prediction.

This paper presents an ILC strategy for a fed-batch fermentation process using linearised models identified from process operational data. The control policy updating is calculated using a model linearised around a reference batch. In order to cope with process variations and disturbances, the

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

reference batch can be taken as the immediate previous batch. In such a way, the model is a batch wise linearised model and is updated after each batch. The newly obtained process operation data after each batch is added to the historical data base and an updated linearised model is re-identified. In order to overcome the colinearity among the predictor variables, this paper proposes that the linearised model can be identified using partial least square regression (PLS) [11].

In order for the updated model to capture the process behavior in the face of process variations, a new technique using a moving window of the historical batches to update batch-wise linearised models is developed in this paper. The historical batches were updated after every batch run but using only the *M* recent number of batches. In other words, after every run the "oldest" batch is forgotten and the new batch is included into the sliding "window" of historical batches.

The proposed strategy is applied to a simulated fed-batch fermentation process. The results show that enhanced control performance is obtained under the proposed approach. Model updating using PLS leads to better control performance than model updating using MLR. Different window sizes were studied and the performances were evaluated.

The paper is organized as follows. Section II presents batch-to-batch ILC with updated linearised model. Application to a simulated fed-batch fermentation process is presented in Section III. Section IV concludes this paper.

## II. Batch-to-batch Iterative Learning Control with Updated Linearised Model

### *A. Linearised Models for Batch Processes*

Consider batch processes where the batch run length ($t_f$) is fixed and consists of *N* control intervals. For simplicity in implementation, the manipulated variable, $u \in R^m$ ($m$=1 in this work), is kept constant within each control interval and, thus, the control policy for a batch is a vector with *N* elements. Product quality variables (outputs), $y \in R^n$ ($n \geq 1$), can be obtained off-line by analysing the samples taken during the batch run. The product quality and control trajectories are defined, respectively, as

$$\mathbf{Y}_k=[y_k^{\mathrm{T}}(1), y_k^{\mathrm{T}}(2),\ldots, y_k^{\mathrm{T}}(N)]^{\mathrm{T}} \tag{1}$$

$$\mathbf{U}_k=[u_k(0), u_k(1),\ldots,u_k(N\text{-}1)]^{\mathrm{T}} \tag{2}$$

where the subscript *k* denotes the batch index. The desired reference trajectories of product quality are defined as

$$\mathbf{Y}_d=[y_d^{T}(1), y_d^{T}(2),\ldots, y_d^{T}(N)]^{T} \tag{3}$$

A batch process is typically modelled with a dynamic model, but it would be convenient to consider a static function relating the control sequence to the product quality sequences over the whole batch duration [10].

$$\mathbf{Y}_k = \mathbf{F}(\mathbf{U}_k) + \mathbf{v}_k \tag{4}$$

where $\mathbf{F}(\cdot)$ represents the non-linear static functions between $\mathbf{U}_k(t)$ and $y_k(t)$ at different sampling times and $\mathbf{v}_k=[v_k^T(0),v_k^T(1),\ldots, v_k^T(N\text{-}1)]^T$ is a vector of measurement noises. Linearising the non-linear batch process model described by Eq(4) with respect to $\mathbf{U}_s$ around the nominal trajectories ($\mathbf{U}_s$, $\mathbf{Y}_s$), the following can be obtained.

$$\mathbf{Y}_k = \mathbf{Y}_s + \left.\frac{\partial \mathbf{F}(\mathbf{U}_k)}{\partial \mathbf{U}_k}\right|_{\mathbf{U}_s} (\mathbf{U}_k - \mathbf{U}_s) + \mathbf{w}_k + \mathbf{v}_k \tag{5}$$

where $\mathbf{w}_k=[w_k^T(1), w_k^T(2), \ldots, w_k^T(N)]^T$ is a sequence of model errors due to the linearization (i.e., due to neglecting the higher order terms) and $\mathbf{v}_k$ represents the effects of noise and unmeasured disturbances. Define the linearised model $\mathbf{G}_s$ as

$$\mathbf{G}_s = \left.\frac{\partial \mathbf{F}(\mathbf{U}_k)}{\partial \mathbf{U}_k}\right|_{\mathbf{U}_s} \tag{6}$$

The structure of $\mathbf{G}_s$ is restricted to the following lower-block-triangular form due to the causality.

$$\mathbf{G}_s = \begin{bmatrix} g_{10} & 0 & \cdots & 0 \\ g_{20} & g_{21} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ g_{N0} & g_{N1} & \cdots & g_{NN-1} \end{bmatrix} \tag{7}$$

The linearised model can be identified from historical process operation data using MLR [10]. Let **X** and **Y** be the deviations from the reference trajectories of historical data in the manipulated variables and product quality variables respectively, then $\mathbf{Y} = \mathbf{G}_s\mathbf{X}$ and the linearised model $\mathbf{G}_s$ can be obtained through MLR as

$$\mathbf{G}_s = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \tag{8}$$

To cope with process drift, the linearised model can be re-identified after each batch run with data from the most recent batch added to the historical process data. Furthermore, the control trajectory and quality variable trajectory from the most recent batch can be used as the reference trajectories.

### *B. Partial Least Squares*

PLS projects the **X** and **Y** matrices to a subset of latent variables, **t** and **u**, respectively.

$$\mathbf{X} = \sum_{j=1}^{k} \mathbf{t}_j \mathbf{p}_j^T + \mathbf{E} \tag{9}$$

$$\mathbf{Y} = \sum_{j=1}^{k} \mathbf{u}_j \mathbf{q}_j^T + \mathbf{F} \tag{10}$$

In the above equations, **E** and **F** are residual matrices of unfitted variations in the **X** and **Y** data respectively. If sufficient numbers of latent variables (sufficiently large *k*) are used, then both **E** and **F** can be made zero or close to zero if the modeled relationship is linear. The objective is to fit a linear relationship between **X** and **Y** by performing least square regression between each pair of corresponding **t** and **u** latent vectors while making $\|\boldsymbol{F}\|$ as small as possible.

$$\hat{\mathbf{u}}_j = \mathbf{t}_j \mathbf{b}_j \qquad j=1,2,\ldots k \tag{11}$$

where $\mathbf{b}_j$ is the coefficient from the inner linear regression between the j^th^ latent variables $\mathbf{u}_j$ and $\mathbf{t}_j$ which is

$$\mathbf{b}_j = (\mathbf{t}_j^T \mathbf{t}_j)^{-1} \mathbf{t}_j^T \mathbf{u}_j \tag{12}$$

PLS provides a bilinear decomposition of the **X** and **Y** matrices into a number of rank-one matrices. The decomposition can be defined as the product between each pair of input scores vector, **t**, and predicted output scores vector, $\hat{\mathbf{u}}_j$, and a set of corresponding input and output loading vectors **p** and **q**.

The number of latent variables, *k*, to be retained in the model for PLS is usually determined through cross-validation [12]. The data set for building a model is partitioned into a training data set and a testing data set. PLS models with different number of principal components are developed on the training data and then tested on the testing data. The model with the smallest testing errors is then selected.

### C. Model Updating Using a Sliding Window of Historical Batches

Let *M* be the size of a sliding window of the past batches and use the immediate previous batch, the (k-1)th batch, as the nominal batch, then the deviations of the process input and output trajectories from their nominal trajectories in the sliding window can be represented as:

$$\Delta \mathbf{X}_{k=1} = \begin{bmatrix} \mathbf{U}_{k-M} - \mathbf{U}_{k=1} \\ \mathbf{U}_{k-M+1} - \mathbf{U}_{k=1} \\ \cdots \\ \mathbf{U}_{k-1} - \mathbf{U}_{k=1} \end{bmatrix} \tag{13}$$

$$\Delta \mathbf{Y}_{k=1} = \begin{bmatrix} \mathbf{Y}_{k-M} - \mathbf{Y}_{k=1} \\ \mathbf{Y}_{k-M+1} - \mathbf{Y}_{k=1} \\ \cdots \\ \mathbf{Y}_{k-1} - \mathbf{Y}_{k=1} \end{bmatrix} \tag{14}$$

The updated model parameters can be obtained using MLR or PLS. If correlations exist among the control actions at different stages of a batch, then PLS will give robust and reliable estimation of the model parameters.

### D. Iterative Learning Control

The batch to batch iterative learning control strategy was developed in [10] and is briefly introduced here.

The prediction of perturbation model is defined as

$$\hat{\overline{\mathbf{Y}}}_k = \hat{\mathbf{G}}_s \overline{\mathbf{U}}_k \tag{15}$$

and the absolute model prediction is defined as

$$\hat{\mathbf{Y}}_k = \mathbf{Y}_s + \hat{\overline{\mathbf{Y}}}_k = \mathbf{Y}_s + \hat{\mathbf{G}}_s \overline{\mathbf{U}}_k \tag{16}$$

After completion of the *k*^th^ batch run, prediction errors between off-line measured or analysed product qualities and their model predictions can be calculated as

$$\varepsilon_k = \mathbf{Y}_k - \hat{\mathbf{Y}}_k = \overline{\mathbf{Y}}_k - \hat{\overline{\mathbf{Y}}}_k \tag{17}$$

Based on the prediction errors of the *k*^th^ batch run, the modified prediction of perturbation model in the (*k*+1)^th^ batch run is obtained as

$$\tilde{\overline{\mathbf{Y}}}_{k+1} = \hat{\overline{\mathbf{Y}}}_{k+1} + \varepsilon_k \tag{18}$$

The absolute modified model prediction is defined as

$$\tilde{\mathbf{Y}}_{k+1} = \hat{\mathbf{Y}}_{k+1} + \varepsilon_k = \mathbf{Y}_s + \hat{\overline{\mathbf{Y}}}_{k+1} + \varepsilon_k \tag{19}$$

The modified prediction error is defined as

$$\tilde{\varepsilon}_{k+1} = \mathbf{Y}_{k+1} - \tilde{\mathbf{Y}}_{k+1} = \overline{\mathbf{Y}}_{k+1} - \tilde{\overline{\mathbf{Y}}}_{k+1} \tag{20}$$

From the definitions in Eq(17) and Eq(18), we have

$$\tilde{\varepsilon}_{k+1} = \varepsilon_{k+1} - \varepsilon_k \tag{21}$$

We assume that the prediction error of the perturbation model is bounded by a certain small positive constant $B_m$ such that

$$|\varepsilon_k| < B_m \tag{22}$$

The prediction error bound $B_m$ is a measure to represent the deviation of $\hat{\overline{\mathbf{Y}}}_k$ from $\overline{\mathbf{Y}}_k$ or $\hat{\mathbf{Y}}_k$ from $\mathbf{Y}_k$. The higher the value of $B_m$ is, the poorer the identified model is. The modified prediction error is bounded by $2B_m$ as follows

$$|\tilde{\varepsilon}_k| < |\varepsilon_k| + |\varepsilon_{k-1}| < 2B_m \tag{23}$$

The tracking errors of process and perturbation model are respectively defined as

$$\mathbf{e}_k = \mathbf{Y}_d - \mathbf{Y}_k = \overline{\mathbf{Y}}_d - \overline{\mathbf{Y}}_k \tag{24}$$

$$\hat{\mathbf{e}}_k = \mathbf{Y}_d - \hat{\mathbf{Y}}_k = \overline{\mathbf{Y}}_d - \hat{\overline{\mathbf{Y}}}_k \tag{25}$$

where $\overline{\mathbf{Y}}_d$ is the deviated desired trajectory and defined as

$$\overline{\mathbf{Y}}_d = \mathbf{Y}_d\text{-}\mathbf{Y}_s \tag{26}$$

The tracking errors of modified prediction of perturbation model is defined as

$$\tilde{\mathbf{e}}_k = \mathbf{Y}_d - \tilde{\mathbf{Y}}_k = \overline{\mathbf{Y}}_d - \tilde{\overline{\mathbf{Y}}}_k \tag{27}$$

From the definitions in Eq(17), Eq(24) and Eq(27), the following relationship among these three tracking errors can be obtained

$$\varepsilon_k = \hat{\mathbf{e}}_k - \mathbf{e}_k \tag{28}$$

$$\tilde{\mathbf{e}}_k = \hat{\mathbf{e}}_k - \boldsymbol{\varepsilon}_{k-1} \tag{29}$$

From Eq(25) and Eq(15), an iterative relationship for $\hat{\mathbf{e}}_k$ along the batch index $k$ can be obtained as

$$\hat{\mathbf{e}}_{k+1} = \hat{\mathbf{e}}_k - \hat{\mathbf{G}}_s \Delta\overline{\mathbf{U}}_{k+1} \tag{30}$$

where $\Delta\overline{\mathbf{U}}_{k+1}$ is defined as

$$\Delta\overline{\mathbf{U}}_{k+1} = \overline{\mathbf{U}}_{k+1} - \overline{\mathbf{U}}_k \tag{31}$$

From the definition of perturbation variables, we can have

$$\Delta\overline{\mathbf{U}}_{k+1} = \overline{\mathbf{U}}_{k+1} - \overline{\mathbf{U}}_k = \mathbf{U}_{k+1} - \mathbf{U}_k \tag{32}$$

Substitute Eq(28) and Eq(30) to Eq(29), we have

$$\tilde{\mathbf{e}}_{k+1} = \hat{\mathbf{e}}_{k+1} - (\hat{\mathbf{e}}_k - \mathbf{e}_k) = \mathbf{e}_k - \hat{\mathbf{G}}_s \Delta\overline{\mathbf{U}}_{k+1} \tag{33}$$

On the other hand, Eq(28) can be rewritten as

$$\mathbf{e}_k = \hat{\mathbf{e}}_k - \varepsilon_k \tag{34}$$

From Eq(34) and Eq(30), an iterative relationship for $\mathbf{e}_k$ along the batch index $k$ can also be obtained as

$$\mathbf{e}_{k+1} = \mathbf{e}_k - \hat{\mathbf{G}}_s \Delta\overline{\mathbf{U}}_{k+1} - \tilde{\varepsilon}_{k+1} \tag{35}$$

Given the error transition model in the form of Eq(33) and Eq(35), the objective of ILC is to design a learning algorithm to manipulate the control policy so that the product qualities follow the specific desired reference trajectories. The following quadratic objective function based on the modified prediction errors upon the completion of the $k^{\text{th}}$ batch run is minimised to update the input trajectory for the $(k+1)^{\text{th}}$ batch run

$$J_{k+1} = \min_{\Delta\overline{\mathbf{U}}_{k+1}} \frac{1}{2}[\tilde{\mathbf{e}}_{k+1}^T \mathbf{Q} \tilde{\mathbf{e}}_{k+1} + \Delta\overline{\mathbf{U}}_{k+1}^T \mathbf{R} \Delta\overline{\mathbf{U}}_{k+1}] \tag{36}$$

where **Q** and **R** are positive definitive matrices. Note that the objective function, Eq(36), has a penalty term on the input change $\Delta\overline{\mathbf{U}}_{k+1}$ between two adjacent batch runs, the algorithm has an integral action with respect to the batch index $k$ [10]. The weighting matrices **Q** and **R** should be selected carefully. A larger weight on the input change will lead to more conservative adjustments and slower convergence. There are also other variants of the objective function. For example, the weighting matrices **Q** and **R** may be set as **Q** = diag{$Q(1)$, $Q(2)$, …, $Q(N)$}, **R** = diag{$R(0)$, $R(1)$, …, $R(N\text{-}1)$}, where $Q(i)$ and $R(j)$ increase with respect to the time intervals $t$ in proportion to its effect of the final product quality. For the sake of simplicity, **Q** and **R** are selected in this study as $\mathbf{Q}=\lambda_q\cdot\mathbf{I}_N$ and $\mathbf{R}=\lambda_r\cdot\mathbf{I}_N$.

By finding the partial derivative of the quadratic objective function Eq(36) with respect to the input change $\Delta\overline{\mathbf{U}}_{k+1}$ and through straightforward manipulation, the following ILC law can be obtained

$$\Delta\overline{\mathbf{U}}_{k+1} = \hat{\mathbf{K}}\mathbf{e}_k \tag{37}$$

where $\hat{\mathbf{K}}$ is defined as the learning rate

$$\hat{\mathbf{K}} = [\hat{\mathbf{G}}_s^T \mathbf{Q} \hat{\mathbf{G}}_s + \mathbf{R}]^{-1} \hat{\mathbf{G}}_s^T \mathbf{Q} \tag{38}$$

From Eq(32) and Eq(37), the ILC law can be written as

$$\mathbf{U}_{k+1} = \mathbf{U}_k + \hat{\mathbf{K}}\mathbf{e}_k \tag{39}$$

## III. Application to a Fed-batch Fermentation Process

### A. A fed-batch Fermentation Process

The process considered in this paper is a fed-batch yeast fermentation process taken from [14], where a detailed kinetic and dynamic model is presented. The kinetic model of yeast metabolism is based on the bottleneck hypothesis by [15] and a dynamic model is developed based on mass balance equations for glucose, ethanol, oxygen and biomass concentrations. In this study, a simulation programme is developed in MATLAB using the kinetic and dynamic model and is verified with the results presented in [14]. The operation objective is to produce maximum amount of biomass by adjusting the glucose feed rate subject to operation constraints.

In this case study, each batch had a finite run time of 16.5hrs. The batch duration was divided into 10 equal stages and the feed rate remains constant within each stage. An initial feed rate profile was obtained from [14]. Then, 20 historical batches were generated by adding random variations to the initial feed rate profile. The end-batch biomass concentration of the historical batches ranged between 45-60g/L. Then, MLR and PLS regression methods were used in estimating the linearised model parameters from these historical process data. Batch –to-batch iterative learning control with updated historical batches was applied to the simulation. The **Q** and **R** values were fixed at 1**I** and 0.0001**I** respectively. The **R** and **Q** weighting were decided using a trial and error method.

### *B. Results for Batch to Batch ILC with Updated Models*

A batch-to-batch control study using linearised models from updated historical batches was conducted. The number of historical bathes used to develop a current batch process model keeps building up after every batch run. In other words, after every batch trial, the data is added into the pile of historical batches. Then all previous batches are used to identify a new process model which is used to generate a new control policy for the current batch. The cycle repeats and the process model is developed using both old and new batch data.

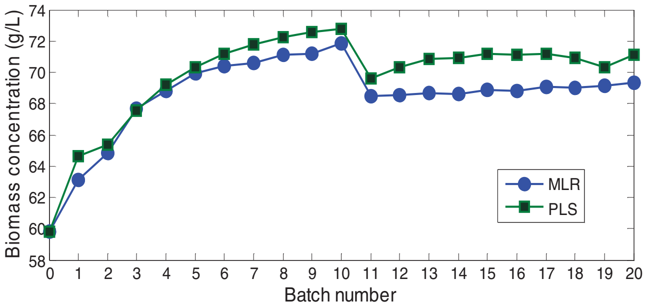


Figure 1. End of batch biomass concentration under ILC with batch-wise updated models

Figure 1 shows the performance of batch to batch ILC using updated MLR and PLS models identified from growing number of historical batches. The desired final biomass concentration was set at 74g/L. Batch 0 represents the last historical batch before implementing ILC. It is used as reference point to show process improvements due to the implementation of ILC. Batches 1 to 10 were used to test the ability of tracking desired trajectory without the presence of disturbances. From batch 11, a disturbance was introduced in that the initial substrate concentration was changed to 305g/l from its nominal value of 325g/l. Note that the initial substrate concentration is not measured and, hence, this is an unmeasured disturbance. Overall, MLR and PLS model revealed improving process operation with slight instability for both with and without disturbances. Comparing MLR and PLS models, it is evident that ILC based on the PLS model delivered higher biomass concentration for almost all the batches. This is due to the ability of PLS model to alleviate co-linearity issue. Although PLS model resulted in higher biomass concentration in the presence of disturbance, the convergence rate and stability was quite unsatisfactory. MLR model exhibited steadier performance from batch 11 to batch 20 because larger number of historical batches is favourable to obtain a better MLR model, thus better performance. The PLS model based ILC gives higher biomass but it does not always improve from batch to batch.

### *C. Results for Batch to Batch ILC with Updated Models and Moving window Historical Batches*

Further improvement was done for PLS model using a sliding window of historical batches to develop process models. After each batch run, the new batch data is added into the window of historical batches. The oldest batch in the window is removed. The idea is to use latest information to update the model and calculate the control policy for the current batch. Three sliding window sizes of 10, 15 and 20 historical batches were studied.

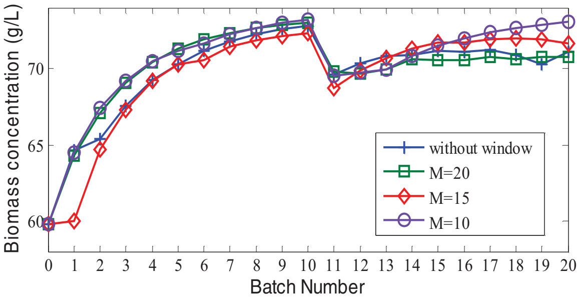


Figure 2. End of batch biomass concentration under ILC with batch-wise updated models using a sliding window of historical batches

Figure 2 shows that all three windows sizes exhibit improving results with varying stability before and after the disturbance was introduced. Performances of different window sizes (N) were compared with the one without using sliding window. For window size of 20, the convergence rate and stability were better and satisfactory when there is no disturbance. In the presence of disturbance, the results fluctuated and were not any better compared to ILC without using a sliding window. As for window size of 15, the biomass concentrations were lesser for most of the batches with no disturbances but in the presence of disturbance, the convergence rate were improving steadily from batch 11 to batch 15. From batch 16 to batch 20 slight fluctuations were noticed though the biomass concentrations were higher than the plot with no window. For window size of 10, the performance with no disturbance is as good as window size of 20. For batch 12 and 13 the biomass yield was lesser than that under ILC without using sliding window, but in the following 7 batches the performance improved steadily. The convergence rate was very satisfactory. The PLS model was able to attain final output almost as good as without disturbance within 10 batches. Amongst the three window sizes, window size of 10 gave the most stable and fastest converging performance. It is shown in the results that PLS method does not need a growing number of historical batches to develop a reliable model. An updated historical batch data with window size equal to the number of control policies used in the fed-batch fermentation process is able to generate optimal process model by using the PLS method.

## IV. Conclusions

An ILC technique with model adaptation using a sliding window of historical batches is developed in this paper. PLS is used to estimate model parameter in order to address the colinearity issues. The proposed method is applied to a simulated fed-batch fermentation process. Application results show that ILC based on batch-wise updated model using a sliding window of recent historical batches improves the control performance with and without disturbance. The effect of window sizes is studied. It is shown that model updating using PLS does not need large window size in providing enhanced control performance.

## Acknowledgment

The work is supported by University Malaya and Ministry of Higher Education, Malaysia, through a scholarship for the first author and the EU through the project iREMO (grant No. NMP2-SL-2009-228662).

## References

[1] K. S. Lee, I. S. Chin, H. J. Lee and J. H. Lee, "Model predictive control technique combined with iterative learning for batch processes", AIChE Journal, vol.45 (10) October, 1999.

[2] W .J. Campbell, S .K. Firth, A. J. Toprac and T .F. Edgar, "A comparison of run-to-run control algorithms", Proceedings of the American Control Conference, May 8-10, 2002.

[3] Z. Xiong and J. Zhang, "A batch-to-batch iterative optimal control strategy based on recurrent neural networks models", Journal of Process Control, vol.15, pp. 11-21, 2005.

[4] M. Garden, "Learning control of actuators in control systems", US Patent 3555252, 1971.

[5] M. Uchiyama, "Formation of high-speed motion pattern of a mechanical arm by trial", Transactions of Society for Implementation and Control Engineers, vol.14, pp.706-712, 1978.

[6] S. Arimoto, S. Kawamura and F. Miyazaki, "Bettering operation of robots by learning", Journal of Robotic Systems, vol.1, pp.123-140, 1984.

[7] F. Gao, Y. Yang and C. Shao, "Robust iterative learning control with application to injection molding process", Chemical Engineering Science, vol.56, pp. 7025-7034, 2001.

[8] D. A. Bristow, M. Tharayil and A. G. Alleyne. "A survey of iterative learning control: a learning-based method for high-performance tracking control", IEEE Control Systems Magazine, pp. 96-114, June 2006.

[9] K. S. Lee and J. H. Lee, "Iterative learning control-based batch process control technique for integrated control of end product properties and transient profiles of process variables", Journal of Process Control, vol.13: pp.607-621, 2003.

[10] Z. Xiong and J. Zhang, "Product quality trajectory tracking in batch processes using iterative learning control based on time-varying perturbation models", Industrial Engineering Chemical Research, vol. 42, pp. 6802-6814, 2003.

[11] P. Geladi, and B. R. Kowalski, "Partial least squares regression: a tutorial", Analytica Chimica Acta, vol. 185, pp.1-17, 1986.

[12] Wold, S., "Cross validatory estimation of the number of components in factor and principal components models", Technometrics, vol. 20, pp. 397-404, 1978.

[13] K. S. Lee and J. H. Lee, "Model predictive control for nonlinear batch processes with asymptotically perfect tracking", Computers and Chemical Engineering, vol. 21, pp. s873-s879, 1997.

[14] U. Yuzgec, M. Turker and A. Hocalar, "On-line evolutionary optimization of an industrial fed-batch yeast fermentation process", ISA Transactions, vol. 48, pp.79-92, 2009.

[15] B. Sonnleitnert and O. Kappeli, "Growth of saccharomyces cerevisiae is controlled by its limited respiratory of capacity: formulation and verification of a hypothesis", Biotechnology and Bioengineering, vol. 28, pp. 927-937, 1986.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# An investigation into sub-optimal control on the downstream processing of a large scale industrial process

Stephen Goldrick*[†‡], Barry Lennox[†], Keith Smith [‡], David Lovett[‡] and Gary Montague*
* Biopharmaceutical Bioprocess Technology Centre, Merz Court, Newcastle University, Newcastle-upon-Tyne.
[†] Control Systems Group, School of Electrical and Electronic Engineering, University of Manchester.
[‡] Perceptive Engineering Limited, Daresbury Innovation Centre, Keckwick Lane, Daresbury, Cheshire.

*Abstract*—**Industrial processes can be divided into two main areas: upstream processing; involving the manufacture of product and downstream processing; product separation and purification. Although both of these operations are necessary to obtain the final product; optimisation and improvement efforts are generally biased towards upstream processing. The importance of downstream processing can often be neglected to concentrate on the "more important"aspect of upstream processing. However, if the purification and separation steps aren't controlled effectively it can result in a significant reduction in the overall process yield. This investigation focuses on the control of a two stage counter current liquid-liquid extraction unit downstream of a batch process. A reduced separation capacity of this process is observed due to a flow oscillation in the solvent stream. The cause of this oscillation is investigated by analysing process variable behaviour, paying particular attention to the interplay between the control strategy and disturbances of the process. The effect on separation efficiency is shown through high frequency analysis of product concentrate in the output stream.**

**The aim of this work is to highlight the importance of the process control strategy. Reviewing the linkages between the process controls, product yield and quality are essential as part of a continuous improvement strategy. Such reviews can highlight opportunities for significant increases in yield that are often masked out through infrequent sampling procedures.**

## I. Introduction

Downstream processing is an integral and essential aspect of many industrial processes. Its application can be varied from industry to industry but the goal remains the same: to separate and purify a product. The downstream processing operation of concern here is liquid-liquid extraction. This is used across a large range of industries; including the biotechnology sector to separate out enzymes [1] and antibotics [2], removal of toxic metals from waste water systems [3], and in the reprocessing of nuclear fuel waste [4]. Generally liquid-liquid extraction processes involve the extraction of a product from an aqueous phase into a solvent phase. This is achieved through the physical contact between the two liquids. There are many different types of liquid-liquid extraction equipment; but the centrifuge extractor will be the focus here. This work will look at the current control strategy implemented on a liquid-liquid extraction process on a large scale manufacturing plant, depicted in Fig. 1. The unit operation is a batchwise continuous process with each batch lasting approximately 15 hours. The aqueous stream flows in the direction of Pod-A to Pod-B and the solvent phase flows counter currently in the direction of Pod-B to Pod-A. The product from the aqueous phase is extracted through physical contact with the solvent phase. Separation of the product is promoted by a chemical adjustment in the aqueous phase. Pod-A is the primary extractor; removing approximately 85% of the product and Pod-B is the secondary extractor; removing approximately 13%. Improving the separation capacity of Pod-A is the primary goal of this work.

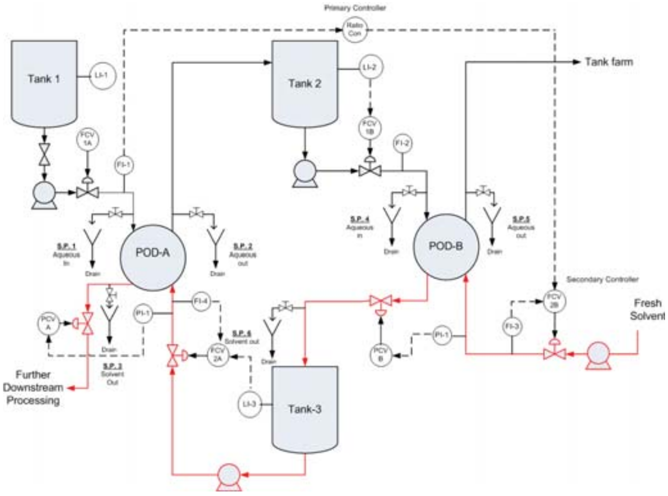


Fig. 1. Schematic of liquid-liquid extraction unit operation

Before the process trends are considered, it is important to understand the fundamentals of liquid-liquid extraction. One of the most important variables to control during operation is the back pressure on the solvent phase. This back pressure is controlled by manipulating the two pressure control valves PCV-A for Pod-A and PCV-B for Pod-B. Fig. 2 illustrates the steady state operation of one of the centrifuges/pods during normal operation. The heavy liquid in (HLI: aqueous phase) enters near the shaft, the light liquid in (LLI: solvent) enters near the rim. Due to the centrifugal forces and the differences in specific gravity of the two phases, the heavy liquid is forced towards the rim and the light liquid is forced towards the centre. This results in the formation of three phases:

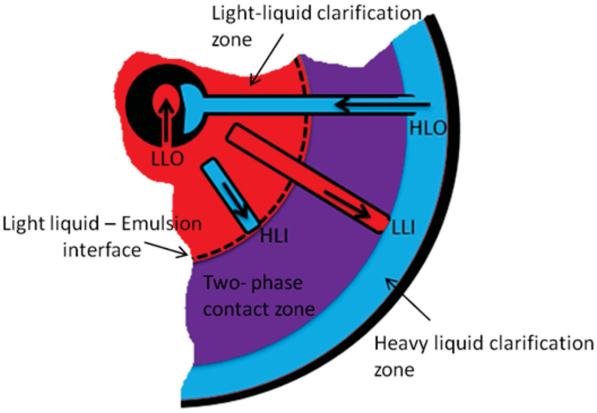

Light-liquid clarification zone
HLO
LLO
Light liquid – Emulsion interface
HLI
LLI
Two- phase contact zone
Heavy liquid clarification zone

Fig. 2. Schematic of Pod internals

1) Heavy liquid phase towards the rim.
2) An emulsion of both heavy and light liquids mixed together (contact zone).
3) Light liquid phase towards the shaft.

The position of the interface between the emulsion phase and the light liquid out (LLO) is primarily controlled by imposing a back pressure on the LLO. The position of the interface is also affected by the ratios of the two phases entering the pod. Controlling this interface is key and it is important to ensure adequate separation of the two liquids. If the back pressure is too low, heavy liquid can become entrained in the light liquid out; known as shaft flooding. And if the back pressure is too high, light liquid can't leave the shaft and can become entrained in the heavy liquid out; known as rim flooding. As shown on Fig. 1, the back-pressure of the solvent exiting the pods is controlled using a pressure control valve which is manipulated using the output from the pressure indicator on the solvent flow in. The outlet pressure is adjusted to ensure the ratio of the pressure in to pressure out is kept at 65%.

## II. Discussion

The overall ratio of the streams entering the unit is controlled through feed forward ratio control. The aqueous flow entering Pod-A produces the set point for the primary controller; the output of this is used by the secondary controller to manipulate the valve position of FCV-2B to ensure the aqueous to solvent ratio of 6.75:1 is kept as constant as possible. Fig. 3 shows a typical representation of the trends associated with this cascade control loop. The flow of aqueous entering Pod-A is shown to be relatively constant at 12,000 L/hr, changing by a maximum of 2.3%. The solvent flow entering Pod B is controlled by the secondary flow controller FCV-2B. The valve position of this secondary loop oscillates from 75% to 67% every 2-3 minutes. These small changes in the valve position result in the solvent stream fluctuating repeatedly from 1600 L/hr to 2000 L/hr; a 22.5% change . Although the output of controller FCV-2B only tries to make small adjustments to the flow, the indicator FI-3 shows that the resulting changes in the flow are significant. The difference between the flow set-point and the actual flow may be the result of numerous reasons including a high back pressure on solvent entering the process, poor tuning of the control loop or a result of a faulty or oversized valve.

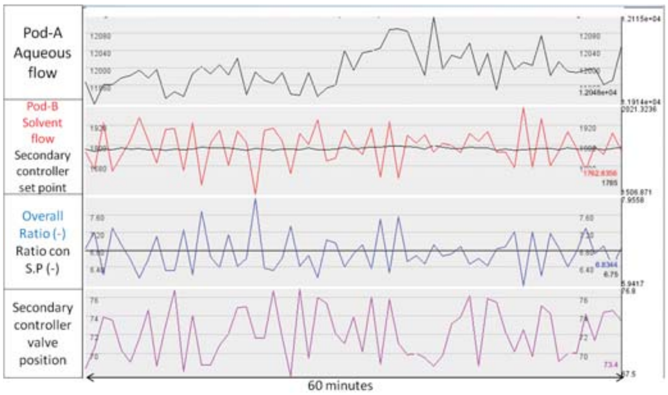


Fig. 3. Trends for overall flow ratio control

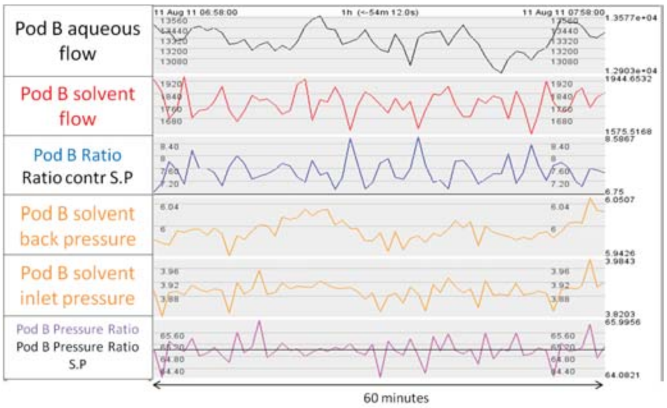


Fig. 4. Trends of flow ratio and pressures across Pod B

Fig. 4, looks at the actual ratio of the two streams entering Pod B. As the process is a continuous one, the above mentioned trends would be expected to be similar as long as the levels of the tanks 2 and 3 are kept constant. However after comparing figures 3 and 4, the flow ratio of the streams entering Pod-B is shown to be consistently higher than its set-point of 6.75:1. The pressures of the solvent entering and exiting the pods are shown to fluctuate around their set point of 65%. Some of the larger fluctuations are shown to correlate with the peaks shown by the flow ratio of Pod-B, however since both signals are quite noisy it is difficult to show a direct relationship.

Fig. 5, shows the process trends for Pod-A; a large fluctuation is seen by indicator FI-4 on the solvent stream. The ratio of the aqueous to solvent phase entering the pod is shown to periodically oscillate around its set point of 6.75, ranging from 5.45 to 7.54. This oscillation is shown to repeat throughout each batch and has a complete cycle of approximately 22

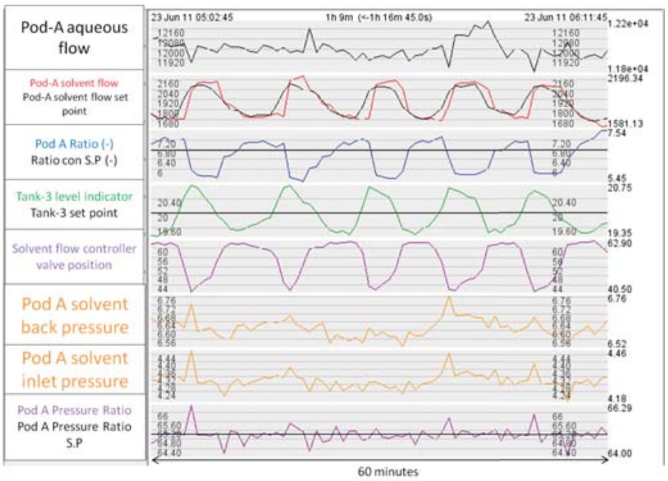


Fig. 5. Trends of flow ratio, pressures and levels associated for Pod A

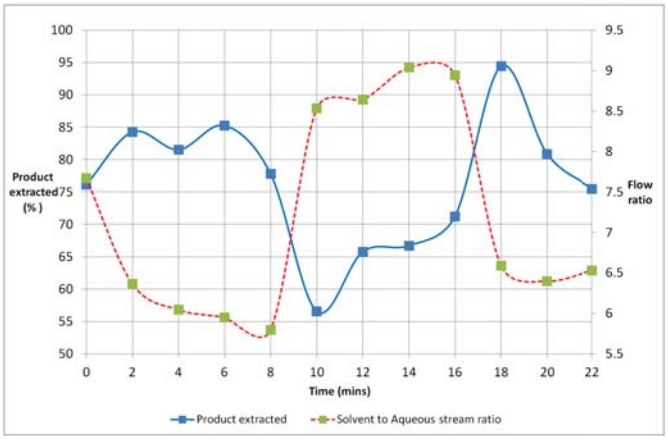


Fig. 6. High frequency sampling campaign modelling the pod efficiency during normal operation

minutes. In order to examine the root cause of this fluctuation the level controller of Tank-3 (LI-3) is examined.

The flow of solvent entering Pod-A is manipulated based on the level in Tank-3. The level in this tank is shown to oscillate around its set point of 20% , peaking at a high of 21.07% and a low of 19.23%. Although this deviation from set point is very minor, it causes the output of the valve FCV-2A to change significantly. The valve position of FCV-2A is shown to change by 20% in direct response to the small changes in the level of Tank-3. This consistent opening and closing of the valve results in the solvent stream entering Pod-A to oscillate from 1600 to 2300 L/hr with the same frequency as the level controller.The actual solvent flow to Pod-A matches the cycling set point closely showing that the root cause of the cycle to be the level controller output of Tank-3. Although the pressure ratio of the two pods are shown to be reasonably well controlled to their set points, some spikes on the flow ratio are shown to align up with the valve positions of FCV-2A. This implies that the solvent flow ratio is also having an effect on the position of the light liquid-emulsion interface which is of primary importance to the separation efficiency of the pods.

Although the above mentioned flow oscillation has been an inherent problem on this plant for many years, its effect on the separation capacity of the pods had never been quantified. The current sampling procedure is too infrequent to capture the effect of this flow oscillation on the separating efficiency of the process. The current sampling procedure takes three process samples from the sampling points 1, 3 and 5 outlined in Fig. 1 every two hours. These sample points allow one to calculate the overall efficiency of the process by performing a simple mass balance but it is too infrequent to capture the effect of the solvent flow fluctuation.

In order to quantify the effect of the solvent flow oscillation on the process, a high frequency sampling campaign was undertaken. This campaign concentrated on Pod-A, as this pod accounts for approximately 85% of the product separation. It was carried out through two separate sampling campaigns. The first campaign involved sampling the process over a 24 minute period during normal operating conditions, with the aim to model the effect of the solvent flow oscillation. The second involved fixing the ratio of the two streams entering the pod with the aim to model the pod's behaviour when the ratio of the two streams entering the pod is held constant.

The high frequency sampling campaign involved taking process sample every two minutes from the sample points 1, 2, 3 and 4 shown in Fig. 1. Once the process samples were collected, the product concentration was measured in these samples using off line analysis techniques. Fig. 6 shows the effect of the oscillating flow ratio on the separating efficiency of Pod-A, based on the product concentrations measured during the high frequency sampling campaign. The percentage of product extracted is calculated using the mass flow rate of product extracted into the solvent phase in Pod-A divided by the total mass flow rate of product entering the process.

The second high frequency sampling campaign is shown by Fig. 7, it shows the separation capacity of Pod-A when the flow ratios entering the unit are kept constant. This shows that by controlling the ratio of the two fluids to set point, the separation capacity of the pod also remains constant.

Comparing the results shown in Fig. 6 with the process trends shown in Fig. 5, an approximate average concentration for the entire batch can be calculated to be just above 75%. This can be compared against the average separation efficiency of 85% approximated by Fig. 7 when the ratio of the flows entering the pods is held constant. This highlights the significant effect of the flow fluctuation on the separation capacity of Pod-A, and is summarised by Fig. 8.

Fig. 8 highlights the opportunity to increase the overall efficiency of Pod-A by approximately 10%. This is a significant improvement to the overall efficiency of the process as the product quality extracted by Pod-A is of higher quality than that extracted from Pod-B.

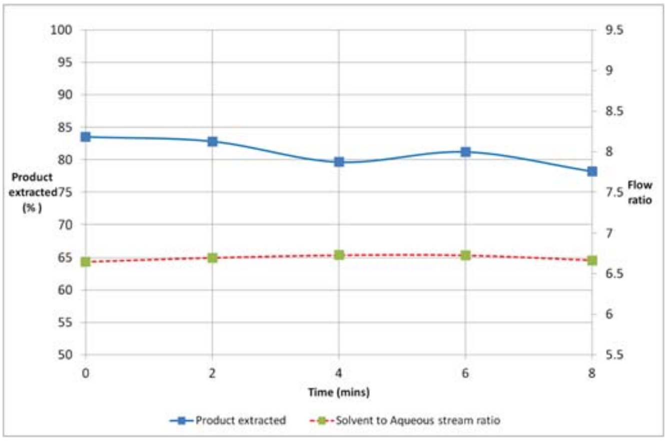


Fig. 7. High frequency sampling campaign modelling the pod efficiency during normal operation

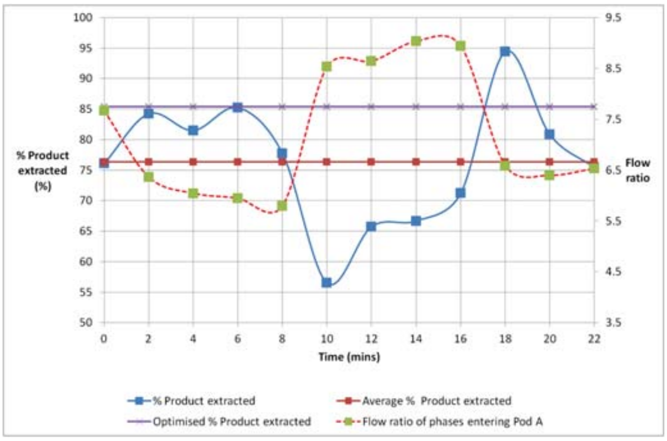


Fig. 8. Approximated average and optimised % product extracted

## III. Future work

Investigate some of the possible ways to improve the flow ratio across the two pods. One possible solution is seen by analyzing the ranks of the individual control loops. The overall ratio of flows entering the unit operation and the levels of the tanks, are shown to take precedence over the actual ratio of flows entering the pods. This ranking of the control loops does not reflect the key process parameters of the operation. This could be overcome by installing individual cascade control loops across the two pods; FCV-2A cascaded against FI-1 for Pod A and FCV-2B cascaded against FI-2 for Pod B. Although the levels in the tanks 2 and 3 will fluctuate this control strategy will ensure that the stability of the flow ratios across the pods will take priority. Once the solvent flow fluctuation have been fixed, a design of experiment will be conducted on the process with the aim of finding the optimum flow ratio set point and back pressure for the individual pods.

## IV. Conclusion

This work highlights the importance of regularly reviewing the process control strategy. These reviews must consider the interplay between the different control loops; failure to do so may result in an undesired interaction and may lead to unwanted disturbances. The disturbance highlighted here between the level controller and the flow controller was shown to have a significant effect on the separating capacity of this liquid-liquid extraction process. The work also highlights the importance of considering the key performance parameters of the process when designing the control strategy. These should take preference in the control strategy allowing the variability to be absorbed by the less important parameters such as tank levels.

## References

[1] Maria-Regina Kula, Karl Heinz Kroner and Helmut Hustedt,*Purification of enzymes by liquid-liquid extraction*,Advances in Biochemical Engineering/Biotechnology, Vol. 24,73-118, 1982.

[2] M. Reschke and K. Schgerl, *Reactive extraction of penicillin II: Distribution coefficients and degrees of extraction*,The Chemical Engineering Journal, Vol. 28, 1, B11-B20, 1984.

[3] Curtis W. Mcdonalda and Raghbir S. Bajwaa,*Removal of Toxic Metal Ions from Metal-Finishing Wastewater by Solvent Extraction*, Separation Science,Vol. 12, 4, 1977.

[4] Isabelle Billard, Ali Ouadi and Clotilde Gaillar,*Liquidliquid extraction of actinides, lanthanides, and fission products by use of ionic liquids: from discovery to understanding*,Anal Bioanal Chem, 400:15551566, 2011.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Intelligent Software Sensors and Process Prediction for Glass Container Forming Processes based on Multivariate Statistical Process Control Techniques

Dean Butler
Process Control
Ardagh Group, Glass: Doncaster
United Kingdom
E-mail dean.butler@ardaghgroup.com

Hongwei Zhang
Faculty of Arts, Computing, Engineering and Science
Sheffield Hallam University
Howard Street, Sheffield, S1 1WB, UK
E-mail H.Zhang@shu.ac.uk

***Abstract*— Glass container forming processes have attracted more attention over the past years due to the problem of lacking process information and correlation for key variables within the processes. In this paper an approach to develop process modeling and intelligent software sensing is presented for application based on multivariate statistical process control methods. The intelligent software sensors are able to provide real time estimation of key variables, and Partial Least Squares (PLS) techniques have allowed for forward prediction of final product quality variables. An application of software sensors used for container forming blank temperature is presented along with PLS being applied to predict the wall and base dimensions of glass container products. Initial results show that these methods are very promising in providing a significant improvement within this area which is usually unmonitored and is susceptible to long time delays between forming and quality inspection.**

***Keywords- Glass Container Forming; Batch Control, Software Sensors; Condition Monitoring; MSPC, PLS, PCA***

## I. Introduction

The Glass container manufacturing industry is a huge global sector of package manufacturing which produces products for various industrial sectors such as cosmetics, pharmaceutical and the food and beverage industry. Glass containers are produced by the melting of raw materials mainly sand, soda ash and limestone in a gas or electric fired furnace at temperatures above 1200 °C. The glass once molten is then distributed away from the furnace along channels known as fore-hearths which are at set controlled temperatures to the relevant glass forming machines, where upon arrival the molten glass is allowed to drain through an orifice. The molten glass is then cut as it leaves the orifice at a precise set rate to allow for enough molten material for the particular product being formed. This ‘cut’ molten glass is then delivered to the glass forming machine, typically in double or triple cavities.

At this stage high capacity, high volume production machinery is used to mould the molten glass into the various batches of containers that one manufacturing facility may produce. It is this part of the Glass Container Forming Processes which have been a difficult challenge to engineers. Difficulties with the harsh nature of the environment and obtaining accurate information relating to its process make it extremely challenging. The lack of on-line sensing for process variables has been a serious obstruction and has left the process to somewhat a ‘black art’ over the years. Now with higher expectations and demands from customers this process has never been as important to monitor and refine. At present almost all control methods and policies applied are based on off line information for the process operators and supervision, this compromises quality as the delays between product formation and inspection is so large that abnormalities can go undetected for a while. It is also relatively unknown at this stage as to the actual limits and constraints that exist on the variables within this part of the process or as to the actual combined contribution that each of these variables actually has upon the final product being made.

Multivariate Statistical process control methods based upon linear projection have attracted interest and have been are proven method for producing empirical models for Industrial processes. Principle Component Analysis (PCA) and Partial Least Squares (PLS) techniques have been applied to many practical regression problems to estimate quality related variables. Zhang and Lennox [1] applied these techniques to batch fed fermentation processes with the focus on the adoption and application of PLS and PCA techniques for sensor failure detection and prediction. Other promising results were achieved within batch fermentation processes with various publications made [2][3][4][5], applications have also been deployed within the steel industries and used for advanced monitoring of plant functions to determine the relationship between process variables and production quality [6].

The applications of the above-mentioned PLS and PCA techniques have proved successful in providing soft-sensing techniques and linear regression model prediction for process variables.

This paper aims to develop software sensors and also linear regression model prediction of glass container forming quality related variables using PLS techniques. Also the paper will demonstrate the ability of PCA to provide abnormal condition detection and isolation within glass container forming processes.

## II. Statistical Modelling and Soft-sensing using Multivariate Statistical Process Control Techniques

### A. *Principle Component Analysis*

Principle Component analysis (PCA) is a multivariate statistical method for identifying patterns within data sets by highlighting similarities and differences within the data presented. PCA attempts to find combinations of factors or variables that describe trends within the data, after data assembly PCA mathematically it is a method of writing a matrix of X variables of rank R as the sum of R matrices of rank 1 initially assuming the data are mean cantered [7].

$$X = M1 + M2 + ... + M3 + ... + Mr \tag{1}$$

Each matrix with *m* rows and *n* columns, and each variable being a column and each sample a row

PCA decomposes the matrix X as the sum of *r* $t_i$ and $p_i$ pairs where *r* is the rank of the matrix X

$$X = t_1pT_1 + t_2pT_2 + ... + t_kpT_k + ... + t_rpT_r \tag{2}$$

The $t_i$, $p_i$ pairs are ordered by the amount of variance captured. The $t_i$ vectors are known as scores and contain information on how the samples relate to each other. The $p_i$ vectors are known as loadings and contain information on how the variables relate to each other. Mathematically, PCA relies upon an eigenvector decomposition of the covariance or correlation matrix of the process variables. For a given data matrix X with m rows and n columns, the covariance matrix of X is

$$\mathrm{cov(X)} = \frac{X^T X}{m-1} \tag{3}$$

provided that the columns of X have been “mean-cantered" by subtracting off the original mean of each column.

In the PCA decomposition, the $p_i$ vectors are eigenvectors of the covariance matrix;

$$\mathrm{cov(X)}p_i = \lambda_i p_i \tag{4}$$

where $\lambda_i$ is the eigenvalue associated with the eigenvector $p_i$. The score vector $t_i$ is the linear combination of the original X variables defined by $p_i$.

Another way to look at this is that the $t_i$ are the projections of X onto the pi. The $t_i$, $p_i$ pairs are arranged in descending order according to the associated $\lambda_i$. The $\lambda_i$ are a measure of the amount of variance described by the $t_i$, $p_i$ pair. The first pair captures the largest amount of information in the decomposition and each subsequent pair captures the greatest possible amount of variance remaining after subtracting $t_ip_i$ from X.

### B. *Partial Least Squares*

PLS is a system identification tool that is capable of identifying the relationships between input (X) and output (Y) variables. The advantage that this approach offers over more traditional identification techniques, such as ordinary least squares, is that it is able to extract robust models even in applications involving large numbers of highly correlated and noisy process variable measurements.

The approach works by selecting factors of cause variables in a sequence that successively maximizes the explained covariance between the cause and effect variables. Given a matrix of cause data, **X**, and effect data, **Y**, a factor of the cause data, $\mathbf{t}_k$, and effect data, $\mathbf{u}_k$, is evaluated, such that:

$$X = \sum_{k=1}^{np<nx} t_k p_k^{\ T} + E \tag{5}$$

and

$$Y = \sum_{k=1}^{np<nx} u_k q_k^{\ T} + F \tag{6}$$

where **E** and **F** are residual matrices, *np* is the number of inner components that are used in the model and *nx* is the number of causal variables. $\mathbf{p}_k$ and $\mathbf{q}_k$ are referred to as loading vectors.

These equations are referred to as the *outer relationships*. The vectors $\mathbf{t}_k$ are mutually orthogonal. These vectors and $\mathbf{u}_k$ are selected so as to maximise the covariance between each pair, ($\mathbf{t}_k$, $\mathbf{u}_k$). Linear regression is performed between the $\mathbf{t}_k$ and the $\mathbf{u}_k$ vectors to produce the inner relationship, such that:

$$u_k = b_k t_k + \varepsilon_k \tag{7}$$

where $b_k$ is a regression coefficient, and $\varepsilon_k$ refers to the prediction error. The PLS method provides the potential for a regularised model through selecting an appropriate number of latent variables, $\mathbf{u}_k$ in the model (*np*). The number of latent variables is typically generated through the use of cross validation.

## III. Advanced Monitoring Techniques using PCA

### A. *PCA Arrangement*

PCA does not attempt to resolve any relationship between input and output data but it identifies patterns within data, as the glass container forming process is constant, by analysing the data present at one condition would give a suitable set of data to train a PCA model for identification of process changes. The data was constructed into a Matrix (X), where each column was a variable and each row a sample of the variables at a specific point in time.

$$X = \begin{bmatrix} & & \\ x_{11} & x_{21}\cdots & x_{n1} \\ x_{12} & x_{22}\cdots\cdot & x_{n2} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots \\ \cdots\cdots & \cdots\cdots & \cdots\cdots \\ x_{1(86)} & x_{2(86)}\cdots & x_{n(86)} \end{bmatrix}$$

### B. *Implementation Of Advanced Monitoring Using PCA*

A PCA Model was generated which consisted of 4 principle components which described 95% of the entire variation within the data sets , this model was then used to test against known faulty conditions which occurred within the glass container forming process.

Figure 1 shows the impact of the erroneous data upon the first two principle components. It depicts the new loaded data to exist well outside that of known good limits, this breach alone could be used to detect process abnormality. There are other components of further interest when performing Principle Component Analysis these are that of the Q residuals and Hoitelling T2 charts. The Q residuals give an indication of the measure of difference 'or residual' between a sample of data and its projection onto the components retained in the original model and the Hoi telling T2 contributions describe how much variation is within each sample to that retained within the  model. Figure 2 shows this information.

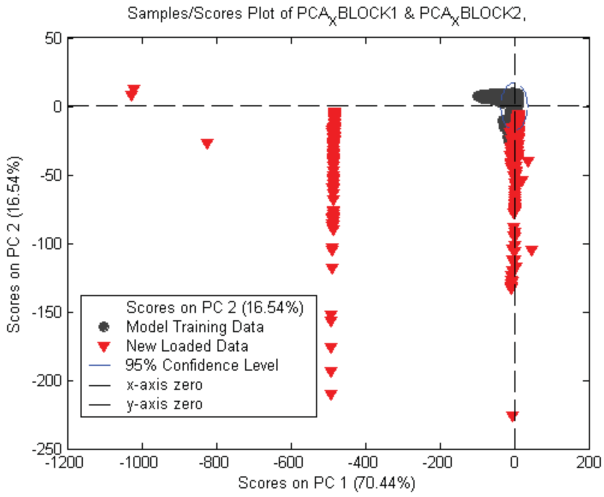


Figure 1 First two PC's loaded against Model Data

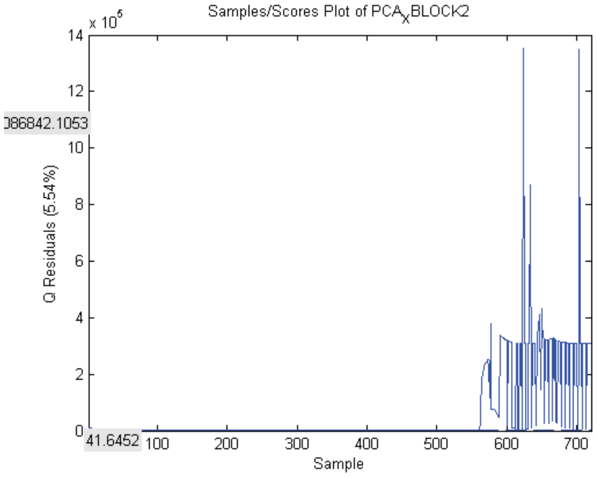


Figure 2 Q residual chart of new loaded PCA data

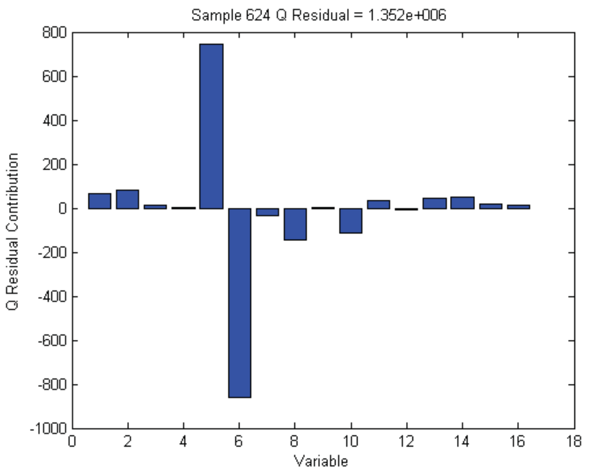


Figure 3 Q Residual Contributions at point of detection

From Figure 2, it can be seen that at around sample 570 a difference between the data in the model and the data now loaded is detected, this difference becomes even worse, at around sample 624 where the Q residual value goes extremely large. What we are actually experiencing here is an indication that something is wrong, to what we deem normal operating conditions for blank temperatures. Further investigation into the data at sample 624 shows that variables 5 and 6 are the main cause of this anomaly.

Looking at the data at this point indicates a sensor failure on IS station 3, variables 5 and 6. The above shows promising results that PCA techniques can be used for identifying process abnormalities.

## IV. Construction of Intelligent Software Sensors

### A. *PLS Model Development*

The first stage in the development of a suitable PLS Model is to obtain relevant training and model testing data. In this application Data was collected from two main process areas, the actual process variables themselves and the final product quality data measured and recorded by the operatives themselves.

There was also a further problem introduced within this process as there is huge transport lag between the process data recorded and the quality data obtained from the operatives, here an empirical approach was undertaken in order to calculate the desired transport lag and construct the data matrix for analysis to begin.

Finally a PLS Model was generated which contained 3 latent variables was developed using this data.

In this model the following measurements were used as input variables (X) 50 Psi  I-S Station operating Air , 35 Psi I-S Station Operating Air , Plunger Cooling I-S Station Air and IS Station Blank Temperatures , with Output Variables (Y) of container wall thickness measurements and container base thickness Measurements.

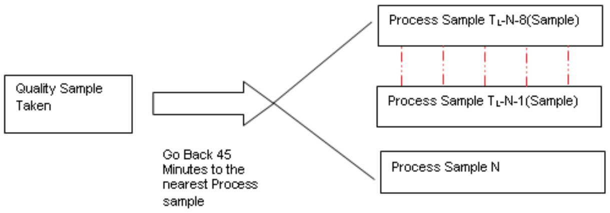

Quality Sample Taken
Go Back 45 Minutes to the nearest Process sample
Process Sample T$_L$-N-8(Sample)
Process Sample T$_L$-N-1(Sample)
Process Sample N

Figure 4 PLS Model Data Construction

The input Variables were constructed into the input variable matrix X and the output variables into the column vectors.

$$X=\begin{bmatrix} x_{11} & x_{21}\cdots & x_{n1} \\ x_{12} & x_{22}\cdots & x_{n2} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots \\ \cdots\cdots & \cdots\cdots & \cdots\cdots \\ x_{1(86)} & x_{2(86)}\cdots & x_{n(86)} \end{bmatrix}, \quad Y_1=\begin{bmatrix} y_1 \\ y_{12} \\ \cdots\cdots \\ \cdots\cdots \\ .y_{1(86)} \end{bmatrix} Y_2=\begin{bmatrix} y_2 \\ y_{22} \\ \cdots\cdots \\ \cdots\cdots \\ .y_{2(86)} \end{bmatrix}$$

These data matrices were then split into two, half utilised for model training data and the remaining half utilised for model verification.

### B. *Implementation of Software Sensors*

Based on the PLS model generated, a software sensing approach was investigated as due to the harsh nature of the environment it was common to experience sensor failures, so with the ability to predict these variables we would be at an advantage if required to adopt such soft sensing techniques for further applied control loops around this process. A 'software sensor' was developed to estimate a container forming IS station Blank temperature. The accuracy of this software sensor is illustrated in Figure 5, although there is some modelling error present this model development is a big step forward and as a basis to providing feasibility towards such applications. The figure gives good evidence that PLS based model prediction is feasible for 'soft sensing' applications on glass container forming processes.

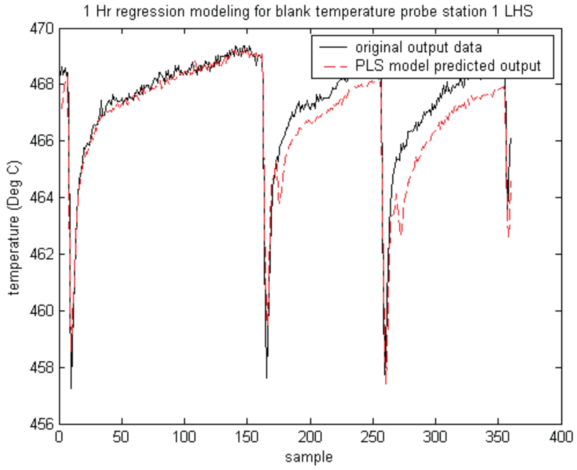


Figure 5 PLS Blank Temperature Prediction

## V. Model Prediction of Glass Container Wall and Base Thickness

### A. *PLS Model Development*

Using the data Matrices X, Y1 and Y2 previously constructed it is also possible to utilise the obtained process data to predict product quality variables.

### B. *Implementation of Linear regression Modellling using PLS fitting*

In the previous section it was demonstrated that PLS techniques can be used for software sensing techniques within glass container forming processes, in this section it will be shown that the same techniques are able to be applied to predict two important glass container quality variables - container wall thickness and base thickness. Based upon the PLS model generated, real data measured for container base thickness and wall thickness are compared with what are predicted using the container forming process input variables, this is of great benefit as the ability to predict such variables would remove the transport times involved before any feedback is generated on glass container variables.

The PLS Model generated is found to be very good in this application and has the ability to predict the container wall thickness, the model generated actually does so to within 0.5mm. Further tests gave similar promising results with the ability to predict container base thickness as shown in Figures 6 and 7. Comparing with the real data collected from production (solid green line), PLS Model is found to predict these product quality variables and shows promising evidence that these techniques and methods are applicable to glass container forming processes.

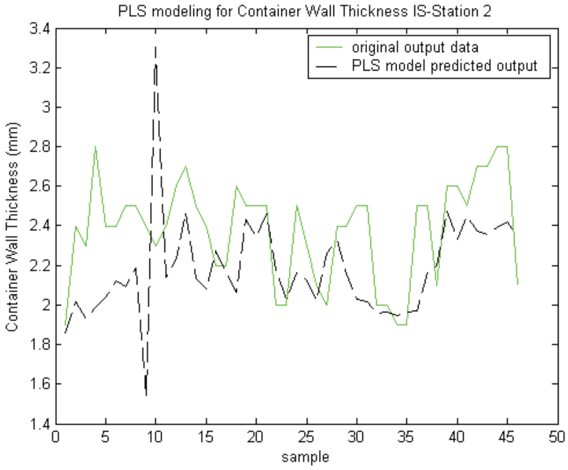


Figure 6 PLS Model Wall Thickness Prediction

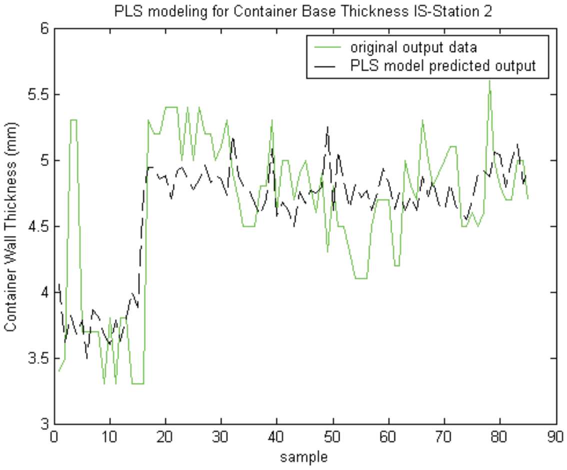


Figure 7 PLS Model Base Thickness Prediction

## VI. Conclusions

In this paper, statistical models of a glass manufacturing IS machine processes have been produced. Comparisons between the real-world data and that produced from the developed models were produced and gave promising results towards the application of such techniques. Intelligent software sensors were also developed and discussed. All the initial results show that these methods are very promising in providing a significant improvement within this area which is usually unmonitored and is susceptible to long time delays between forming and quality inspection.

## Acknowledgment

The authors of this paper would like to thank Ardagh Group for the support and the funding required for the hardware and software tools that were vital for this project.

## References

[1] H. Zhang and B. Lennox, “Integrated condition monitoring and control of fed-batch fermentation processes, ” Journal Of Process Control, vol. 14, pp. 41-50, 2004.

[2] P. Nomikos and J. F. MacGregor, “Multi-way partial least squares in monitoring batch processes,” Chemometrics and Intelligent Laboratory Systems, vol. 30, pp. 97-108, 1995.

[3] M. Ignova, J. Glassey, A. C. Ward and G.A. Montague, “Multivariate statistical methods in bioprocess fault detection and performance forecasting,” Transactions of the Institute of Measurement and Control, vol. 19, pp. 271-279, 1997.

[4] A. AlGhazzawi and B. Lennox, “Model Predictive Control monitoring using multivariate statistics,” Journal of Process Control, vol. 19, pp. 314-327, 2009.

[5] H. Zhang and B. Lennox, “Multi way optimal Control of a Benchmark fed-batch fermentation process,” Transactions of the Instititute of Measurement and Control, vol. 25, pp. 403-417, 2003.

[6] M. Kanoa and Y. Nakagawa, “Data-based process monitoring, process control, and quality improvement: Recent developments and applications in steel industry,” Computers and Chemical Engineering, vol. 32, pp. 12-24, 2008.

[7] P. Geladi and B. R. Kowalski, “Partial Least Squares Regression: A Tutorial,” Analytica Chimica Acta, vol. 185, pp.1-17, 1986.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Towards a Fully Autonomous Swarm of Unmanned Aerial Vehicles

Jeremie Leonard
Cranfield University
Cranfield, MK430AL
Email: j.leonard@cranfield.ac.uk

Dr. Al Savvaris
Cranfield University
Cranfield, MK430AL
Email: a.savvaris@cranfield.ac.uk

Prof. Antonios Tsourdos
Cranfield University
Shrivenham, Swindon, SN6 8LA
Email: a.tsourdos@cranfield.ac.uk

***Abstract*—With advances in UAS technologies the quadrotor was given a special interest for its manoeuvrability and payload capacity. These assets are amplified when more of them are deployed simultaneously in order to improve the situational awareness over areas of interest. As the number of agents operating in the same environment grows, a common intelligence is needed to optimize their cooperation and ensure their safety throughout the completion of the missions. This paper presents the results of experiments conducted to demonstrate a set of algorithms on a surveillance system employing a swarm of quadrotor UAVs to track detected targets. It was initially assumed that the UAV paths are generated at constant altitude to replace the complicated quadrotor dynamics by ones of a point mass entity. The system is then extended to the third dimension to allow for a more complex guidance and navigation scheme. Several simulations were performed under various circumstances to validate the accuracy and robustness of the system.**

**Keywords:** Autonomous, swarm, UAV, quadrotor, control.

## I. Introduction

Recent events have highlighted the ever-changing face of modern warfare and the need for strategically urban battle techniques to be developed and improved upon. The increase in asymmetric warfare shows that the current forces have particular weaknesses exploited by the enemy, due to an increased knowledge of the urban battlespace, putting lives in greater danger. With advances in UAV and UGV technologies together with the improvements in sensor capabilities there is a definite domain that allied forces can use to their advantage. Covert operations, disposable resources and getting closer to the enemy are all assets that these technologies carry on. Using multiple platforms to detect and inspect targets in regions of interest could provide precious information of the surrounding urban area, paving safer corridors, increasing the situational awareness. A thorough analysis of the data gathered by the swarm would help mapping safer and more effective missions. Ultimately, an on-board processing of these information would enable the system to function in real time applications to execute such recognition missions.

The state of the art in swarm behaviour has changed in the past few years and an increasing number of teams focus their research on formation flying and multi-agent control [1][2]. However, most of their work is based on off-the-shelf platforms modified to grab certain objects, land on a given surface or fly alongside its pears in an ideal environment. Resource optimisation, conflict resolution and failure diagnostic are still challenging fields of study to draw the need of human interaction further back.

## II. Two dimensional simulation

### A. The simplified model

Initially, the 3D environment is reduced to a two dimensional space and the 6 Degrees of Freedom (DoF) quadrotor is replaced by a 3 DoF entity. The simplified kinematics enables the testing of the following algorithms in real time and in more complex scenarios involving more units and several types of targets. In order to match the behaviour of its flying counterpart, the two dimensional model must be non-holonomic. For bench testing purposes, the point mass replacement was based on differential drive vehicle for which the user can easily control the rotation speeds $\omega_L$ and $\omega_R$ of the left and right wheels to direct the unit. This gives the unit's kinematics the advantage of being linear:

$$\begin{aligned} V_L &= R_{wheel}\,\omega_L \\ V_R &= R_{wheel}\,\omega_R \end{aligned} \tag{1}$$

$$\begin{aligned} V &= \tfrac{V_R+V_L}{2} \\ \omega &= \tfrac{V_R-V_L}{d} \end{aligned} \tag{2}$$

Thanks to this linearity, $\omega_L$ and $\omega_R$ can be monitored by a simple PI controller, easy to implement and computationally quick. The emphasis will be put on individual guidance and swarm cooperation.

### B. Guidance and navigation

In mobile robotics, it is important that the environment in which the vehicles will be driving is eciently converted to a road map that the system can use to optimize the units movements. Even if the map is only partially known or uncertain, the conversion to a C-space representation needs to be simple enough to run in real-time, but still precise enough to guarantee the optimality of the calculated path. Looking forward a little bit, the main application of the system developed in this research would be a permanent surveillance setup in a dense urban environment. In this context, a cell decomposition coupled with an A* algorithm

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

would guarantee a quick and easy way to find the shortest path between two points separated by several obstacles. Its computational lightness also makes it the most fitting solution to be implemented on the on-board processor. The A* algorithm is based on a step-by-step effort to minimize the path [3]. The idea is very intuitive: at each step of the way, it favours the solutions directly closer to the targets and leaves the other cells on the side. Those unfit solutions can be temporarily ignored but not deleted because there is no way to be sure that a path is going to be the right one. They are stored in a list of unexplored possibilities that the algorithm can search later on if the previously chosen path leads to a dead end. The path can then be cut down in the middle and restarted from a point previously left aside.

Developing an algorithm capable of finding the optimal path to a given point is essential to make the system intelligent. To make it autonomous, the quadrotor would then have to reach the goal point on its own by following the path allocated to it. Once again we want to implement on-board the algorithm in charge of guiding the quadrotor along its path. The pure pursuit (or carrot following) method answers the simplicity requirement and presents three additional assets.

- The guidance algorithm itself does not need to consider the dynamics of the vehicle.
- The method generates a virtual point, called carrot, on the path in front of the vehicle it wants to move which then has to determine the sequence of commands in speed and heading to stay at a look-ahead distance away from the virtual point while permanently aiming towards it (Fig.1). Since the point has no reality it can be created, deleted or moved with ease and the quadrotor will always follow. An appropriate control of the carrots behaviour could prove to be very beneficial for the upcoming conflict detection and resolution in 3D.

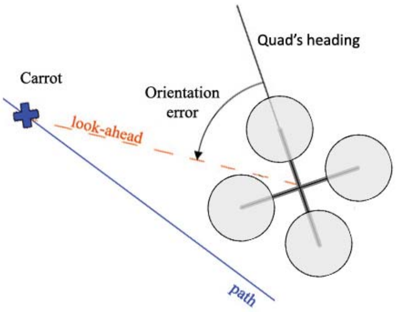


Fig. 1. Pure pursuit principle

- Though the A* generated path is made of straight segments, tracking the carrot leads the vehicle to cut the corner resulting in a much smoother trajectory and an even shorter trajectory. To avoid bumping into one of the obstacles while cutting a corner, the map of the environment is pre-processed to widen every obstacle. The path generated by the A* will therefore have a bigger clearance distance with the obstacles in the environment to ensure the safety of the quadrotor throughout the mission.

It is assumed that the vehicles have a relatively accurate knowledge of their environment. The number of unexpected obstacles should therefore be limited and a brute-force A* replaner is sufficient to guarantee an optimized avoidance. Once the system is aware of the threat, it adds the obstacle on the map and runs the A* once again. The next commands sent to the quadrotor will take the presence of the new obstacle into account and avoid it.

### *C. Cooperation*

To match the competences of their manned competitors, automated vehicles need to be developed for numerous applications which require a larger range of action and long-term possibilities. By increasing the number of units deployed, one can also increase the systems capabilities. In the scope of this work, the augmented number of agents widens the surveillance area and enables the system to reach more targets in less time [4]. Cooperative protocols are then implemented for versatility, so that the system could guarantee an optimized assignment of the quadrotors regardless of the number and types of the detected targets.

*1) Single fixed target:* When a target is spotted in the environment, the system attempts to send an agent to the targets location as fast as possible. If only one vehicle is deployed, it is automatically sent to any target that might appear in the area. The path is planned using the A* algorithm in order to reach the target with the shortest path possible. Although if several agents are used simultaneously, the system calculates the path of all the units to the target and only the closest one is assigned to the target. The others can ignore the assignment and focus on something else as shown in Fig.2.

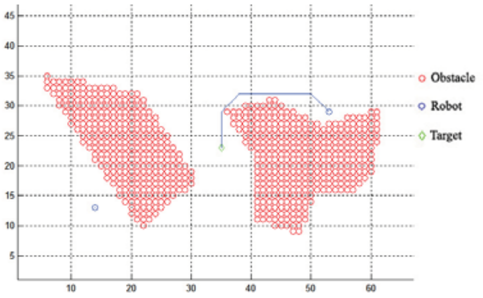


Fig. 2. Single assignment

This assignment method assumes that a target can be covered by a single unit. It would however be easy to imagine scenarios where several agents need to rendezvous at the target's location (Fig.3). For that purpose, the pure pursuit algorithm was modified to create a carrot for unit $i$ of variable speed $V_{C_i}$. Hence if we need all the vehicles to reach the target at the exact same time, each $V_{C_i}$ can be set to travel the path of unit $i$ in a given time $\tau$. The travel time $\tau$ is fixed by the central intelligence based on the kinematic limitations of the slowest unit. It could also be possible to establish a hierarchy within the vehicles forcing them to reach the target in a certain order. $V_{C_i}$ would then vary depending on the priority level of unit $i$ as well as its distance to the target.

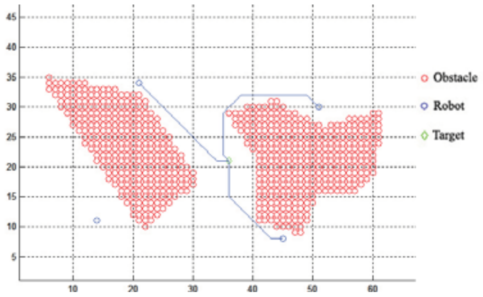


Fig. 3. Multiple assignment

*2) Multiple fixed targets:* To complete the target tracking protocols it is important to take into account a multi-target assignment. If the number of targets is smaller than the number of quarotors, the objective is to minimize the overall travelled distance. Rather than treating each target individually as they appear, the system tries to optimize the movements of the entire swarm. Every time a new target is detected, it is added to a list of previously detected targets and the allocation process is reiterated. This way each assignment is coherent with the rest of the system. To do so, the user has to deal with all the possible combinations unit/target and when the numbers increase, the optimal allocation quickly stops to be evident. An auction algorithm is therefore implemented to deal with this type of situations.

The Auction algorithm is a method used to solve classical assignment problems and is very well suited for parallel computation [5]. In the basic auction problem, there are n people and n objects (quadrotors and targets respectively) that have to be paired. Each combination person i - object j is associated to a benefit $a_{ij}$ so that the user will try and maximize the total benefit. Mathematically, the aim is to construct n pairs person-object $\{(1,j_1),(2,j_2),...,(n,j_n)\}$, where all the objects $j_1$,...,$j_n$ are distinct, that maximizes the total benefit $\Sigma_{i=1}^{n} a_{ij_i}$. The basic algorithm is modified to account for a higher number of vehicles than targets. The aim of the algorithm remains the same but in order to optimize the assignment it is capable of leaving agents out of it.

On the other hand if the number of targets becomes greater than the number of quadrotors, these targets are no longer seen as individual missions but rather as checkpoints (Fig 4). The units have to find the optimal path going through all the targets as quickly as possible. Since the number of quadrotors/targets is not limited, the solution for that problem needs to be versatile enough to answer any type of configuration. To meet this requirement, it was decided to use a genetic algorithm to handle the allocation of the targets.

Even though the method is largely based on the resolution of the multiple Travelling Salesman Problem (mTSP), parts of the method were modified to fit the overall system. In the original Traveling Salesman Problem, the salesman is supposed to come back to the initial location after visiting all the cities thereby creating a looped path. It is also common for a mTSP to fix a single point of origin for all the salesmen and a single end point. This final point can sometimes even be the same as the initial location. For the target-tracking application, each unit has a different initial position and is not required to terminate its path on the exact same node. Once again the priority is given to a minimum time constraint.

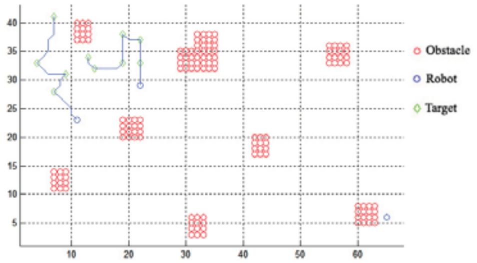


Fig. 4. Targets as checkpoints

To calculate the paths, the Genetic algorithm needs the number of targets $n_{targets}$, the number of quadrotors $n_{quads}$, and the initial positions of all of the above to create a list of all the targets and generate an array of random break points with $(n_{quads} - 1)$ elements. When applied to the target list, these break points will divide the list into $n_{quads}$ sub-lists representing the $n_{quads}$ paths travelled by the vehicles. The algorithm then needs to evaluate the fitness of the solutions and mutate the fittest genes to converge towards an optimal solution. Similarly to the Auction process, the genetic algorithm was modified so that it can choose to keep units out of the assignment if necessary.

*3) Maneuvering target:* The next step was to introduce moving targets (also called free-agents) in the environments and have the quadrotors react to it. Here, the objective is not to intercept the target anymore but rather to follow it from a given distance. To avoid predictability, the targets drive through the environment in random arcs and at variable speeds. Since the units are only moving according to the targets whereabouts, it is impossible for the system to plan a collision-free path ahead. An obstacle avoidance scheme based on the potential field method was therefore added to the system to safely follow the free-agents without hitting its surroundings (obstacles or other agents) [6].

To begin the pursuit, the user needs to fix a trailing distance $d_{trail}$ that the vehicle H needs to keep at all time with the target. Depending on the distance D to the target, this unit adapts its speed to keep up with it which is done by creating a distance-to-the-target dependent component $K_{H_d}$ to control its speed. To follow a target moving at $V_{free}$ this component would be:

$$\begin{aligned} K_{H_d} &= V_{free} + k_{H_d} \times (D - d_{trail}) \quad if\ D \geq d_{trail} \\ K_{H_d} &= V_{free} \times k_{H_d} \times (D - d_{trail}) \quad if\ D < d_{trail} \end{aligned} \tag{3}$$

To keep the unit aiming at the target, the vehicle has to maintain its heading aligned with the quad-to-target axis to keep the orientation error $e_{H_o}$ to zero. The term in Eq.4 is added to correct the orientation:

$$K_{H_\phi} = k_{H_\phi} \times e_{H_o} \tag{4}$$

The potential field obstacle avoidance is a method that converts the map known by the system into areas of attractive and repulsive potentials. The attractive field is usually generated by the target we are trying to reach but in our case, the vehicle wants to keep a certain distance with the free agent so the attractive force is excluded. The repulsive forces are generated by the known threats in order to push the vehicle away from them. Their magnitude decreases as the vehicle gets further away from them until a vicinity distance $d_0$ after which the obstacle will have no repulsive effect.

$$U_{rep_i}(q) = \begin{cases} \frac{1}{2}\gamma_{obst_i}\left(\frac{1}{d_{obst_i}(q)} - \frac{1}{d_0}\right)^\beta & if\ d_{obst_i}(q) < d_0 \\ 0 & if\ d_{obst_i}(q) \geq d_0 \end{cases} \tag{5}$$

Consider the quadrotor at the location $q = (x_{rob}, y_{rob})$ and given the linear nature of the problem, the overall potential results from the sum of the repulsive effects of all the obstacles:

$$U_{rep}(q) = \sum_{i\ \in obst} U_{rep_i}(q) = k_q \tag{6}$$

With that field in place, the vehicle is rejected as soon as it approaches an obstacle but with the absence of an attractive force from the target, the unit does not know in which direction to avoid the obstacle. To overcome that issue, the repulsive potentials from each obstacle are given an orientation to push the vehicle in the right direction as illustrated in Fig.5.

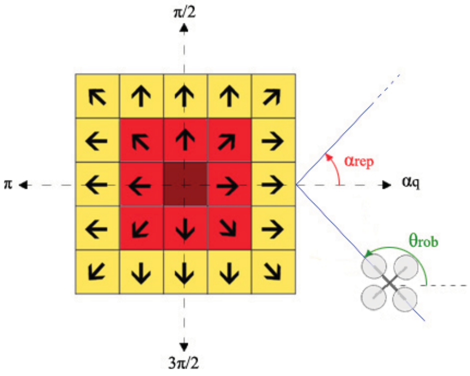


Fig. 5. Directions of rejection around an obstacle

Once again the effect on the quadrotor is converted into a new factor $K_{obst}$:

$$\begin{aligned} K_{H_{obst}} &= \frac{k_{obst} \times k_q}{\alpha_{rep}} \\ with\quad \alpha_{rep} &= mod\left[\theta_{rob} + \pi, 2\pi\right] - \alpha_q \end{aligned} \tag{7}$$

The 2D point mass model used so far is controlled through differential drive. For this particular case, the coefficients presented would be used as follow:

$$\begin{aligned} V_{H_L} &= K_{H_d} - (K_{H_\phi} - K_{obst}) \\ V_{H_R} &= K_{H_d} + (K_{H_\phi} - K_{obst}) \end{aligned} \tag{8}$$

$k_{H_d}$, $k_{H_\phi}$ and $k_{obst}$ were calibrated to ensure a smooth path around the obstacles while constantly aiming at the target ($k_{H_\phi} = R2D/5$, $k_{H_d} = 1$ and $k_{obst} = 100$).

*4) Surveillance:* Throughout this paper, the system has the ability the pick which units to assign and which to keep out of the assignment. The unassigned units can then divide the area to survey amongst them to cover it faster and more efficiently. When a new target is spotted (most likely a moving one for a realistic scenario) the system can be set to react in 2 ways.

- The closest unassigned unit is sent to the target. The remaining vehicles restart the division of the area to include the part left out by the assigned unit. This unit will follow the target wherever it goes.
- The unit whose surveillance zone includes the targets location is assigned to it while the others stay on their current path. When the target gets to another zone, the chasing agent goes back to surveillance and the agent whose zone has been entered starts chasing the target. This solution becomes interesting when each quadrotor is pre-assigned to a surveillance zone and cannot fly out of it. The reason can be that the vehicle needs to stay at all times close to a station (to charge or communicate information) in its own area.

## III. Initial Results

All the methods introduced previously were then centralized into a single coherent system accessible through a graphical user interface (GUI). The user can change the number of units deployed, add/remove fixed targets and free-agents, modify the collaboration method to deal with manoeuvring targets, and take control of a live vehicle. At this stage, the algorithms were tested in 2D scenarios so the physical platforms used for testing were ground robots.

A camera tracking device is installed to get the position of every object in the surveillance area. That position is sent to the central computer that can follow the movements of the live vehicle in real-time while the computer simulates the robot's odometry on its own. By fusing both information through a Kalman filter the system could accurately locate the vehicles and correct potential errors in their paths via the Kalman updates. For the implementation of the filter, it is important to specify that the kinematics of the system are linear and that each iteration of the filter will occur at fixed discrete time intervals, meaning that the time evolution of the state vector can be calculated by means of a state transition matrix. The accuracy of the filter was tested on several types of trajectories, from the simple straight line to a figure of 8, and the physical vehicle was always kept on its path. But in order to successfully extend to the quadrotor platform the filter needs to be robust.

The first series of tests consisted in creating systematic errors as would the IMU bias create on the quadrotor platform.

To simulate that, Fig.6 shows the path travelled by the robot $ROB_{real}$ with a wheel larger than the other (x1.5). The original path of the simulated vehicle $ROB_{sim}$ is travelled three times in a row to make sure the physical one returns to its original position.

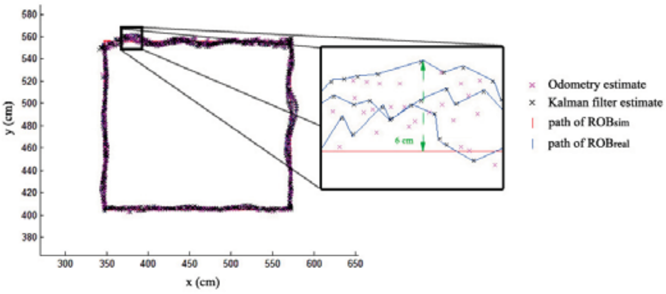


Fig. 6. Influence of a systematic error on a square path

Oscillations appear after the turns but they are rapidly damped in the straight segment. The maximum deviation from the original path was of 6cm representing only a third of the platform's width. This shows that the Kalman filter and on-board controller can keep the agent on track and on time even with sensor bias or physical damage.

To test the effects of non-systematic errors, the vehicle is returned to its calibrated state to travel along the same path. During a turn, one of the wheels is blocked to induce an error in heading then the robot is manually moved in the middle of its path to create a significant error in position. In both cases the vehicle finds its way back to the original path.

The discrete Kalman filter proves to be robust enough to overcome the effects of systematic and non-systematic errors.

## IV. Extension to the third dimension

Once the algorithms were extensively tested in their 2D environment to prove their accuracy and robustness, they were extended to the third dimension. The model of the vehicles was switched back to a proper quadrotor helicopter [7][8] and the previous methods were optimized for 3D applications towards health management.

### A. Quadrotor model

The quadrotor is very simply modelled as four rotors equipped at each end of a cross. All the propellers axes are parallel and fixed directly on the DC motors shaft. The blades have a fixed pitch and push the air downwards. Fig.7 shows the quadrotor structure in hover. Each propeller is defined by an orientation of rotation (blue circular vector), a rotating speed ($\Omega_1, \Omega_2, \Omega_3, \Omega_4$ for respectively the front, right, rear and left rotor) and a velocity (blue vertical vector) representing the amount of thrust created by the motor/propeller assembly. The fixed-body B-frame, in red, is set up to be front / left / up so that the altitude is measured positively as the quadrotor is climbing.

To take advantage of the body symmetry and keep the inertia matrix time-invariant, the equations of motion are formulated in the B-frame hence two assumptions can easily be made to simplify the model:

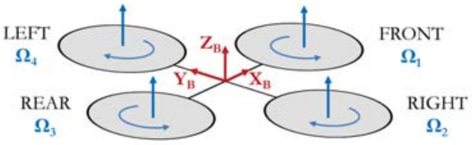


Fig. 7. Simplified quadrotor in hover

- The origin of the B-frame is coincident with the quadrotor's centre of mass.
- The quadrotor's axes of inertia coincide with the axes of the B-frame.

Under these assumptions and following the Newton-Euler formalism, the dynamics of a rigid body subject to external forces applied to its centre of mass can be expressed in the B-frame by Eq.9.

$$\begin{bmatrix} mI_{3\times3} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \dot{V} \\ \dot{\omega} \end{bmatrix} + \begin{bmatrix} \omega.mV \\ \omega.I\omega \end{bmatrix} = \begin{bmatrix} F \\ \tau \end{bmatrix} \tag{9}$$

To comply with real-time computation capabilities, the less influent effects such as the hub forces, ground effect and rolling moments were neglected and the thrust/drag coefficients were assumed constant. The matrix system (9) can then be rewritten in a state-space form $\dot{X} = f(X, U)$ where U is the input vector and X the state vector chosen as follows:

$$\begin{array}{l} X = [\,\phi\ \dot{\phi}\ \theta\ \dot{\theta}\ \psi\ \dot{\psi}\ x\ \dot{x}\ y\ \dot{y}\ z\ \dot{z}\,]^T \\ U = [\,U_1\ U_2\ U_3\ U_4\,]^T \end{array} \tag{10}$$

with

$$\begin{array}{l} U_1 = b\,(\Omega_1^2 + \Omega_2^2 + \Omega_3^2 + \Omega_4^2) \\ U_2 = l\,b\,(-\Omega_2^2 + \Omega_4^2) \\ U_3 = l\,b\ -(\Omega_1^2 + \Omega_3^2) \\ U_4 = d\,(-\Omega_1^2 + \Omega_2^2 - \Omega_3^2 + \Omega_4^2) \\ \Omega = -\Omega_1 + \Omega_2 - \Omega_3 + \Omega_4 \end{array} \tag{11}$$

From which we can derive the following equations of motion:

$$\begin{array}{l} \ddot{\phi} = \dot{\theta}\,\dot{\psi}\,\frac{I_{yy}-I_{zz}}{I_{xx}} - \dot{\theta}\,\Omega\,\frac{J_R}{I_{xx}} + U_2\,\frac{l}{I_{xx}} \\ \ddot{\theta} = \dot{\phi}\,\dot{\psi}\,\frac{I_{zz}-I_{xx}}{I_{yy}} + \dot{\phi}\,\Omega\,\frac{J_R}{I_{yy}} + U_3\,\frac{l}{I_y} \\ \ddot{\psi} = \dot{\phi}\,\dot{\theta}\,\frac{I_{xx}-I_{yy}}{I_{zz}} + \frac{U_4}{I_{zz}} \\ \ddot{x} = (s_\psi\,s_\phi\ + c_\psi\,s_\theta\,c_\phi)\,\frac{U_1}{m} \\ \ddot{y} = (-c_\psi\,s_\phi\ + s_\psi\,s_\theta\,c_\phi)\,\frac{U_1}{m} \\ \ddot{z} = -g + c_\theta\,c_\phi\,\frac{U_1}{m} \end{array} \tag{12}$$

A PID controller was added to the model to stabilize the model and then tested on the physical quadrotor. Without feedback from the camera tracking device presented above, the position controller was turned off for the first flights. The objective was to see if by simply controlling the throttle the attitude controller can stabilize the platform in hover. With the exact same PID coefficient as in the simulation, the flight was successful. A slight drift appeared on the X axis just like the simulation of a lone attitude controller would suggest but it was easily decreased by tuning the ESCs and the PID coefficients (all less than 10% away from their simulation value) and will be completely cancelled once the position controller is in place.

### B. Improvements on the 2D protocols

The A* path planner and the carrot following method were both kept for the guidance of the quadrotor platform. Compared to their previous centralized solution adopted for the 2D application, the decision-making is much more distributed. The quadrotor offers more processing power to run more complex algorithm so that both the path planning and tracking techniques could be run on-board. Apart from increasing the number of primitive movement at each iteration, the A* process stayed the same. Fig.8 shows the 3D path generated by the improved A* as well as the quadrotor following the carrot point as it travels the path.

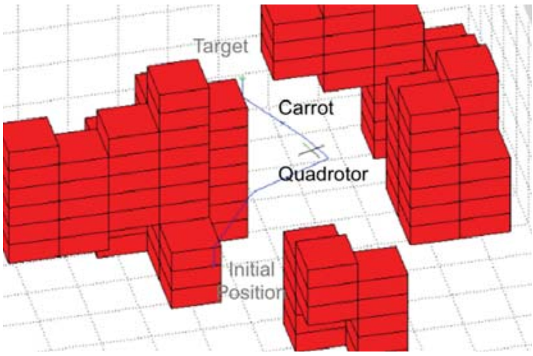


Fig. 8. Pure pursuit path following

On the other hand, significant changes were made to the pure pursuit to take more control over the carrot and facilitate the integration of a new Conflict Detection and Resolution scheme. The Distributed Reactive Collision Avoidance (DRCA) algorithm [9] is being developed to guarantee a conflict resolution scheme for $n$ nonholonmic vehicles, based on speed changes and lateral manoeuvres. It also distributes the computation load among the different subsystems. Consequently, the central computer can focus on a high-level task management and would be able to handle more units. Hence the overall system does not depend exclusively on the centralized intelligence making it more robust to eventual bugs.

Originally the DRCA takes into account the dynamic limitations of the platform in the conflict resolution scheme. Here the carrot is assumed to be a projection of quadrotor in the near future so that the avoidance process can be applied to the carrot instead of the vehicle. Since the carrot has no reality, it has no dynamic limitations. It can be created, moved and deleted freely to go around an obstacle or avoid an incoming friendly agent. The dynamics of the quadrotor will be handled by the controller and path tracking methods already in place. This will enable the generation of new avoidance trajectory and extend the effectiveness of the DRCA to more complex scenarios.

The Auction algorithm was kept to deal with the target assignment problems, and was improve to incorporate the case of $n_{targets} > n_{units}$ previously covered by the genetic algorithm. Though it was giving good results, the genetic algorithm is computationally heavy and does not fit the real time expectations of this work. The new Auction algorithm also has the ability to assign units as to optimize time rather than energy. The 2D version tried to minimize the added length of all the paths thus minimizing the power consumed by the swarm. Now the algorithm can choose to minimize the length of the longest path thus trying to optimize the time needed to reach all the targets.

## V. Further Work: Towards full autonomy

The quadrotor should be able to detect a faulty behaviour from its sensors and actuators. Thus the on-board decision making would be performed based on the current state of the vehicle and its priority level could be changed depending on the significance of the recorded problem. Once the central computer is aware of the issue, it can reassign the other agents to account for the vehicle in trouble.

The integrity of the swarm, the safety of its direct environment and the success of the mission will always prevail on the individual. In case of a minor issue however, the central intelligence can land the agent and warn the operator that it needs to be fixed or replaced.

## VI. Conclusion

In this paper we presented a set of algorithms grouped into a single method in order to control the behaviour of a multi-agent system. The work was first restricted to a two dimensional scenario to focus on the versatility of the target assignment and its real time capabilities. All the scenarios implemented were successfully analysed by the task manager and properly carried out by the vehicles. The algorithms were then extended to the third dimension and optimized to account for the unstable quadrotor dynamics. Significant modifications on the pure pursuit method enabled a smooth and safe trajectory in a dense environment. The path tracking algorithm will be further improved to incorporate a DRCA-inspired conflict resolution and avoidance protocols in partially known environments.

## References

[1] E. Saad, J. Vian, G. J. Clark, and S. Bieniawski, *Vehicle Swarm Rapid Prototyping Testbed*. The Boeing Company, Seattle, WA, 98124.

[2] G. M. Hoffmann, H. Huang, S. L. Waslander, C. J. Tomlin, *Precision Flight Control for A Multi-Vehicle Quadrotor Helicopter Testbed*. 2011.

[3] Khayyam, *Recherche de chemin par l'algorithme A**. 2008.

[4] A. Tsourdos, *Cooperative Path, Planning of Unmanned Aerial Vehicles*. 2011.

[5] D. P. Bertsekas, *Auction algorithms for network flow problems*. Computational Optimization and Applications. 1992.

[6] M. Becker, C. Dantas and W. Macedo, *Obstacle Avoidance Procedure for Mobile Robots*. Mechatronics, vol. 2. 2006.

[7] T. Bresciani, *Modelling , Identification and Control of a Quadrotor Helicopter*. 2008.

[8] S. Bouabdallah, *Design and Control of Quadrotors with Application to Autonomous Flying*. PhD thesis, 2007.

[9] E. Lalish, K. A. Morgansen, *Distributed reactive collision avoidance*. Autonomous Robots, special issue, 2012.

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# Development of an autopilot system for rapid prototyping of high level control algorithms

Matthew Coombes, Owen McAree, Wen-Hua Chen, Peter Render
Department of Automotive and Aeronautical engineering
Loughborough University
Loughborough, LE11 3TQ UK
Email: {ttmjc2@lboro.ac.uk, ttom@lboro.ac.uk, w.chen@lboro.ac.uk, p.m.render@lboro.ac.uk}

***Abstract*—This paper describes the development of a system for the rapid prototyping of high level control algorithms using an Arduino based commercial off the shelf autopilot called ArduPilot. It is capable of controlling multiple vehicle types, including fixed, and rotary wing aircraft as well as ground vehicles. The inner loop control is performed by ArduPilot, so the high level control can be rapidly prototyped and tested in Simulink, or an embedded system. The ability to conduct tests in software and hardware in the loop has also be developed, to enable safe testing of algorithms, which will speed up the development process. To show its functionality and ability to assist with the development process of algorithms, ArduPilot is used with a remote controlled aircraft in simulation and in real world testing to verify newly developed high level algorithms for UAVs.**

***Index Terms*—Autopilot; rapid prototyping; algorithm development; hardware in the loop; software in the loop.**

## I. INTRODUCTION

It is important to be able to test high level algorithms, in simulation first, and then eventually in the real world. To be able to test high level control, a system needs to be in place for this to be tested on. The advantage of performing Software In the Loop (SIL) and then Hardware In the Loop (HIL) testing before performing real world testing is widely recognized as much cheaper, less time consuming, and safer [1]. A system that allows a seamless transition from one stage to the next without reprogramming, or reconfiguring any part of the system would hugely reduce the time from conception to the final design. This paper describes a system for developing high level autonomous control functions of UAVs.

Developing a full autopilot system from scratch is difficult, as specialist knowledge is needed in electronic engineering, systems integration, and software engineering. By purchasing an open source Commercial Off The Shelf (COTS) autopilot system much time and effort was saved. ArduPilot [2] is a hobbyist “do-it-yourself” autopilot meant for use on remote controlled aircraft. It is capable of flying simple waypoints, or more importantly taking roll ($\phi$), pitch ($\theta$), rudder, and throttle commands from an external source. By using a communication protocol called MAVLink [3], any device, or system can give ArduPilot these commands over a number of different connection types. As the low level control is performed by this autopilot, it leaves the user the freedom to concentrate on the high level control.

This paper describes the ArduPilot autopilot, and the system developed to control it. SIL, and HIL testing is discussed, as well as the X-Plane simulation environment used in this testing. The use of Simulink, and embedded systems to perform high level control is described. An example is given to illustrate the full development cycle from SIL to HIL to real world testing, where a simple PID waypoint tracking algorithm is developed and tested safely and easily.

## II. ARDUPILOT

ArduPilot is a Arduino [4] based autopilot designed for the use by remote controlled aircraft hobbyists. It is designed to control fixed wing aircraft, and various rotary wing platforms, including single, tri, quad, hexa, and octa copters. The software is written in C++, and is completely open source with an active development community. The system enables the user to fly a series of predefined waypoints using a simple cross track error trajectory following algorithm. Or the user can fly the aircraft on a Fly by Wire (FBW) mode, where pitch and roll angles are commanded over the transmitter using ArduPilots inner loop instead of directly commanding servos. ArduPilot can communicate with a ground control station, where data can be gathered, waypoints, or even control gains can be updated. It communicates over a wireless serial connection, using a communication protocol called MAVLink [**?**], which was designed specifically for Micro Aerial Vehicles (MAVs).

The system hardware consists of two circuit boards shown in Fig. 1. The lower board has an ATmega2560 processor which runs the software, a failsafe multiplexer (that means there is always a hardware override if the processor fails) and all servo and receiver connections. The upper board houses the sensors, and telemetry ports. The sensors include a triple axis accelerometer, a dual and single axis gyro, barometric pressure sensor, triple axis magnetometer, also an externally connected GPS module. It is also capable of having an external pitot static sensor, for airspeed measurements. The full system on an aircraft set up for flight testing is shown in Fig. 2.

The ATmega2560 processor is relatively slow and only capable of executing 256 Kb of code. The autopilot software is written in C++ which is not that accessible a programming language. In its default state it is not an effective system for research.

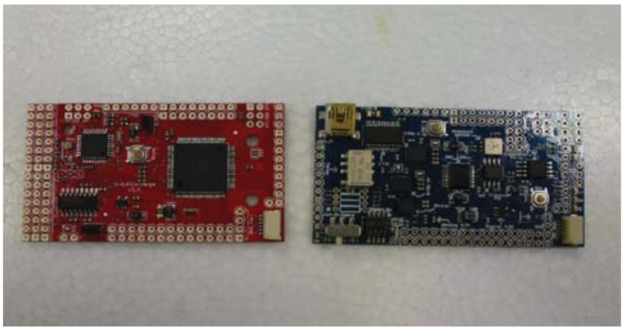


Fig. 1. Left: ArduPilot ATmega2560 processor board Right: IMU board

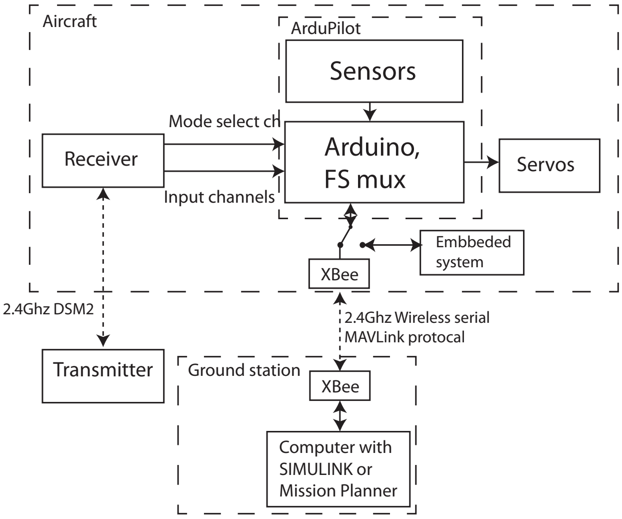


Fig. 2. Flow diagram for ArduPliot on and aircraft ready for flight testing

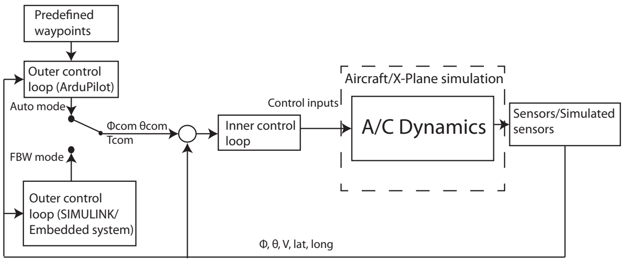

Predefined waypoints
Outer control loop (ArduPilot)
Auto mode
Φcom θcom
Tcom
FBW mode
Outer control loop (SIMULINK/ Embedded system)
Inner control loop
Control inputs
Aircraft/X-Plane simulation
A/C Dynamics
Sensors/Simulated sensors
Φ, θ, V, lat, long

Fig. 3. Flow diagram of the control structure on the whole system

There are many examples of autopilot systems developed by research organisations, to assist with their research goals. They are often made in house at considerable expense, and many run on embedded systems making for high development times [1], [5], [6], and [7]. There are other autopilots available, like Paparazzi [8], or MicroPilot [9]. These systems are expensive, they are not easy to manipulate and interface with, and by no means as cross platform compatible. Paparrazi has no standard hardware as the schematics are open source [10], requiring one to be specially ordered, or made in house. MicroPilot is a very small open architecture autopilot, which has an integrated Inertial Measurement Unit (IMU) and GPS. Although all the parameters and control gains can be altered, the code is closed source and can not be altered. An autopilot system meant for hobbyists is Attopilot [11], this system is quite expensive and is quite dated. AttoPilot uses thermopiles for attitude control in instead of the much more accurate and reliable IMU [12].

## III. System Configuration

ArduPilot has a number of different operating modes. The outer loop waypoint following mode (auto), the inner loop pitch, and roll angle hold (FBW), and manual override. If roll and pitch commands are sent to ArduPilot from an external source while in FBW mode, the higher level control loop can be performed on any platform of the users choice that is able to communicate over a serial connection. This gives huge flexibility to the user. As the ATmega2560 is simply not fast enough, and does not have enough storage, an external system to run more advance, more demanding code is needed.

The external system that this system has been primarily designed for is Simulink. A program that is used and understood by many engineers. It enables rapid prototyping of algorithms, and has an accessibility that is simply not offered when doing the same on an embedded system where the user must be familiar with C or C++.

The layout of the control system is shown in Fig. 3. ArduPilot's Attitude Heading Refrence System (AHRS) conducts sensor fusion and transmits the data over its serial telemetry port encoded in MAVLink format. A Simulink block has been developed for communication over MAVLink, it receives and decodes MAVLink messages, Simulink conducts high level control on the block outputs, to give pitch, roll angle, and throttle commands. The block then encodes these commands and resends them over the same serial connection. To enable wireless serial communication for flight tests, XBee wireless modules are used. As data and control signals are being transmitted wirelessly, this system relies on the wireless link to have high data integrity and low latency.

An embedded system can be used in place of Simulink if the wireless data link is a concern. By simply using the MAVLink wrapper discussed in Section V an embedded system can do exactly the same job as Simulink by directly connecting its serial port to that of ArduPilot's. This enables all the control to be performed on board the aircraft but over two components, mitigating XBee signal issues, also enabling a higher data rate.

## IV. Software and hardware in the loop

Before performing real world testing, it is important to verify code functionality, which is commonly done through SIL, and then HIL testing. Putting algorithms through the full development cycle of SIL, HIL, and then flight testing is a good systematic debugging method, which significantly reduces risk. This process can be shown in Fig. 4. SIL enables any bugs in the software, to be ironed out, then HIL brings to light bugs which come about due to the software's interaction with the hardware, and the wireless communication

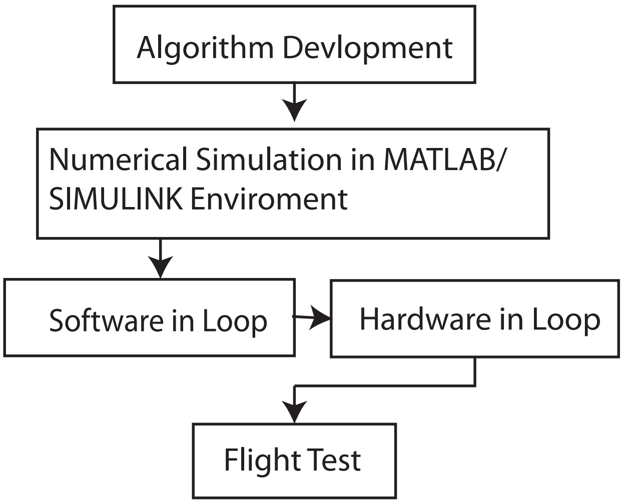


Fig. 4. The development cycle for high level control for UAS

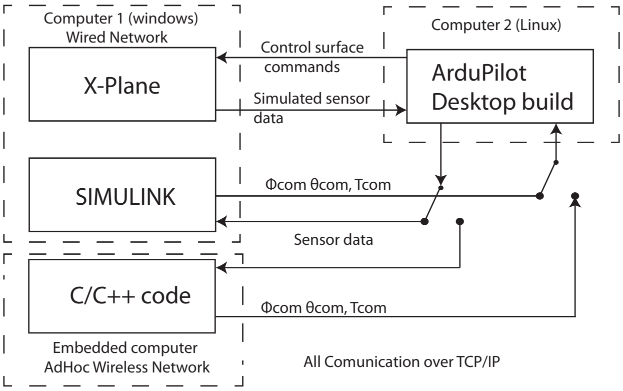


Fig. 5. Flow diagram for ArduPliots SIL testing, use with Simulink or and embedded system

system. Finally fight testing shows the functionality of the communications system at range, and issues the system has in the real environment, like wind.

The SIL system is shown in Fig. 5. The open source ArduPilot code has a desktop build, which means that, instead of the code being compiled on the Arduino on the ArduPilot, it is complied on a normal desktop computer running LINUX. This simulated ArduPilot communicates with a simulated aircraft in the X-Plane flight simulator using a plugin that enables communication over a TCP/IP network connection using the MAVLink Protocol. The X-Plane aircraft model is controlled by the desktop build of ArduPilot, and in turn ArduPilot is controlled by an external system like Simulink or an embedded system running the users high level control algorithms. This SIL method enables development to be undertaken without any ArduPilot hardware.

HIL (shown in Fig. 6) is much the same, but the actual ArduPliot hardware is used and sensor data is faked. The communication between X-Plane and ArduPilot is now done over a virtual serial connection, provided by the USB to serial chip. ArduPilot is now controlled over its telemetry port connected to Simulink, or a ground station using a XBee wireless serial transmitter, or to an embedded system directly.

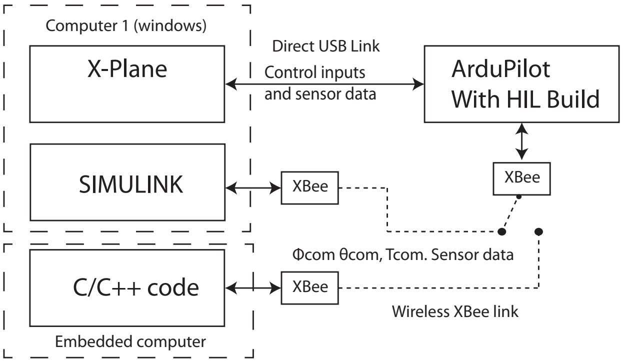


Fig. 6. Flow diagram for ArduPliots HIL testing, use with Simulink or and embedded system

When it comes time to conduct actual flight tests there is no need for reconfiguration of software hardware or communications, apart from ArduPilot needs to be mounted to the aircraft, and wired to the receiver and servos. As ArduPilot and MAVLink have been abstracted to such a high level, throughout the whole development cycle none of the ArduPilot code, or high level control ran externally needs to change at all to move to the next stage. All that changes is if the aircraft is real or simulated, where ArduPilot software is ran, and how the components of the system physically communicate with one another.

## V. Communication

As has been mentioned previously, the communications between all components of the system use a protocol called MAVLink. MAVLink is a very lightweight, header-only message marshalling library for MAVs, in C/C++. It encodes data structures into high efficiency data packets which use binary instead of ASCII encoding, yielding faster data transfer and higher data integrity. Any device that can communicate in MAVLink can talk to ArduPilot. A C++ wrapper has been developed that abstracts MAVLink so it can communicate over any physical transport layer, currently available is communication over serial, TCP/IP, UDP, and Write to file. This has facilitated the seamless transition in the testing phase from SIL through HIL to flight testing, by simply having each component change the physical means of sending MAVLink encoded data.

XBee converts a serial stream to wireless using a protocol called Zigbee. A large outdoor 15 dB omni directional 2.4Ghz antenna is used on the ground, and a striped down 8 dB antenna on the aircraft. This means that the small 60 mW power of the XBee can communicate with an aircraft a mile away in any direction and orientation. The series 2.5 XBees used are point to multi point in a communication mesh. A master module is attached to the ground station receiving data from as many as 255 other XBees. Due to the high gain

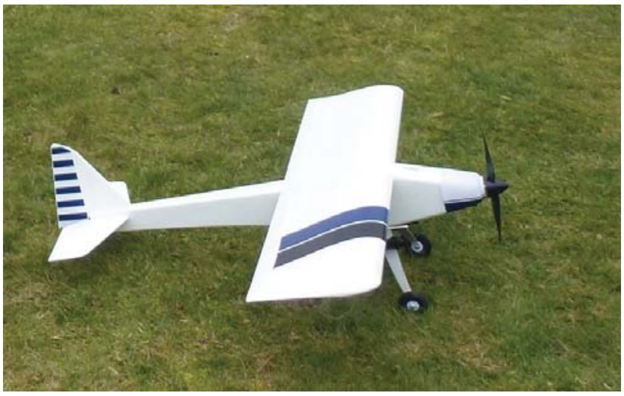


Fig. 7. WOT4 testing aircraft

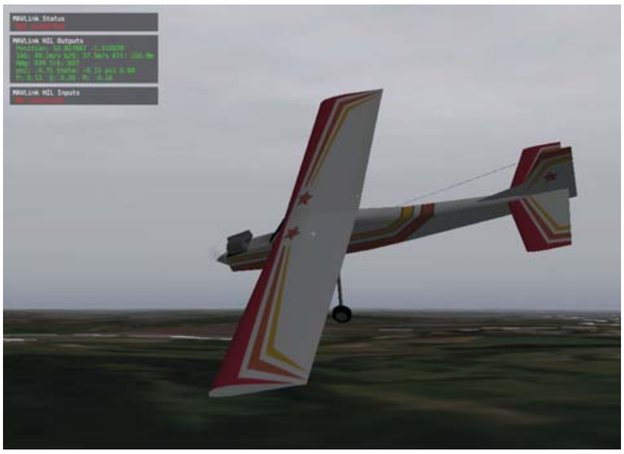


Fig. 8. X-Plane RC Plane model of WOT4

antennas, and the meshing ability of XBee 2.5 multi vehicle test are possible.

## VI. Experimental Studies

The platform chosen for the initial experiments is a WOT4 model aircraft. The WOT4 is an inexpensive option for the initial flight tests. This aircraft has been chosen as it is large enough to carry Ardupilot but small enough to keep risk at an acceptable level during the initial flight testing. It is made from Expanded Poly Olefin (EPO) foam with is extremely strong, durable, and very light. The WOT4 is capable of carrying of 500g payload, so would be capable of carrying a camera and video transmitter equipment. Ardupilot is easily able to fit inside the fuselage, close to the center of gravity of the aircraft, and out of the airflow to minimise drag. A pitot static probe, and sensor is mounted half way along the wing out of prop wash, to measure the airspeed of the aircraft. The WOT4 is shown in Fig. 7.

For SIL and HIL, an X-Plane model of the WOT4 has been developed, so algorithms could be tested on a representative model. The model was developed using X-Plane's plane maker, using the aircrafts dimensions, power, and wing cross section, to give an approximate representation of the aircraft. A graphical representation of the model can be seen in Fig. 8.

To test the functionality of the full system from SIL to real world testing, step tests are performed as well as tests on speed hold, and altitude hold controllers. Finally a simple PID waypoint tracking algorithm is performed.

Using the MAVLink Simulink block aircraft state information can be read, and roll, pitch angle, rudder and throttle commands sent to and from ArduPilot. The speed, altitude hold, and waypoint tracking are also programed in Simulink.

Illustrated in Fig. 9 is the step response of the WOT4 in roll, in SIL, HIL, and real world. From flying straight and level a 60° roll angle is commanded, then a -60 ° angle is commanded. The response in SIL and HIL were of course similar as they use the same model in X-Plane, the real world response only differed slightly. This simple test shows the fundamentals of the system functioning, including Ardupilots inner loop control, MAVLink, and communications downlink functionality.

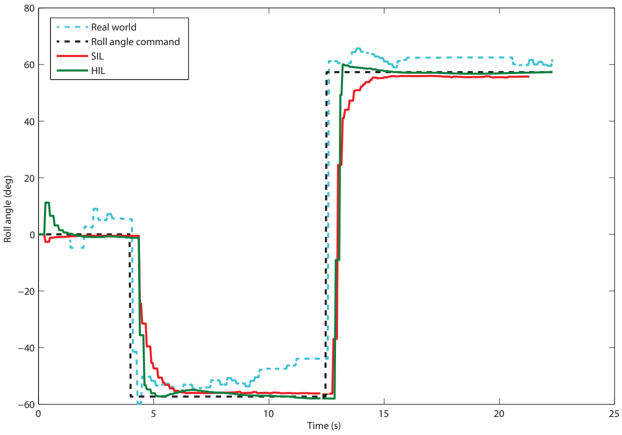


Fig. 9. Step response for roll angle for the WOT4, in SIL, HIL, and real world

To demonstrate the ability of Simulink to perform outer loop control, both speed hold, and heading hold control are tested. Also they are needed for more advanced outer loop control. Both are based on simple PID controllers, where speed is controlled with pitch angle, and heading is controlled with roll angle. Once again this was put through the whole development cycle. Fig. 10 shows the aircraft responding to step commands in airspeed. Fig. 11 shows the aircraft responding to step commands in heading. As good data integrity, and speed is important in off board control, these tests show that Xbee, and Simulink are capable of this. As wireless communication is the weak link in this system and as simple off board control functions perfectly, there is no reason that the system could not now perform more complex control algorithms.

A simple PID waypoint tracking algorithm was implemented in Simulink, to make an aircraft fly around user defined waypoints. It uses the haversign formula [13] to calculate the heading between the aircraft's, and waypoint's latitude and

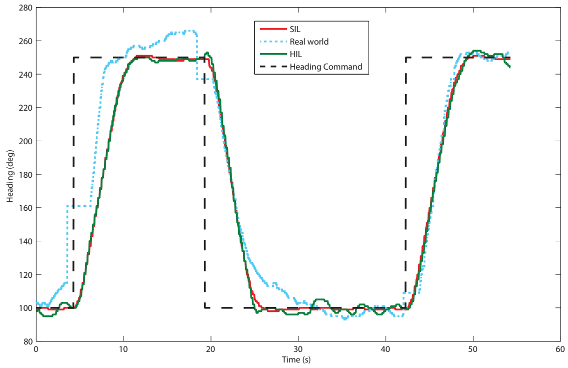


Fig. 10. Step response for heading hold controller for the WOT4, in SIL, HIL, and real world

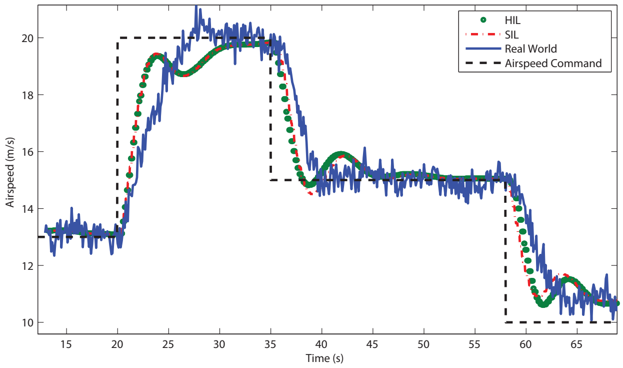


Fig. 11. Step response for speed hold controller for the WOT4, in SIL, HIL, and real world

longitude. Eq. (1), and Eq. (2) shows how the bearing ($\theta$) and the distance (d) between two points on the earth surface defined by latitude and longitude can be calculated.

$$a = s^2(\frac{\delta lat}{2}) + c(lat_1)c(lat_2)s^2(\frac{\delta lng}{2})$$

$$c = 2\arctan(\frac{\sqrt{a}}{\sqrt{1-a}}) \qquad (1)$$

$$d = R_e c$$

$$\theta = \arctan(\frac{s(\delta lng)c(lat_2)}{c(lat_1)s(lat_2) - s(lat_1)c(lat_2)c(\delta lng)}) \qquad (2)$$

Where cos and sin are abbreviated to c and s respectively. Where $R_e$ is the radius of the earth which is 6378.1 Km, and $\delta lat$, and $\delta lng$ is difference between the aircraft's latitude and longitude and that of the origin or the waypoint the distance and heading is to be measured. $lat_2$, and $lng_2$ are the latitude, and longitude of the aircraft, and $lat_1$, and $lng_1$ are the latitude, and longitude of the next waypoint.

This bearing is then used as the heading command to make the aircraft fly directly at the next waypoint. The aircraft is flown around a sqaure circuit at 200 m height above ground, and at 15 m/s. Fig. 12 shows the 2D path flown by the WOT4 in SIL HIL, and in real world testing. It is in this flight test that the greatest difference between the simulated WOT4 and the real WOT4 is observed, but the simulated tests proved that the algorithms worked successfully.

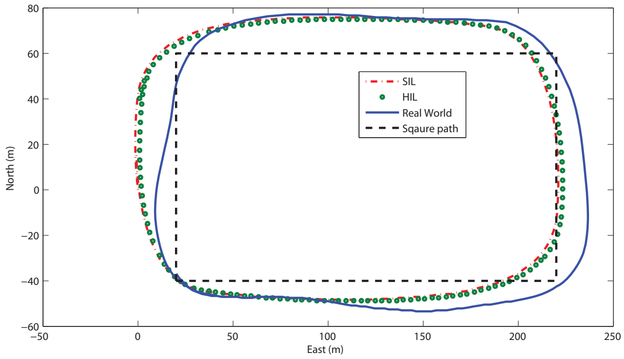


Fig. 12. Simple PID waypoint tracking algorithm tracking a 100x200m square path, in SIL, HIL, and real world

## VII. Conclusion

A system has been developed that enables research in the field of UAS to quickly develop and test high level control algorithms. By using a COTS autopilot, the full system is cheap and easy to integrate into a range of vehicles. As ArduPilot is a well developed system it is extremely reliable, and robust which makes flight testing swift and safe. The results from SIL, HIL testing have with relative accuracy, predicted the performance of the WOT4 in real world tests. The control system developed in Simulink needed no modification between simulation and real world tests. These show that this is a good technique for safely testing algorithms in simulation before moving the to actual vehicles. As the high level control was carried out off board, on a ground control station, and relied on a wireless data link occasionally there were a few data packets drops. The aircraft never got further than 300m away so if operating at further distances, or at higher data rates, an embedded system can be used to conduct the high level control with the same ease.

## References

[1] G. Cai, B. M. Chen, T. H. Lee, and M. Dong, "Design and implementation of a hardware-in-the-loop simulation system for small-scale uav helicopters," *Mechatronics*, vol. 19, no. 7, pp. 1057 – 1066, 2009.

[2] [Online]. Available: http://diydrones.com/notes/ArduPilot

[3] [Online]. Available: http://qgroundcontrol.org/mavlink/start

[4] [Online]. Available: http://www.arduino.cc/

[5] D. Kingston, R. Beard, A. Beard, T. McLain, M. Larsen, and W. Ren, "Autonomous vehicle technologies for small fixed wing uavs," in *AIAA Journal of Aerospace Computing, Information, and Communication*, 2003, pp. 2003–6559.

[6] Y. C. Paw and G. J. Balas, "Development and application of an integrated framework for small uav flight control development," *Mechatronics*, vol. 21, pp. 789 – 802, 2011.

[7] A. Mehta and K. Pister, "Warpwing: A complete open source control platform for miniature robots," in *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, oct. 2010, pp. 5169 –5174.

[8] [Online]. Available: http://paparazzi.enac.fr/wiki

[9] [Online]. Available: http://www.micropilot.com/

[10] M. G. P.-S. H. Pascal Brisset, Antoine Drouin and J. Tyler. (2006, October) The paparazzi solution. ENAC.

[11] [Online]. Available: http://www.attopilotinternational.com/

[12] M. A. O. Simes, "Development of an aerial robot for inspection and surveillance," in *Mestrado em Engenharia Mecnica*. Universidade de Aveiro, 2009.

[13] R. W. Sinnott, "Virtues of the haversine," in *Sky and Telescope*, vol. 68, Dec 1987, p. 158.

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# Resource Allocation with Cooperative Path Planning for Multiple UAVs

Hyo-Sang Shin,Cédric Leboucher, and Antonios Tsourdos
Department of Informatics and Systems Engineering
Cranfield University, Defence Academy of the UK
Swindon SN6 8LA, UK
Email: h.shin; c.leboucher; a.tsourdos@cranfield.ac.uk

***Abstract*—This study proposes an optimal resource allocation algorithm of multiple UAVs with cooperative path planning using a geometric approach. The focus of the resource allocation is on mission and task completion, also known as feasibility whilst coping with operational and physical constraints of UAVs. Therefore, this study first introduces a geometric path planning algorithm based on Pythagorean Hodgraphs (PH). Using Bernstein Bzier polynomials, the path planning algorithm can generate feasible and safe (obstacle and inter-collision free) paths which can also meet position and orientation constraints of UAVs. We then optimise the resource allocation based on Evolutionary Game Particle Swarm Optimisation (EGPSO) and paths generated by the geometric planning. The input parameter of the optimal allocation problem is the allocation policy and the performance index is chosen to be the total flight time of the UAVs. Here the flight time is computed from the path produced by the path planning algorithm. The optimal allocation algorithm changes the allocation policy and finds the best allocation policy which minimise the performance index. The performance of the proposed algorithm is investigated by numerical examples simulated under realistic scenarios.**

## I. Introduction

Inexpensive unmanned aerial vehicles (UAVs) have considerable potential for use in remote sensing operations. They are cheaper and more versatile than manned vehicles, and are ideally suited for dangerous, long and/or monotonous missions that would be inadvisable or impossible for a human pilot. Groups of UAVs are of special interest due to their ability to coordinate simultaneous coverage of large areas, or cooperate to achieve common goals. Specific applications under consideration for groups of co-operating UAVs include, but not limited, border patrol, search and rescue, surveillance, mapping and environmental monitoring. In these applications, the group of UAVs becomes a mobile resource/sensor and consequently routes and tasks for each UAV need to be properly and optimally assigned in order to cooperatively achieve their mission. Therefore, this study addresses the vehicle routing problem of of multiple UAVs.

The vehicle routing problem, has been mainly handled in the operational research area ([1], [2], [3], [4]) and can be generally classified by two categories: one is the Traveling Salesman Problem (TSP) which finds a shortest circular trip through a given number of cities, and the other is the Chinese Postman Problem (CPP) finding the shortest path with considering path constraints on an entire network of road. The TSP using multiple UAVs can be considered as a task assignment problem to minimise the cost of time or energy for a certain mission by assigning each target to an UAV, for which binary linear programming ([5]), iterative network flow ([6]), tabu search algorithm ([7]) and receding horizon control ([8]) have been proposed. Recently, [9] proposed a route optimisation algorithm for multiple searchers to detect one or more probabilistically moving targets incorporating other factor such as environmental and platform-specific effects. Meanwhile, the CPP is normally used for ground vehicle applications such as road maintenance, snow disposal ([10]), boundary coverage ([11]), and graph searching and sweeping ([12], [13]). Since the general vehicle routing algorithms approximate their path to a straight line shape to reduce computational load, the physical constraints imposed on the vehicle are not to be addressed.

In order to mitigate this issue, this study divides the routing problem into two parts: the first part is to design cooperative path planning and the second one is to find the optimal resource allocation policy based on the paths obtained in the first part. Cooperative path planning algorithm is designed using the differential geometry concepts, especially Pythagorean Hodographs (PH) curves, which was proposed in our previous study [14]. Path planning algorithms based on differential geometry examine the evolution of guidance geometry over time to derive curvature satisfying the guidance goals. Guidance command such as a manoeuvre profile can be then computed using the derived curvature of the guidance geometry. One of main advantages of this approach is that the number of design parameters can be significantly reduced whilst maintaining the guidance performance. Therefore, this approach will enable us not only to design fast and more lightweight algorithms, but also to generate safe and feasible paths for multiple UAVs. This would be preferred for integration of path planning with the optimal resource allocation. Since reaching targets at the same instant with specific orientations could improve the overall effectiveness and survivability of the UAVs, simultaneous arrivals with predefined orientations are considered as constraints with the physical ones such as obstacle avoidance and the maximum turning rate of of the UAVs.

The performance index for the optimal resource allocation problem is the total flight time of the multiple UAVs since

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

this is necessary to have a manageable task in the available time. The total flight time for each candidate allocation is computed by using the velocity profile and paths generated by cooperative path planning and this is used to find the optimal allocation policy. The optimisation method implemented in this paper is Evolutionary Game Particle Swarm Optimisation (EGPSO) which is proposed in our previous study [15]. The proposed EGPSO algorithm integrates the Evolutionary Game Theory (EGT) concepts with those of Particle Swarm Optimisation to find the optimal weight of the coefficients considering the entire swarm fitness. Moreover, it is shown that this algorithm efficiently works in the general allocation problem [15].

The overall structure of this paper is given as: Section II briefly introduces a target tracking filter design, trajectory classification to model the behaviour of ground vehicles, and behaviour recognition algorithm using string matching theory. Section III introduces rule-based decision making algorithm to find suspicious or anomalous behaviour based on a fuzzy logic. Section IV presents numerical simulation results of behaviour monitoring for both military and civilian traffic scenario using realistic ground vehicle trajectory data. Lastly, conclusions and future works are addressed in Section V.

## II. Problem Formulation

### A. Scenario

The scenario considered in this study is similar to that in our previous paper [14]. In the scenario, it is assumed that a group of $N$ UAVs leaves from a base and they have to reach the target area at the same time with predetermined orientations. The individual start and finish points for each UAV are represented by position coordinates $(x, y)$ and orientation by angle $\theta$. These are assumed to be known a priori. The UAVs are assumed to be of same type and are flying at same the speed at constant altitude. Each UAV has the same maximum bound on its curvature and the environment has static obstacles. The UAVs are required to avoid collision with other UAVs and with other objects in the air-space, as well as avoiding the static obstacles. The allocation of the UAVs needs to be optimised.

### B. Optimal Allocation Problem

Let us first consider a path between a single UAV from the base to the target position with no constraints. The starting point $P_s$ is at the base position and the finishing point $P_f$ is at the target position. The path connecting the poses is represented by the label $r$. The path planner produces a path connecting the start pose $P_s(x_s, y_s, \theta_s)$ and the finish pose $P_f(x_f, y_f, \theta_f)$.

$$P_s(x_s, y_s, \theta_s) \xrightarrow{r(t)} P_f(x_f, y_f, \theta_f) \tag{1}$$

where $t$ is a path length parameter.

Extending equation (1) to account for a group of $N$ UAVs gives:

$$P_{si}(x_{si}, y_{si}, \theta_{si}) \xrightarrow{r_i(t)} P_{fi}(x_{fi}, y_{fi}, \theta_{fi}), \tag{2}$$

$$\max|\kappa_i| < \kappa_{max}, \coprod_{safe}, \coprod_{length}, i = 1 \dots N, \tag{3}$$

where $\kappa$ is the path curvature, $\kappa_{max}$ is the maximum curvature bound obtained from the maximum turning rate, and $\coprod_{\text{safe}}$ and $\coprod_{\text{length}}$ are the constraints on safety and path length respectively.

The safety constraints are described as:

$$\begin{aligned} \coprod_{length} \quad : \quad & d(P_i(t), P_j(t)) > d_{sep}, \\ & d(P_i(t), O_k(t)) > d_{sep} \\ & i \neq j = 1 \dots N, k = 1, \dots, n_o \end{aligned} \tag{4}$$

where $P_i(t)$ and $O_k(t)$ are the positions of the $ith$ UAV and the $kth$ obstacle, $n_o$ the number of obstacles, and $d_{sep}$ the minimum separation distance. To enable simultaneous arrivals, the constraint on length is given as:

$$\begin{aligned} \coprod_{length} : s_i(t_{fi}) &= s_{cm}, i = 1 \dots N \\ s_i(t_{fi}) &= \int_{t_{si}}^{t_{fi}} \sqrt{\dot{x}_i(t)^2 + \dot{y}_i(t)^2} dt, \end{aligned} \tag{5}$$

where $s_{cm}$ is a common path length which is automatically obtained regarding to the allocation.

The optimal allocation problem is then formulated as: minimising the following performance index

$$J = \Sigma_1^N \int_{t_{si}}^{t_{fi}} dt \tag{6}$$

subject to equation (2) and (3).

## III. Cooperative Path Planning Using Pythagorean Hodographs Curves

One of well known path planning approaches based on the differential geometry concepts is Dubins path planning [14], [16]. The Dubins trajectory ([17]) is the shortest path connecting two configurations represented by position and pose under the constraints of a bound on curvature or turning radius. The Dubins path is a composite curve of both lines and circles and is easy to produce. However, it lacks a smooth variation of curvature. Mathematically, the Dubins path provides only tangent continuity, $C^1$. The curvature continuity is important as the curvature is proportional to the lateral acceleration of the UAV. Therefore, curvature discontinuity results in an abrupt maneouvre of the UAV. A smooth motion needs curvature continuity $C^2$. Therefore, it is necessary to seek for an alternate path with curvature continuity. The equation of curvature is:

$$\kappa(t) = \frac{\dot{\mathbf{r}} \times \ddot{\mathbf{r}}}{|\dot{\mathbf{r}}|^3}, \quad \dot{\mathbf{r}} = \frac{d\mathbf{r}}{dt}, \ddot{\mathbf{r}} = \frac{d^2\mathbf{r}}{dt^2} \tag{7}$$

From the equation (7), the curvature is a function of first two derivatives of a curve, $\mathbf{r}(t)$, so the path needs to be at least twice continuously differentiable, that is $C^2$ continuity. There are many polynomial curves which can provide $C^2$ continuity. However, we choose Pythagorean Hodograph (PH) curve known for its rational properties.

A smoother curve can be produced by using techniques such as PH, where basis curves are used to piece together using Bernstein Bézier polynomials. For a planar parametric curve, $\mathbf{r}(t) = \{x(t), y(t)\}$, the hodographs are $\dot{x}(t)$ and $\dot{y}(t)$ so the velocity vectors of a curve are its hodograph. The path-length of the curve $\mathbf{r}(t)$ is:

$$s = \int_{t_1}^{t_2} \sqrt{\dot{x}^2 + \dot{y}^2} dt = \int_{t_1}^{t_2} |\dot{\mathbf{r}}(t)| dt \tag{8}$$

where $s$ is the path-length, $t$ is a parameter such that $t \in [t_1, t_2]$ and $\dot{x} = \frac{dx}{dt}$, $\dot{y} = \frac{dy}{dt}$, and $\dot{r} = \frac{dr}{dt}$ . The term inside the square root in equation (8) is the sum of the square of the hodographs of the curve, $\mathbf{r}(t)$. For the PH path the path-length is an integral of a polynomial $\sigma(t)$ such that:

$$\begin{aligned} \sigma(t) &= \sqrt{\dot{x}^2 + \dot{y}^2} \\ s &= \int_{t_1}^{t_2} |\sigma(t)| dt \end{aligned} \tag{9}$$

The PH curve can be produced by selecting two Bernstein polynomials $u(t)$ and $v(t)$ such that:

$$\begin{aligned} \dot{x}(t) &= u^2(t) - v^2(t) && (10) \\ \dot{y}(t) &= 2u(t)v(t) && (11) \\ & && (12) \end{aligned}$$

This gives:

$$|\sigma(t)| = u^2(t) + v^2(t) \tag{13}$$

Note that the PH path thus provides exact calculation of path length and it's curvature, as well as the orders of the two polynomials, $u(t)$ and $v(t)$, determine the order of PH curves. In this research, we use a fifth order PH curve as this is the lowest order curve which has inflexion points which can provide sufficient flexibility [**?**]. For a fifth order PH curve, $u(t)$ and $v(t)$ can be approximated as second order polynomials:

$$\begin{aligned} u(t) &= \sum_{k_u=0}^{2} u_k \binom{2}{k_u} (1-t)^{(2-k_u)} t^{k_u}; && (14) \\ v(t) &= \sum_{k_v=0}^{2} b_k \binom{2}{k_v} (1-t)^{(2-k_v)} t^{k_v}; && (15) \end{aligned}$$

Hence the curve, $\mathbf{r}(t)$, is given by Berstein form:

$$\mathbf{r}(t) = \sum_{k=0}^{5} P_k \binom{5}{k} (1-t)^{(5-k)} t^k; \tag{16}$$

where $P_k(x_k, y_k),\ k = 0, 1, 2, 3, 4, 5$ are control points. Note that these control points determine the curve $\mathbf{r}(t)$ and can be derived as:

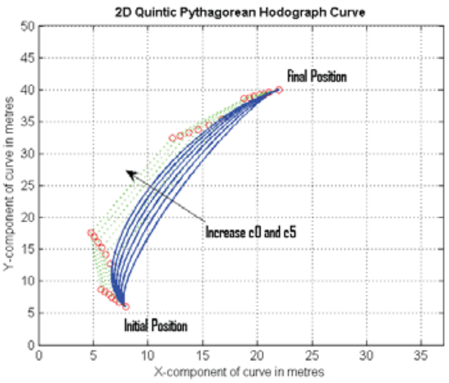


Fig. 1. PH paths

$$P_1 = P_0 + \frac{1}{5}\begin{pmatrix} u_0^2 - v_0^2 \\ 2u_0v_0 \end{pmatrix} \tag{17a}$$

$$P_2 = P_1 + \frac{1}{5}\begin{pmatrix} u_0u_1 - v_0v_1 \\ u_0v_1 + u_1v_0 \end{pmatrix} \tag{17b}$$

$$P_3 = P_2 + \frac{1}{5}\begin{pmatrix} u_1^2 - v_1^2 \\ 2u_1v_1 \end{pmatrix} + \frac{1}{15}\begin{pmatrix} u_0u_2 - v_0v_2 \\ u_0v_2 + u_2v_0 \end{pmatrix} \tag{17c}$$

$$P_4 = P_3 + \frac{1}{5}\begin{pmatrix} u_1u_2 - v_1v_2 \\ u_1v_2 + u_2v_1 \end{pmatrix} \tag{17d}$$

$$P_5 = P_4 + \frac{1}{5}\begin{pmatrix} u_2^2 - v_2^2 \\ 2u_2v_2 \end{pmatrix} \tag{17e}$$

When the initial and final configurations (pose and heading) for each vehicles are known, $P_k(x_k, y_k),\ k = 0, 1, 2, 3, 4, 5$ can be specified. From the configurations, $P_0$, $P_1$, $P_4$, $P_5$ are directly obtained as:

$$\begin{aligned} P_0 &= (x_s, y_s) && (18a) \\ P_5 &= (x_f, y_f) && (18b) \\ P_1 &= P_0 + (1/5) * d_0 && (18c) \\ P_4 &= P_5 - (1/5) * d_5 && (18d) \end{aligned}$$

where

$$\begin{aligned} d_0 &= c_0(\cos(\theta_s), \sin(\theta_s)) && (19) \\ d_5 &= c_5(\cos(\theta_f), \sin(\theta_f)) && (20) \end{aligned}$$

where $c_0 \in (0, \infty]$ and $c_5 \in (0, \infty]$. Note that when $c_0$ and $c_5$ are specified, the control points $(P_0, P_1, P_4, P_5)$ in (18) are fixed by configuration. Moreover, $u_i$ and $v_i$ for $i = 1, 2, 3$ are also uniquely derived from equation (17), which implies that $P_2$ and $P_3$ are also fixed. Therefore, the number of control parameters reduces to two of $c_0$ and $c_5$ whilst maintaining the continuity of the curve. Increasing the values of $c_0$ and $c_5$ will increase the length of tangent vectors $P_0 = |\boldsymbol{P_0P_1}|$ and $P_5 = |\boldsymbol{P_5P_4}|$ and in turn $P_2$ and $P_3$ get shifted to meet the PH condition (17). This is shown in figure 1. As shown in this figure, varying the two control parameters controls the curvature, which in turn will determine the space curve.

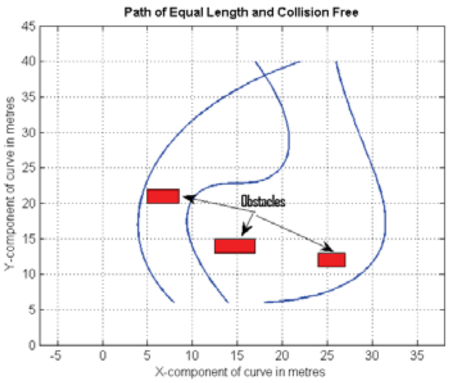


Fig. 2. PH paths, all equal in length and avoiding obstacles

For simultaneous arrival, the paths are required to be made equal in path if UAV speeds are constant and same. The variable speed UAVs can have difference in path lengths. This is achieved by increasing the shorter paths to that of the longest one. The path lengths of the flyable and safe paths are calculated using (8). For $N$ number of UAVs, with the length of each path $s_i$, the set of path lengths $\Sigma$ is:

$$\Sigma = \{s_i\}, \quad i = 1, \dots, N \tag{21}$$

The longest of the safe flyable path is the reference path. That is the maximum of $\Sigma$. The path lengths of $(N-1)$ UAVs are increased to that of the reference path. Lengths of the PH path is increased by changing the control parameters of $c_0$ and $c_5$:

$$\text{Find } c_0 \text{ and } c_5\text{, such that} \quad s_i - \max s_i = 0, \quad i = 1, \dots N-1 \tag{22}$$

Obstacles as well as collisions can be also avoided by using the parametric freedom of $c_0$ and $c_5$. Such a set of paths (equal length and collision free) is shown in figure 2

## IV. Resource Allocation using EGPSO

The approach, which is used to solve the allocation of the tasks to the UAVs in this paper, is a Discrete Particle Swarm Optimisation (DPSO) combined to Evolutionary Game Theory (EGT). One of the main implementation issues of DPSO is the choice of inertial, individual and social coefficients. In order to resolve this problem, those coefficients are optimised by using a dynamical approach based on EGT. The strategies are either to keep going with only inertia, or only with individual, or only with social coefficients. Since the optimal strategy is usually a mixture of the three coefficients, the fitness of the swarm can be maximized when an optimal rate for each coefficient is obtained. This algorithm is described in our previous study [15]. In this method, all the particles $X = (t_i)_{i \in [1 \dots T]}$ are considered as a vector of feasible solution, where the $t_i$ denote the tasks $i \in [1 \dots T]$ and $T$ is the number of tasks to achieve. The index of the vector represents the id of the UAV. Moreover, permuting the elements of $X$ in this representation gives all the possible solutions so it could enable to deal with high dimensional problems. The single constraint of assigning one UAV to one task does not enable us to deal with the common DPSO algorithm described by Eberhart and Kennedy in [18]. Due to the context and the application, it was required to adapt the form of the particles according to the problem.

### A. Swarm organisation

In the PSO algorithm, the establishment of the networks is a key point to maximise the exploration and the global efficiency of the algorithm to solve a problem. In this paper, in order to assigned the tasks to the UAVs, three different swarms are used. Each of them has its own features. One will adopt only inertial behaviour ($c_1 \neq 0$ and $c_2 = c_3 = 0$). One will adopt only selfish behaviour ($c_2 \neq 0$ and $c_1 = c_3 = 0$). One will adopt only social behaviour ($c_3 \neq 0$ and $c_1 = c_2 = 0$). Then the last one is following the common behaviour of the PSO with the coefficients determined by the result of EGT and the four previous swarms. The coefficients are chosen dynamically according to the current state of all the other swarms.

### B. Particle movement

Using the common process of the PSO, the probability of movement toward another solution is introduced. To guarantee that the particles are moving on the feasible solution space and won't need to be repaired, we use an a priori method. In fact, the particles are built in such a way that the solution space is obtained by permuting feasible solutions. One of the key issue when it is required to convert the PSO into DPSO, is the computation of the velocity. Indeed, in discrete space, the velocity does not really makes sense and it was essential to adapt it to discrete case. The main approach used in [18], [19], [20], [21] is the sigmoid function which enables to convert a velocity into a probability. The proposed method is based on that principle, and set the sigmoid function as $s(v_{id}^t) = 1 - \frac{2}{1+e^{v_{id}^t}}$, $(\forall v_{id}^t \in \mathbb{R}_+)$ $(s_{id}^t \in [0,1])$. Once all the probabilities of change for a particle are obtained, the final step is to draw a random number and compare it to each coefficient of probability. All the coefficients greater than the random number are selected as potential candidate for a permutation. Then we draw randomly two particle coefficients and permute them. (In case we obtain a random number greater than only one coefficient, we consider the particle won't move).

### C. Principle of EGT in the determination of the DPSO coefficients

In order to improve the convergence speed of the DPSO, it is proposed to combine it with EGT. If this way has already been investigated by Miranda and Fonseca in [22] to improve the local exploration of the particles, then by Di Chio in [23] and Liu and Wang [24]. The proposed approach is considering the global swarm's welfare instead of the particle's welfare. Thus, like described in IV-A, there are 3 available strategies to play: inertial, individual, or social. Each swarm will play one pure strategy and will provide its welfare to the others. Then, the EGT process, which is based on the replicator dynamic [15], find the equilibrium strategy which enables to the main swarm to improve or keep its mean welfare.

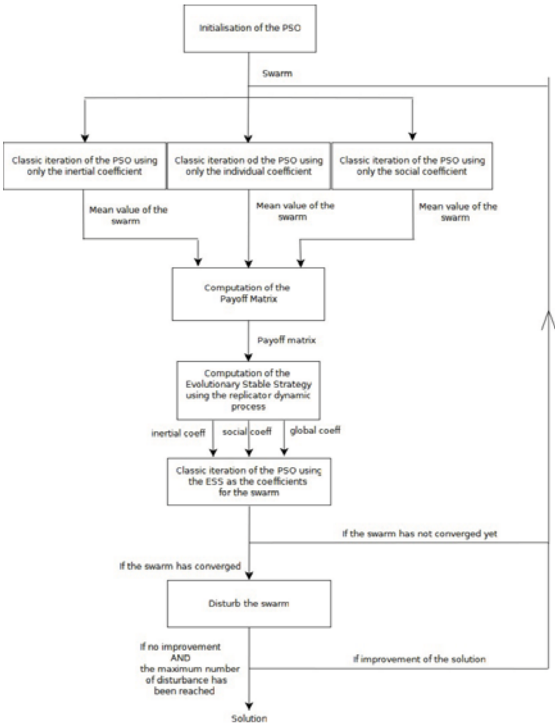

Initialisation of the PSO
Swarm
Classic iteration of the PSO using only the inertial coefficient
Classic iteration od the PSO using only the individual coefficient
Classic iteration of the PSO using only the social coefficient
Mean value of the swarm
Mean value of the swarm
Mean value of the swarm
Computation of the Payoff Matrix
Payoff matrix
Computation of the Evolutionary Stable Strategy using the replicator dynamic process
inertial coeff
social coeff
global coeff
Classic iteration of the PSO using the ESS as the coefficients for the swarm
If the swarm has not converged yet
If the swarm has converged
Disturb the swarm
If no improvement AND the maximum number of disturbance has been reached
If improvement of the solution
Solution

Fig. 3. Overview of the proposed method

### D. Scheme of the process

The figure 3 shows how is designed the algorithm.

## V. Numerical Simulations

In numerical simulations, a swarm of four UAVs is considered for mission deployment. All the UAVs will start from certain starting point at the same time and will reach the goal point at the same time. The goal points of each UAV will be determined from the optimal allocation. During their flights from starting position to finishing position all the UAVs will avoid inter-collisions and collisions with known obstacles. The proposed path planning algorithm will plan safe and flyable flight paths for all the UAVs in the group. For this simulation, the initial poses, the curvature constraints and the safety radii of all the UAVs are summarised in table I. The final poses, which represent targets, are given in table II. All the known static obstacles are given in the terrain database. For simplicity, all the obstacles are assumed to be of rectangular shape in this simulation. The locations of the obstacle are described by the coordinates of their vertices. The vertices of the obstacle1, obstacle2 and obstacle 3 are given table III. The optimal allocation obtained is $[2, 1, 4, 3]$ which represent the first UAV is allocated to the $2nd$ target, the $2nd$ UAV is assigned to the $1st$ target, and so on. The optimal paths of the UAVs are shown in figure 4. As shown in the figure, there is neither inter-collision between UAVs, nor collision with obstacles. The curvature constraint is also satisfied. The length of all the paths are made equal to the longest path by further change of curvature of each path. In this simulation the lengths of each path comes out to be $[36.8, 37.3, 39.7, 36.2]$ meters.

TABLE I
Numerical simulation conditions

| *UAVs* | *Position (m)* | *Heading (deg)* | *$\kappa$ (m)* | *Safety Radius* |
|---|---|---|---|---|
| UAV1 | (8, 6) | 130 | -1/13, +1/3 | 1 |
| UAV2 | (14, 6) | 124 | -1/13, +1/3 | 1 |
| UAV3 | (18, 6) | 3 | -1/13, +1/3 | 1 |
| UAV4 | (25, 6) | 20 | -1/13, +1/3 | 1 |

TABLE II
Final poses

| *No.* | *Position (m)* | *Heading (deg)* |
|---|---|---|
| 1 | (22, 40) | 24 |
| 2 | (17, 40) | 120 |
| 3 | (26, 40) | 113 |
| 4 | (10, 40) | 40 |

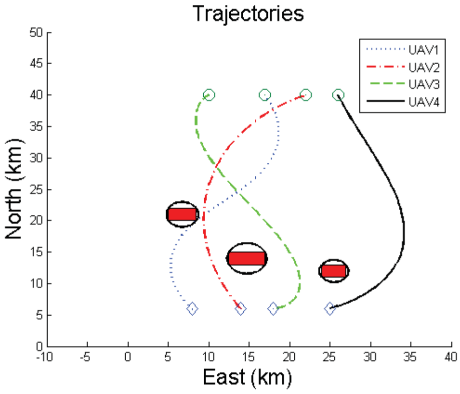


Fig. 4. Optimal PH paths, all equal in length and avoiding obstacles

TABLE III
Location of the known obstacles

| *Obstacles* | *Vertex1* | *Vertex2* | *Vertex3* | *Vertex4* |
|---|---|---|---|---|
| Obstacle1 | (5, 20) | (5,22) | (8.5, 22) | (8.5, 20) |
| Obstacle2 | (12.5, 13) | (12.5, 15) | (17, 15) | (17, 13) |
| Obstacle3 | (24, 11) | (24, 13) | (27, 13) | (27, 11) |

In order to examine the performance of the optimal resource allocation, the optimal result is compared to an arbitrary allocation, $[3, 4, 2, 1]$. The paths for the allocation are shown figure 5. The performance improvement of the optimal allocation compare to the arbitrary allocation is $12.12\%$.

## VI. Conclusions

In this paper, optimal resource allocation with 2D path planning algorithm (2D Path Planner) is proposed. In order to consider operational and physical constraints of the UAVs in the resource allocation design procedure, the proposed algorithm consists of two parts: cooperative path planning based on Pythagorean Hodograph quintic and optimal resource allocation using Evolutionary Game Particle Swarm Optimisation (EGPSO). The algorithm successfully calculated safe

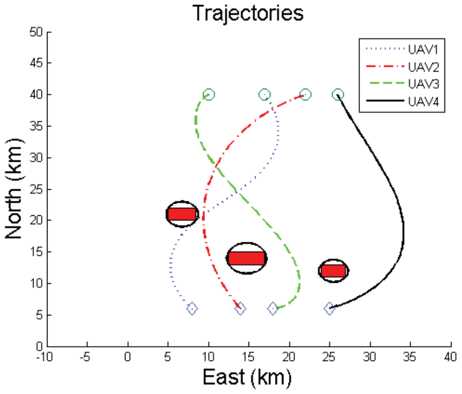


Fig. 5. PH paths for an arbitrary allocation, all equal in length and avoiding obstacles

and flyable paths (feasible paths) for all the UAVs in the group and optimally allocated the UAVs to a group of targets. The performance of the proposed algorithm is evaluated using numerical simulations.

## References

[1] A. Gibbons. *Algorithmic Graph Theory*. Cambridge University Press, 1999.

[2] D. Ahr. *Contributions to Multiple Postmen Problems*. PhD thesis, Heidelberg University, 2004.

[3] J.L. Gross and J. Yellen. *Handbook of Graph Theory*. CRC Press, 2003.

[4] T. Bektas. The multiple traveling salesman problem: An overview of formulations and solution procedures. *The International Journal of Management Science*, 34(3):209–219, 2006.

[5] J. Bellingham, M. Tillerson, A. Richards, and J. How. *Multi-Task Allocation and Path Planning for Cooperating UAVs, Cooperative Control: Models, Applications and Algorithms*. Kluwer Academic Publishers, 2003.

[6] P.R. Chandler, M. Pachter, D. Swaroop, J.M. Hwlett, S. Rasmussen, C. Schumacher, and K. Nygard. Compexity in uav cooperation control. In *American Control Conference*, American Control Conference, Anchorage, AK, 2002.

[7] J. Ryan, T. Bailey, J. Moore, and W. Carlton. Reactive tabu search in unmanned aerial reconnaissance simulations. In *30th Conference on Winter Simulation*, Washington, DC, 1998.

[8] A. Ahmadzadeh, G. Buchman, P. Cheng, A. Jadbabaie, J. Keller, V. Kumar, and G. Pappas. Cooperative control of UAVs for search and coverage. In *Conference on Unmanned Systems*, 2006.

[9] J.O. Royset and H. Sato. Route optimization for multiple searchers. *Naval Research Logistics*, 57:701717, 2010.

[10] N. Perrier, A. Langevin, and J.F. Campbell. A survey of models and algorithms for winter road maintenance. *Computers and Operational Research, Part IV: Vehicle Routing and Fleet Sizing for Plowing and Snow Disposal*, 34:258–294, 2007.

[11] K. Easton and J. Burdick. A coverage algorithm for multi-robot boundary inspection. In *IEEE International Conference on Robotics and Automation*, 2005.

[12] B. Alspach. Searching and sweeping graphs: a brief survey. *Matematiche (Catania)*, 59:5–37, 2006.

[13] T.D. PARSONS. *Pursuit-evasion in a graph*. Theory and Applications of Graphs. Springer-verlagf, Berlin, 1976.

[14] M. Shanmugavel, A. Tsourdos, B. A. White, and R. Zbikowski. Cooperative path planning of multiple uavs using dubins paths with clothoid arcs. *Control Engineering Practice*, 18(9):1084–1092, 2010.

[15] C. Leboucher, R. Chelouah, P. Siarry, and S. Le Ménec. A swarm intelligence method combined to evolutionary game theory applied to ressources allocation problem. In *International Conference on Swarm Intelligence*, 2010.

[16] M. Shanmugavel. *Path planning of multiple autonomous vehicles*. PhD thesis, Cranfield University, 2007.

[17] L.E. Dubins. On curves of minimal length with a constraint on average curvature, and with prescribed initial and terminal positions and tangents. *American Journal of Mathematics*, 79(3):497–516, 1957.

[18] J. Kennedy and R.C. Eberhart. A discrete binary version of the particle swarm algorithm. In *The 1997 IEEE International Conference on Systems, Man, and Cybernetics*, volume 5, pages 4104–4108, Orlando (USA), October 1997.

[19] B. Al-kazemy and C. K. Mohan. Multi-phase discrete particle swarm optimization. In *Fourth International Workshop on Frontiers in Evolutionary*, 2000.

[20] M. Clerc. Discrete particle swarm optimization, illustred by the travelling salesman problem. *New Optimization Techniques in Engineering*, 1:219–239, 2004.

[21] C.J. Liaoa, C.T. Tsengb, and P. Luarnb. A discrete version of particle swarm optimization for flowshop scheduling problems. *Computers & Operations Research*, 34:3099–3111, 2007.

[22] V. Miranda and N. Fonseca. Epso - evolutionary particle swarm optimization, a new algorithm with application in power systems. In *Transmission and Distribution Conference and Exhibition*, volume 2, pages 745–750, October 2002.

[23] C. Di Chio, P. Di Chio, and M. Giacobini. An evolutionary game-theoretical approach to particle swarm optimisation. In *Conference on Applications of evolutionary computing*, pages 575–584, Naples, 2008.

[24] W-B. Liu and X-J. Wang. An evolutionary game based particle swarm optimization algorithm. *Journal of Computational and Applied Mathematics*, 214(1):30–35, 2008.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Design and Control of Novel Tri-rotor UAV

Mohamed Kara Mohamed
School of Electrical and
Electronic Engineering
The University of Manchester
Manchester, UK, M13 9PL
Email: Mohamed.KaraMohamed@postgrad.manchester.ac.uk

Alexander Lanzon
School of Electrical and
Electronic Engineering
The University of Manchester
Manchester, UK, M13 9PL
Email: Alexander.Lanzon@manchester.ac.uk

***Abstract*—Tri-rotor UAVs are more efficient compared to quadrotors in regard to the size and power requirement, yet, tri-rotor UAVs are more challenging in terms of control and stability. In this paper, we propose the design and control of a novel tri-rotor UAV. The proposed platform is designed to achieve six degree of freedom using a thrust vectoring technique with the highest level of flexibility, manoeuvrability and minimum requirement of power. The proposed tri-rotor has a triangular shape of three arms where at the end of each arm, a fixed pitch propeller is driven by a DC motor. A tilting mechanism is employed to tilt the motor-propeller assembly and produce thrust in the desired direction. The three propellers can be tilted independently to achieve full authority of torque and force vectoring. A feedback linearization associated with $\mathscr{H}_\infty$ loop shaping design is used to synthesize a controller for the system. The results are verified via simulation.**

## I. Background and Motivation

In recent decades, Unmanned Aerial Vehicles (UAVs) have attracted growing attention in research due to their wide applications and large potential [1]. Aiming for more efficiency in term of size, autonomy, maneuverability and other factors, various conventional and non-conventional structure designs and configurations of UAV systems are proposed, see [2] and the literature therein. One such design that attracts increasing interest is the vertical-take-off-and-landing (VTOL) tri-rotor configuration.

Tri-rotor vehicles are systems with a three rotors arrangement. This configuration has been proposed as less-expensive with more flexibility and great agility [3], [4]. Compared to quadrotors, tri-rotor UAVs are smaller in size, less complex, less costly and have longer flight time due to the reduction in number of motors [5], which makes tri-rotor vehicles ideal for deployment in various research projects and missions [6].

On another perspective, thrust vectoring has been used in designs to maximize the capability of UAVs [7]. Thrust vectoring is of significant benefit in some applications to arbitrarily orient the vehicle body with respect to the vehicle acceleration vector [8]. In addition, thrust vectoring mechanism is used to give UAVs the capability of taking-off and landing in very narrow areas [9]. In small aircrafts and UAVs, a simple technique of tilt-rotor mechanism can be used to obtain thrust vectoring, where propulsion units are inclined in certain angles using an additional control motor to get the desired thrust in different directions. In general, tilt-rotor mechanism is used in tri-rotor systems to control the horizontal forces and yaw torque of the vehicle. Typically, one rotor only, referred to as the tail rotor, has the ability to tilt to control the yaw moment, see for example [6].

Dynamics of tri-rotor vehicles are highly coupled and nonlinear, which makes the control design of these vehicles the key for successful flight and operations [5]. Compared to quadrotor systems, the yaw control of tri-rotor systems is a further challenge due to the asymmetric configuration of the system. For instance, the reactive yaw moments in quadrotor systems is decoupled from pitch and roll moments which simplifies the yaw control design in such systems. In contrary, pitch, roll and yaw moments are highly coupled in tri-rotor systems. Moreover, attitude control of these vehicles is more challenging compared to quadrotor systems due to gyroscopic and Coriolis terms. In [5], the authors propose a tri-rotor system of which the control design is implemented by four loops for attitude control and guidance. This control design is complicated with coupling between attitude and position control loops and high computation load. The authors in [3] propose a tri rotor configuration in which all rotors of the system tilt simultaneously to the same angle to attain yaw control. The control design considers only the attitude stabilization and neglects the trajectory tracking. The control algorithm in [4] is based on nest saturation for decoupled channels where the configuration of the vehicle makes the separate control of attitude and position possible. The control design of the tri-rotor UAV proposed in [10] discusses only the hovering position. In [11], the attitude of the proposed tri-rotor UAV is controlled by using differential thrust concept between the rotors. The control system design in [12] controls the yaw angle of the proposed tri-rotor UAV by differentially tilting the two main rotors in the plane of symmetry while a fixed up-right propeller is used at the tail to control the pitch moment.

Few researchers have identified the structure of tri-rotor UAV combined with full independent tilt-rotor capability. In this paper, we propose a novel tri-rotor platform, herein referred to as the Tri-rotor UAV, and then we discuss the design and control of the proposed system. The proposed vehicle can achieve full authority of torque and force vectoring by employing three rotors and three servos for tilt-rotor mechanism. This structure gives the vehicle high level of maneuverability and

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

flexibility for translational motion as well as attitude control.

The rest of the paper is organized as follows. In Section II, a functional description of the vehicle and its design is discussed. A mathematical model that captures the dynamics of the UAV and govern the behaviour of the system is derived in Section III. The control system design is presented in Section IV and the simulation results is shown in Section V. The paper ends by conclusion in Section VI.

## II. System Structure and Design

The structure of the proposed Tri-rotor UAV is depicted in Figure 1. The vehicle has a triangular structure of three arms and at the end of each arm, a force generating unit is mounted to produce part of the required controlling force/torque. All three arms are identical of length $l$ and the three force generating units are also identical. Each force generating unit consists of a fixed pitch propeller driven by a Brushless DC (BLDC) motor to generate thrust. The three motors can be powered by a single battery pack or three separate packs located at the center of the body. The propeller-motor assembly is attached to the body arm via a servo motor that can rotate in a vertical plane to tilt the propeller-motor assembly with an angle $\alpha_{s_i}$ in the range $\frac{-\pi}{2} \leq \alpha_{s_i} \leq \frac{\pi}{2}$, $i = 1,2,3$ to produce a horizontal component of the generated force, see Figure 2. All three propellers can be tilted independently to give full authority of thrust vectoring. The system has six degree of freedom in which all movements can be achieved independently and directly by changing the norm of the generated thrust and the tilting angles. This configuration enables the vehicle body to stay aligned in the required direction regardless of the movement the UAV makes.

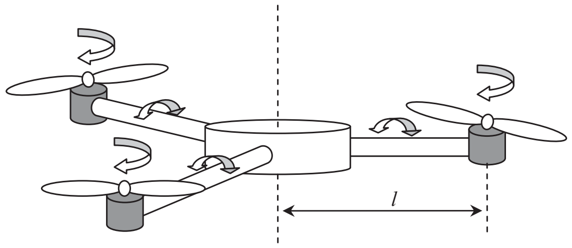


Figure 1. The design of the Tri-rotor UAV (3D view).

## III. Mathematical Modeling

To develop the dynamic model of the UAV, we consider the following right hand coordinate systems shown in Figure 3:
$e$: the generalized earth coordinate system of axes $X_e$, $Y_e$, $Z_e$.
$b$: the body fixed coordinate system in which the origin coincides with the centre of mass of the UAV. The axes of frame $b$ are denoted by $X_b$, $Y_b$, $Z_b$. In addition, we choose three right hand coordinate systems $l_i$ of axes $X_{l_i}$, $Y_{l_i}$, $Z_{l_i}$ with $i = 1,2,3$. These coordinate systems are termed as local coordinate systems and located at the locations of the three propellers, see Figure 4. The origin of each local coordinate system coincides with the joining point between the UAV arm and the propulsion unit where $X_{l_i}$ is extended outside the $i^{th}$ arm of the UAV and $Z_{l_i}$ is along the BLDC motor shaft axis when the tilting angle is zero.

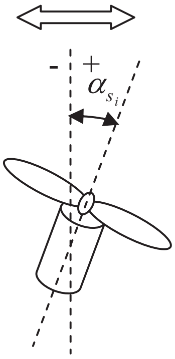


Figure 2. Front view of one arm.

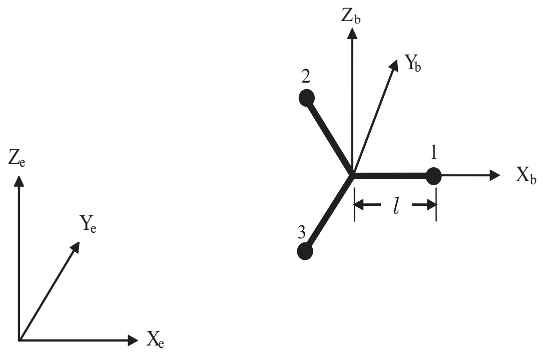


Figure 3. Coordinate systems used to develop the UAV dynamic model.

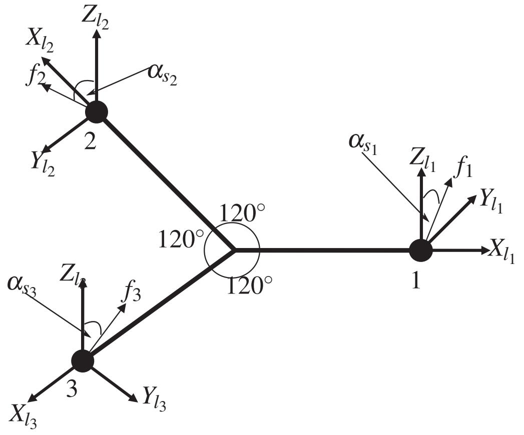


Figure 4. Local coordinate systems at the three propulsion units.

The rotation matrices between the defined coordinate systems are denoted by:
$\boldsymbol{R}_e^b$: the rotational matrix from frame $e$ to frame $b$.
$\boldsymbol{R}_{l_i}^b$: the rotational matrix from coordinate system $l_i$ to coordinate system $b$, $i = 1,2,3$.

In the sequel, we use superscript $b$, $e$ and $l_i$ to denote the coordinate system in which vectors are expressed. The subscript $i$ refers to the $i^{th}$ BLDC motor, servo motor or propeller as applies where $i = 1,2,3$.

In order to obtain the dynamic equations of the UAV, we need to obtain forces and torques acting on the vehicle. We assume very fast actuators and therefore the dynamics of the actuators are neglected.

*Forces*

There are two main forces acting on the UAV which are the propulsive force and the gravitational force.

*The propulsive force:* The total propulsive force $F_{p_\Sigma}$ is equal to the algebraic sum of the three individual propulsive forces generated from propellers. The individual propulsive forces of the three propellers expressed in the local coordinate systems can be written as [13]:

$$F_{p_i}^{l_i} = \begin{bmatrix} 0 \\ k_f \omega_{m_i}^2 \sin(\alpha_{s_i}) \\ k_f \omega_{m_i}^2 \cos(\alpha_{s_i}) \end{bmatrix}, \; i = 1,2,3. \tag{1}$$

where $k_f$ is the thrust to speed constant of the propeller and it is identical for all three propellers, $\omega_{m_i}$ is the rotational speed of the $i^{th}$ BLDC motor (we assume direct driving of the propeller, i.e., the rotational speed of the motor equals the rotational speed of the propeller) and $\alpha_{s_i}$ is the tilting angle of the $i^{th}$ Servo motor.

In the body coordinate system, the individual propulsive forces are given by:

$$F_{p_i}^{b} = \boldsymbol{R}_{l_i}^{b} F_{P_i}^{l_i}, \; i = 1,2,3. \tag{2}$$

From Figure 4, we can obtain the rotation matrices from the local coordinate systems $l_1$, $l_2$ and $l_3$ to the body coordinate system $b$ as:

$$\boldsymbol{R}_{l_1}^{b} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \boldsymbol{R}_{l_2}^{b} = \begin{bmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{2} & 0 \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix}, \boldsymbol{R}_{l_3}^{b} = \begin{bmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} & 0 \\ \frac{-\sqrt{3}}{2} & -\frac{1}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{3}$$

Using Eq.(3), the total propulsive force is:

$$\begin{aligned} F_{p_\Sigma}^{b} &= F_{p_1}^{b} + F_{p_2}^{b} + F_{p_3}^{b} & (4) \\ &= k_f \boldsymbol{H}_f \rho. & (5) \end{aligned}$$

where

$$\boldsymbol{H}_f = \begin{bmatrix} 0 & -\frac{\sqrt{3}}{2} & \frac{\sqrt{3}}{2} & 0 & 0 & 0 \\ 1 & -\frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}, \rho = \begin{bmatrix} \omega_{m_1}^2 \sin(\alpha_{s_1}) \\ \omega_{m_2}^2 \sin(\alpha_{s_2}) \\ \omega_{m_3}^2 \sin(\alpha_{s_3}) \\ \omega_{m_1}^2 \cos(\alpha_{s_1}) \\ \omega_{m_2}^2 \cos(\alpha_{s_2}) \\ \omega_{m_3}^2 \cos(\alpha_{s_3}) \end{bmatrix}. \tag{6}$$

*The gravity force:* The gravitational force in the generalized earth coordinate system is given as:

$$F_g^e = \begin{bmatrix} 0 \\ 0 \\ -gM_{tot} \end{bmatrix}. \tag{7}$$

where $g$ is the gravitational acceleration and $M_{tot}$ is the total mass of the UAV.

In the body coordinate system, we have:

$$F_g^b = \boldsymbol{R}_e^b F_g^e. \tag{8}$$

Using the general notation of rotation angles for the UAV attitude: Roll $\phi_v$, Pitch $\theta_v$ and Yaw $\psi_v$ around the axes $X_e$, $Y_e$ and $Z_e$ respectively, the gravity force in the body system is given by:

$$F_g^b = gM_{tot}H_g \tag{9}$$

where

$$H_g = \begin{bmatrix} \sin(\theta_v) \\ -\sin(\phi_v)\cos(\theta_v) \\ -\cos(\phi_v)\cos(\theta_v) \end{bmatrix}. \tag{10}$$

Now, the total force acting on the UAV and expressed in the body coordinate system is:

$$\begin{aligned} F^b &= F_{p_\Sigma}^b + F_g^b & (11) \\ &= k_f \boldsymbol{H}_f \rho + gM_{tot}H_g. & (12) \end{aligned}$$

*Torques*

The two main torques acting on the UAV are the propulsive torque and the drag torque.

*The propulsive torque:* The propulsive torque is the torque resulting from the generated propulsive force around the center of mass of the vehicle. For the case of the Tri-rotor UAV, we have three identical arms and then the components of the propulsive torque are:

$$\tau_{p_i}^b = \bar{l}_i^b \times F_{p_i}^b, \; i = 1,2,3. \tag{13}$$

where $\bar{l}_i^b$ is the vector of the $i^{th}$ arm between the center of mass of the UAV and the propulsion unit expressed in the body coordinate system. $F_{b_i}^b$ is obtained from Eq. (2).

Now, the total propulsive torque expressed in the body coordinate system is:

$$\begin{aligned} \tau_{p_\Sigma}^b &= \tau_{p_1}^b + \tau_{p_2}^b + \tau_{p_3}^b & (14) \\ &= k_f \boldsymbol{H}_t \rho & (15) \end{aligned}$$

where

$$\boldsymbol{H}_t = l \begin{bmatrix} 0 & 0 & 0 & 0 & \frac{\sqrt{3}}{2} & -\frac{\sqrt{3}}{2} \\ 0 & 0 & 0 & -1 & \frac{1}{2} & \frac{1}{2} \\ 1 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}, \tag{16}$$

$l$ is the length of the vehicle's arm measured between the center of mass of the UAV and the propulsion unit (identical for the three arms) and $\rho$ is defined in Eq. (6).

*The drag torque:* The drag torque is defined as the torque resulting from the aerodynamic drag forces exerted by the ambient fluid (air) on the propeller. Drag torque is in the opposite direction to the direction of rotation. In our case, the resulting drag torque on the $i^{th}$ propeller can be approximated by $\tau_{d_i} = -k_t \omega_{m_i}^2$ [14], where $k_t$ is the drag torque to speed constant resulting from the rotation of the propeller and $\omega_{m_i}$ is the rotational speed of the motor (we consider the BLDC motors drives the propeller directly). In the local coordinate systems $l_i$, the drag torque can be written as:

$$\tau^{l_i}_{d_i} = \begin{bmatrix} 0 \\ -k_t\omega^2_{m_i}\sin(\alpha_{s_i}) \\ -k_t\omega^2_{m_i}\cos(\alpha_{s_i}) \end{bmatrix}, \; i = 1,2,3. \tag{17}$$

In the body coordinate system, the individual drag torques can be represented as:

$$\tau^b_{d_i} = \boldsymbol{R}^b_{l_i}\tau^{l_i}_{d_i}, \tag{18}$$

Using definitions (3), the total drag torque in the body system is given by:

$$\tau^b_{d_\Sigma} = \tau^b_{d_1} + \tau^b_{d_2} + \tau^b_{d_3} \tag{19}$$

$$= -k_t\boldsymbol{H}_f\rho, \tag{20}$$

where $\boldsymbol{H}_f$ and $\rho$ are defined in (6).

Now, the total torque acting on the Tri-rotor and expressed in the body coordinate system is:

$$\tau^b = \tau^b_{p_\Sigma} + \tau^b_{d_\Sigma} \tag{21}$$

$$= \left(k_f\boldsymbol{H}_t - k_t\boldsymbol{H}_f\right)\rho. \tag{22}$$

*Dynamic Model:* Assuming that the Tri-rotor UAV is a rigid body of fixed mass, the vehicle's motion can be described by the Newton-Euler second's law in the body coordinate system as:
for translational motion: $F^b = M_{tot}\left(\dot{\upsilon}^b_v + \boldsymbol{S}(\omega^b_v)\upsilon^b_v\right)$,
for rotational motion: $\tau^b = \boldsymbol{I}^b_v\dot{\omega}^b_v + \boldsymbol{S}(\omega^b_v)\boldsymbol{I}^b_v\omega^b_v$,
where $\upsilon^b_v$ is the translational velocity of the UAV, $\omega^b_v$ is the angular velocity of the UAV, $\boldsymbol{S}(\omega^b)$ is the skew matrix of the vector $\omega^b_v$ and $\boldsymbol{I}^b_v$ is the inertia matrix of the UAV all with respect to the fixed body coordinate system. Assuming no mass change over time, $\boldsymbol{I}^b_v$ is fixed.

Now, Substituting $F^b$ and $\tau^b$ from (12) and (22) gives:

$$k_f\boldsymbol{H}_f\rho + gM_{tot}H_g = M_{tot}\left(\dot{\upsilon}^b_v + \boldsymbol{S}(\omega^b_v)\upsilon^b_v\right) \tag{23}$$

$$(k_f\boldsymbol{H}_t - k_t\boldsymbol{H}_f)\rho = \boldsymbol{I}^b_v\dot{\omega}^b_v + \boldsymbol{S}(\omega^b_v)\boldsymbol{I}^b_v\omega^b_v \tag{24}$$

Let $\eta_v$ and $\lambda^e_v$ be the attitude vector and the position vector of the UAV related to the earth coordinate system and defined as:

$$\eta_v = \begin{bmatrix}\phi_v & \theta_v & \psi_v\end{bmatrix}^T, \; \lambda^e_v = \begin{bmatrix}x_v & y_v & z_v\end{bmatrix}^T.$$

To fully describe the dynamic equations of the UAV, we have the following relations from [15]:

$$\dot{\eta}_v = \boldsymbol{\Psi}\omega^b_v \;,\; \dot{\lambda}^e_v = (\boldsymbol{R}^b_e)^{-1}\upsilon^b_v$$

where $\boldsymbol{\Psi}$ is the rotational matrix between the angular velocity expressed in the body coordinate system $\omega^b_v$ and the angular velocity in the earth coordinate system $\dot{\eta}_v$. $\boldsymbol{\Psi}$ is given in [15] as:

$$\boldsymbol{\Psi} = \begin{bmatrix} 1 & \sin(\phi_v)\tan(\theta_v) & \cos(\phi_v)\tan(\theta_v) \\ 0 & \cos(\phi_v) & -\sin(\phi_v) \\ 0 & \sin(\phi_v)\sec(\theta_v) & \cos(\phi_v)\sec(\theta_v) \end{bmatrix}, \; \frac{-\pi}{2} < \theta_v < \frac{\pi}{2}. \tag{25}$$

From the properties of the rotation matrix we have $(\boldsymbol{R}^b_e)^{-1} = \boldsymbol{R}^e_b$, where $\boldsymbol{R}^e_b$ is the rotation matrix from the body coordinate system $b$ to the earth coordinate system $e$.

Finally, the dynamic model of the UAV can be written as:

$$\dot{\upsilon^b}_v = gH_g - \boldsymbol{S}(\omega^b_v)\upsilon^b_v + \frac{k_f}{M_{tot}}\boldsymbol{H}_f\rho \tag{26}$$

$$\dot{\omega^b}_v = -(\boldsymbol{I}^b_v)^{-1}\boldsymbol{S}(\omega^b_v)\boldsymbol{I}^b_v\omega^b_v + (\boldsymbol{I}^b_v)^{-1}(k_f\boldsymbol{H}_t - k_t\boldsymbol{H}_f)\rho \tag{27}$$

$$\dot{\eta}_v = \boldsymbol{\Psi}\omega^b_v \tag{28}$$

$$\dot{\lambda}^e_v = \boldsymbol{R}^e_b\upsilon^b_v \tag{29}$$

This model of the UAV is written in the compact form in which every state variable is a vector of three components, i.e., $x \in \mathbb{R}^{12}$, where:

$$\upsilon^b_v = \begin{bmatrix} u \\ v \\ w \end{bmatrix}, \; \omega^b_v = \begin{bmatrix} p \\ q \\ r \end{bmatrix}, \; \eta_v = \begin{bmatrix} \phi_v \\ \theta_v \\ \psi_v \end{bmatrix}, \; \lambda^e_v = \begin{bmatrix} x_v \\ y_v \\ z_v \end{bmatrix}.$$

Equations (26) - (29) show a nonlinear model with coupling between the translational and rotational dynamics of the UAV. Moreover, there is coupling between inputs and output channels in which all inputs act on all outputs. The system coupling along with the nonlinearity of the system makes the control design of the proposed Tri-rotor UAV a real challenge compared with other UAV configurations. On the other hand, if we consider the control problem of the UAV to be position tracking with attitude regulating, then the system is square in which we have six actuators (three BLDC motor speeds and three servo angles) and six outputs (3D position and three attitude angles). This highlights the positive aspect of the proposed configuration in terms of controller design compared to other UAV systems that are in general underactuated systems such as quadrotors.

## IV. Control System Design

In this section we consider the control design for the proposed Tri-rotor UAV using input-output feedback linearisation and $\mathscr{H}_\infty$ Loop Shaping Design Procedure (LSDP). The control design of the system can be seen as a tracking problem for the position and attitude of the vehicle via the speed of the BLDC motors and the angles of the servo motors. In this case, the system is fully actuated having six outputs and six inputs. The proposed control algorithm is a centralized $\mathscr{H}_\infty$ controller that stabilizes and tracks simultaneously all outputs, i.e., 3D position and three attitude angles. The motivation behind such a centralized control design is to synthesize a robust controller that can compensate for any unmodeled coupling between channels and attenuate cross-coupling noises and disturbances. Moreover, the implementation of such a design is simple via a single feedback loop structure.

In the sequel and for simplicity of expression, the superscript $b$ and $e$ as well as the subscript $v$ are not written unless it is necessary to avoid ambiguity. We consider the vector $\rho$ as the input vector for the UAV system, i.e., $u = \rho$, and we have the output as $y = \begin{bmatrix}\eta & \lambda\end{bmatrix}^T$

To implement input-output feedback linearization, we have:

$$\dot{y} = y^{(1)} = \begin{bmatrix} \dot{\eta} \\ \dot{\lambda} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Psi}\omega \\ \boldsymbol{R}^e_b\upsilon_v \end{bmatrix} \tag{30}$$

and

$$\ddot{y} = y^{(2)} = \begin{bmatrix} \dot{\boldsymbol{\Psi}}\omega + \boldsymbol{\Psi}\dot{\omega} \\ (\dot{\boldsymbol{R}}_b^e)\upsilon + \boldsymbol{R}_b^e\dot{\upsilon} \end{bmatrix} \tag{31}$$

From the general properties of the rotation matrix, we have $\dot{\boldsymbol{R}}_b^e = \boldsymbol{R}_b^e\boldsymbol{S}(\omega^b)$ [16], and then we write:

$$\begin{aligned} y^{(2)} &= \begin{bmatrix} \dot{\boldsymbol{\Psi}}\omega + \boldsymbol{\Psi}(-\boldsymbol{I}^{-1}\boldsymbol{S}(\omega)\boldsymbol{I}\omega + \boldsymbol{I}^{-1}(k_f H_t - k_t \boldsymbol{H}_f)\rho) \\ \boldsymbol{R}_b^e\boldsymbol{S}(\omega)\upsilon + \boldsymbol{R}_b^e(gH_g - \boldsymbol{S}(\omega)\upsilon + \frac{k_f}{M_{tot}}\boldsymbol{H}_f\rho) \end{bmatrix} \\ &= \begin{bmatrix} \left(\dot{\boldsymbol{\Psi}} - \boldsymbol{\Psi}\boldsymbol{I}^{-1}\boldsymbol{S}(\omega)\boldsymbol{I}\right)\omega \\ g\boldsymbol{R}_b^e H_g \end{bmatrix} + \begin{bmatrix} \boldsymbol{\Psi}\boldsymbol{I}^{-1}\left(k_f\boldsymbol{H}_t - k_t\boldsymbol{H}_f\right) \\ \frac{k_f}{M_{tot}}\boldsymbol{R}_b^e\boldsymbol{H}_f \end{bmatrix}\rho \end{aligned} \tag{32}$$

where

$$\dot{\boldsymbol{\Psi}} = \frac{\partial\boldsymbol{\Psi}}{\partial\phi_v}\dot{\phi}_v + \frac{\partial\boldsymbol{\Psi}}{\partial\theta_v}\dot{\theta}_v \tag{33}$$

and $\dot{\phi}_v$, $\dot{\theta}_v$ are obtained from Eq. (28) as: $\dot{\eta} = \boldsymbol{\Psi}\omega^b$.

We define the decoupling matrix $\boldsymbol{\beta}(x)$ as:

$$\boldsymbol{\beta}(x) = \begin{bmatrix} \boldsymbol{\Psi}\boldsymbol{I}^{-1}\left(k_f H_t - k_t\boldsymbol{H}_f\right) \\ \frac{k_f}{M_{tot}}\boldsymbol{R}_b^e\boldsymbol{H}_f \end{bmatrix} \tag{34}$$

We have $\det[\boldsymbol{\beta}(x)] \neq 0$ and the inverse $\boldsymbol{\beta}^{-1}(x)$ exists always[1] for all $x \in \mathbb{R}^{12}$ where $x$ represents the states of the system. The relative degree of the system in the compact form is $r = r_1 + r_2 = 2 + 2 = 4$ which is equal to the number of states in the compact form of the dynamic equations, and there is no zero dynamics.

To linearize the system, we choose a new control input $\vartheta = \begin{bmatrix} \vartheta_1 \\ \vartheta_2 \end{bmatrix}$, and we write our desired linearized dynamics as: $y^{(2)} = \vartheta$.

From Eq. (32) we can write the feedback linearisation law as:

$$u = \boldsymbol{\beta}^{-1}\left(\vartheta - \begin{bmatrix} \left(\dot{\boldsymbol{\Psi}} - \boldsymbol{\Psi}\boldsymbol{I}^{-1}\boldsymbol{S}(\omega)\boldsymbol{I}\right)\omega \\ g\boldsymbol{R}_b^e H_g \end{bmatrix}\right). \tag{35}$$

The centralized input-output feedback linearization handles the coupling without the need for strict assumption on operating point to decouple the system. The linearized model in the compact form is given as:

$$\dot{\zeta} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}\zeta + \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}\vartheta \tag{36}$$

$$y = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}\zeta \tag{37}$$

where

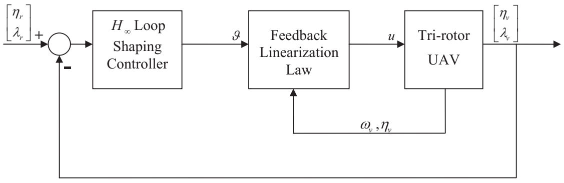
$$\zeta = \begin{bmatrix} \eta \\ \dot{\eta} \\ \lambda \\ \dot{\lambda} \end{bmatrix} \in \mathbb{R}^{12},\ y = \begin{bmatrix} \eta \\ \lambda \end{bmatrix} \in \mathbb{R}^6,\ \vartheta = \begin{bmatrix} \vartheta_1 \\ \vartheta_2 \end{bmatrix} \in \mathbb{R}^6.$$

The linearized plant is a double integrator representing single degree of freedom for transitional and rotational motion.

To control the linearized system, the $\mathscr{H}_\infty$ loop-shaping design is invoked to synthesize a controller for the linearized system. An algorithm proposed in [17] is invoked to simultaneously optimize the synthesis of loop-shaping weights and a stabilizing controller. This algorithm captures the design specification in a systematic manner while trying to maximize the robust stability margin of the closed-loop system. We fix the pre-compensator weight to a low-pass filter on all channels and use the algorithm to optimize an identical post-compensator weights for all channels. The optimized post-compensator for each channel is $w_2 = 105(s+0.6)/(s+8)^2$. The achieved robust stability margin is 0.51 which means a tolerance of approximately 51% of coprime factor uncertainty.

## V. Simulation Results

To demonstrate numerical results, we simulate the Tri-rotor UAV along with the designed controller in Simulink. Figure 5 depicts the block digram for the simulation where $\begin{bmatrix} \eta_r & \lambda_r \end{bmatrix}^T$ is the desired reference attitude and position respectively.



$\begin{bmatrix} \eta_r \\ \lambda_r \end{bmatrix}$
+
-
$H_\infty$ Loop Shaping Controller
$\vartheta$
Feedback Linearization Law
$u$
Tri-rotor UAV
$\begin{bmatrix} \eta_v \\ \lambda_v \end{bmatrix}$
$\omega_v, \eta_v$

Figure 5. Simulation block diagram for the control design of the Tri-rotor UAV.

Figure 6 shows the singular values of the linearized plant, the shaped plant and the synthesized controller.

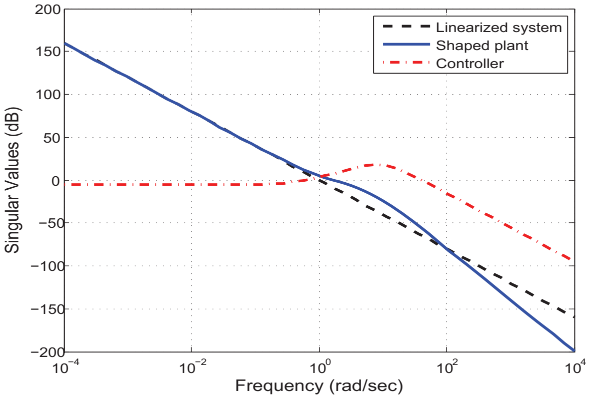


Figure 6. Singular value plots for the linearized system, the shaped system and the controller.

Figure 7 depicts the performance of the UAV for a scenario of horizontal hovering at height of 5 m where the vehicle was at a non-zero initial position and attitude as shown. The speed of the BLDC motors and the angles of the servo motors to stabilize the vehicle and track the references are shown in Figure 8. The controller shows good performance with tracking in all channels. The controller succeeds to maintain the stability of the vehicle and follow the reference trajectory for all initial conditions of the vehicle. The settling time

[1] It is always assumed that $-\pi/2 < \theta_v < \pi/2$.

of the system is about 3 s which is acceptable taking into considering the slow dynamics of the vehicle. The servos and BLDC motors are not saturated and operate within their physical limits of ±90° for the servos and 12000 rpm for the BLDC motors, where these limits come from the technical specifications of the real actuators used in the Tri-rotor UAV.

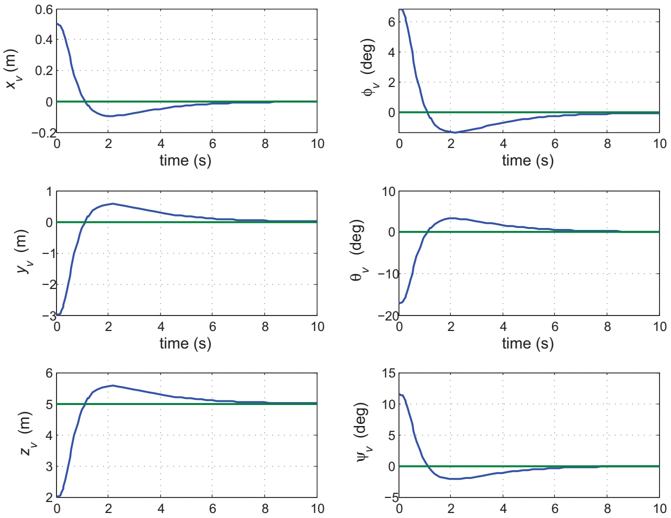


Figure 7. Simulation plots of the UAV position and attitude using the synthesized controller of $\mathscr{H}_\infty$ loop shaping control associated with classical feedback linearization.

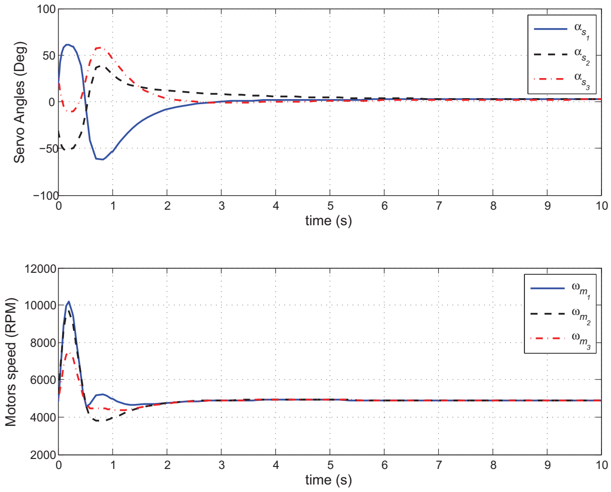


Figure 8. The performance of the actuators (servos and BLDC motors) to track the specified reference input of $(0,0,0)$ deg for attitude, and $(0,0,5)$ m for position coming from non-zero initial point.

## VI. Conclusion

In this paper, a novel tri-rotor UAV is proposed. The proposed UAV has six actuators with full authority of thrust and torque vectoring. The mathematical model of the proposed design is non-linear and it indicates coupling between translational and rotational motion. The nonlinear model of the UAV is linearized by a centralized input-output feedback linearization. This procedure cancels the nonlinearity of all channels simultaneously without further conditions for specific operating point which is the case when we handle channels individually. The linearized plant is a double integrator that is controlled using $\mathscr{H}_\infty$ loop-design procedure. The result is verified via simulations. More complex feedback linearization techniques (such as robust feedback linearization in [18]) can be used in the same manner to avoid linearizing the system to a double integrator.

## References

[1] A. Das, K. Subbarao, and F. lewis, "Dynamic inversion with zero-dynamics stabilisation for quadrotor contrl," *IET Control Theory and Applications*, vol. 3, pp. 303–314, 2009.

[2] K. P. Valavanis, Ed., *Advances in Unmanned Aerial Vehicles: State of the Art and the Road to Autonomy*, ser. International Series on Intelligent Systems, Control, and Automation: Science and Engineering. Springer, 2007, vol. 33.

[3] J. Escareno, A. Sanchez, O. Garcia, and R. Lozano, "Triple tilting rotor mini-UAV: Modeling and embedded control of the attitude," in *American Control Conference*, 2008, pp. 3476–3481.

[4] S. Salazar-Cruz, R. Lozano, and J. Escareño, "Stabilization and nonlinear control for a novel trirotor mini-aircraft," *Control Engineering Practice*, vol. 17, no. 8, pp. 886–894, August 2009.

[5] R. Huang, Y. Liu, and J. J. Zhu, "Guidance, navigation, and control system design for tripropeller vertical-takeoff-and-landing unmanned air vehicle," *Journal of Aircraft*, vol. 46, no. 6, pp. 1837–1856, November-December 2009.

[6] D.-W. Yoo, H.-D. Oh, D.-Y. Won, and M.-J. Tahk, "Dynamic modeling and stabilization technique for tri-rotor unmanned aerial vehicles." *International Journal of Aeronautical and Space Science.*, vol. 11, pp. 167–174, 2010.

[7] P. Vanblyenburgh, "UAVs: An Overview," *Air & Space Europe*, vol. 1, no. 5-6, pp. 43–47, September 1999.

[8] B. Crowther, A. Lanzon, M. Maya-Gonzalez, and D. Langkamp, "Kinematic analysis and control design for a non planar multirotor vehicle," *Journal of Guidance, Control, and Dynamics*, vol. 34, no. 4, pp. 1157–1171, 2011.

[9] F. Lin, W. Zhang, and R. D. Brandt, "Robust hovering control of a PVTOL aircraft," *Control Systems Technology, IEEE Transactions on*, vol. 7, no. 3, pp. 343–351, August 2002.

[10] P. Rongier, E. Lavarec, and F. Pierrot, "Kinematic and dynamic modeling and control of a 3-rotor aircraft," in *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*, 2005, pp. 2606–2611. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1570506

[11] P. Fan, X. Wang, and K.-Y. Cai, "Design and control of a tri-rotor aircraft," in *Control and Automation (ICCA), 2010 8th IEEE International Conference on*, June 2010, pp. 1972 –1977.

[12] Z. Prime, J. Sherwood, M. Smith, and A. Stabile, "Remote control (rc) vertical take-off and landing (vtol) model aircraft," LevelIV Honours Project Final Report, University of Adelaide, Adelaide, Australia, October 2005.

[13] W. F. Phillips, *Mechanics of Flight*. Wiley & Sons, Inc., 2004.

[14] W. Z. Stepniewski and C. N. Keys, *Rotory-Wing Aerodynamics*. Dover Publications, Inc., 1984.

[15] G. D. Padfield, *Helicopter Flight Dynamics: The Theory and Application of Flying Qualities and Simulation Modeling*, ser. Education Series. AIAA, 1996.

[16] A. Isidori, L. Marconi, and A. Serrani, *Robust Autonomous Guidance: An Internal Model Approach*, 1st ed., ser. Advanced in Industrial Control. Springer, 2003.

[17] A. Lanzon, "Weight optimisation in $H_\infty$ loop-shaping," *Automatica*, vol. 41, no. 7, pp. 1201–1208, 2005.

[18] A. L. D. Franco, H. Bourlès, E. R. De-Pieri, and H. Guillard, "Robust nonlinear control associating robust feedback linearization and $H_\infty$ control," *IEEE Transactions on Automatic Control*, vol. 51, no. 7, pp. 1200–1206, 2006.

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# Singular Perturbation Control of the Longitudinal Flight Dynamics of an UAV

Sergio Esteban
Department of Aerospace Engineering
University of Seville
Seville, Spain 41092
Email: sesteban@.us.es

Damián Rivas
Department of Aerospace Engineering
University of Seville
Seville, Spain 41092
Email: drivas@.us.es

***Abstract*—This paper presents a singular perturbation control strategy for regulating the longitudinal flight dynamics of an Unmanned Air Vehicle (UAV). The proposed control strategy is based on a four-time-scale (4TS) decomposition that includes the altitude, velocity, pitch, and flight path angle dynamics, with the control signals being the elevator deflection and the throttle position. The nonlinear control strategy drives the system to follow references in the aerodynamic velocity and the flight path angle. In addition, the control strategy permits to select the desired dynamics for all the singularly perturbed subsystems. Numerical results are included for a realistic nonlinear UAV model, including saturation of the control signals.**

## I. Introduction

Historically, classical linear control techniques have been sufficient to obtain reasonable control responses of aerospace systems, but the evolution of the aerospace industry, and the consequent improvement of technologies, have increased the performance requirements of all systems in general, which has called for better control designs that can deal with more complex systems. Specifically, in the area of aerospace systems, a wide range of different nonlinear control techniques have been studied to deal with the nonlinear dynamics of such systems. From singular perturbation [1], [2], feedback linearization [3], dynamic inversion [4], sliding mode control [5], or backstepping control methods [6], [7], to name a few. Neural Networks (NN) are also included within the realm of nonlinear control techniques, and seem to provide improved robustness properties under system uncertainties. Some of works include Adaptive Critic Neural Network (ACNN) based controls, originally presented by Balakrishnan and Biega [8], and later extended to many other aerospace systems [9].

One of the most challenging tasks in control is the modeling of systems in which the presence of parasitic parameters, such as small time constants, is often the source of a increased order and stiffness [10]. The stiffness, attributed to the simultaneous occurrence of slow and fast phenomena, gives rise to time-scales, and the suppression of the small parasitic variables results in degenerated, reduced-order systems called singularly perturbed systems (SPS), that can be stabilized separately, thus simplifying the burden of control design of high-order systems.

The application of singular perturbation and time-scale techniques in the aerospace industry can be traced back to the 1960s when it was first applied to solve complex flight optimization problems [11]. Since then, singular perturbation and time-scale techniques have been extensively used in the aerospace industry as described in the extense literature review conducted by Naidu and Calise [10]. In recent years these techniques have been also extended to UAVs [2], [12].

The objective of this paper is to develop a singular perturbation control strategy for the longitudinal dynamics of an aircraft, that be able to follow references in aerodynamic velocity and flight path angle, using as control actuators the elevator deflection and the throttle position. In addition, the proposed singular perturbation control strategy permits to select the desired closed-loop dynamics of each of the resulting reduced-order and boundary-layer subsystems using a time-scale analysis similar to those presented in [2], [13]. Simulations are included for a realistic UAV model including nonlinear dynamics and actuator saturation on both the elevator deflection and throttle setting. The model used corresponds to the Cefiro aircraft [14], an UAV recently designed and constructed by the authors at the University of Seville.

This paper is structured as follows: Section II presents the flight dynamics used throughout this work; Section III presents the time scales selection; Section IV describes the proposed 4-time-scale analysis; the singular perturbed control strategies are presented in Section V; numerical results for the UAV model considered in this paper are given in Section VI; and finally, some conclusions are drawn in Section VII.

## II. Model Definition

The problem discussed in this article considers a constant-mass UAV with an electrical propulsion plant, for which the point-mass longitudinal flight dynamics equations are

$$\dot{h} = V\sin\gamma, \tag{1}$$

$$\dot{V} = \frac{1}{m}\left(T - D - mg\sin\gamma\right), \tag{2}$$

$$\dot{\theta} = q, \tag{3}$$

$$\dot{\gamma} = \frac{1}{mV}\left(L - mg\cos\gamma\right). \tag{4}$$

$$\dot{q} = \frac{M}{I_y}, \tag{5}$$

where $h$ is the altitude; $V$ the aerodynamic speed; $\gamma$ the flight path angle; $\theta$ the pitch angle; $q$ the pitch rate; $T$, $D$, and $L$ the thrust, drag, and lift forces, respectively; $M$ the total pitch moment; $m$ and $I_y$ the mass and the moment of inertia of the UAV. The thrust-force model used is given by

$$T = \delta_T(T_0 + T_1 V + T_2 V^2), \qquad (6)$$

where $\delta_T$ is the throttle setting, $0 \leq \delta_T \leq 1$ and $T_0, T_1, T_2$ are known coefficients obtained through wind tunnel experiments. The lift, drag and pitch moment are given by the following expressions

$$L = q_\infty S C_L,\ D = q_\infty S C_D,\ M = q_\infty S c C_M, \qquad (7)$$

where $q_\infty = 1/2\rho V^2$ is the dynamic pressure; $S$ is the reference wing area, $c$ is the wing mean aerodynamic chord, and $C_L$, $C_D$ and $C_M$ are the lift, drag and pitch moment coefficients, which are given by the following standard models [15], [16] that have been widely used in the literature [11], [17], [18]

$$\begin{aligned} C_L &= C_{L_0} + C_{L_\alpha}\alpha + C_{L_\delta}\delta, & (8)\\ C_D &= C_{D_0} + kC_L^2, & (9)\\ C_M &= C_{M_0} + C_{M_\alpha}\alpha + C_{M_\delta}\delta + C_{M_q}q, & (10) \end{aligned}$$

where $\alpha$ is the angle of attack, given by $\alpha = \theta - \gamma$, $\delta$ is the elevator deflection, $-40^\circ \leq \delta \leq 40^\circ$, and $C_{L_0}$, $C_{L_\alpha}$, $C_{L_\delta}$, $C_{D_0}$, $k, C_{M_0}$, $C_{M_\alpha}$, $C_{M_\delta}$, and $C_{M_q}$ are known aerodynamic coefficients. In this paper the simplifying assumption of constant air density is considered, and, therefore, the altitude equation becomes decoupled from the rest, and can be solved a posteriori. Equations (1–5) are expanded using Eqns. (8–10), resulting in

$$\begin{aligned} \dot{h} &= V\sin\gamma, & (11)\\ \dot{V} &= \delta_T(a_1 + a_2V^2 + a_3V) + V^2\left[a_4 + a_5 + a_6(\theta-\gamma)\right.\\ &+ \left. a_7(\theta-\gamma)^2 + a_8\delta + a_9\delta^2 + a_{10}(\theta-\gamma)\delta\right]\\ &+ a_{11}\sin\gamma, & (12)\\ \dot{\theta} &= q, & (13)\\ \dot{\gamma} &= V\left[a_{12} + a_{13}(\theta-\gamma) + a_{14}\delta\right] + \frac{a_{11}}{V}\cos\gamma, & (14)\\ \dot{q} &= V^2\left[a_{15} + a_{16}(\theta-\gamma) + a_{17}\delta + a_{18}q\right], & (15) \end{aligned}$$

where $a_1 = T_0/m$, $a_2 = T_2/m$, $a_3 = T_1/m$, $a_4 = -c_2C_{D_0}$, $a_5 = -c_1C_{L_0}^2$, $a_6 = -2c_1C_{L_0}C_{L_\alpha}$, $a_7 = -c_1C_{L_\alpha}^2$, $a_8 = -2c_1C_{L_0}C_{L_\delta}$, $a_9 = -c_1C_{L_\delta}^2$, $a_{10} = -2c_1C_{L_\alpha}C_{L_\delta}$, $a_{11} = -g$, $a_{12} = c_2C_{L_0}$, $a_{13} = c_2C_{L_\alpha}$, $a_{14} = c_2C_{L_\delta}$, $a_{15} = c_3C_{M_0}$, $a_{16} = c_3C_{M_\alpha}$, $a_{17} = c_3C_{M_\delta}$, $a_{18} = c_3C_{M_q}$ with $c_1 = \rho Sk/(2m)$, $c_2 = \rho S/(2m)$, and $c_3 = \rho Sc/(2I_y)$.

The underactuated structure of the system requires that two variables need to be used as references. In this work, the nonlinear control strategy will seek to drive the system to follow references in the aerodynamic velocity and the flight path angle, that is $V = V_{ref}$ and $\gamma = \gamma_{ref}$.

## III. TIME SCALES SELECTION

The appropriate selection of time scales is an important aspect of the singular perturbation and time-scales theory [10], [19]–[21], and can be categorized into three approaches: 1) direct identification of small parameters (such as small time constants); 2) transformation of state equations; and 3) linearization of the state equations. Ardema [19] proposes a rational method of identifying time scales separations that does not rely on an *ad hoc* selection of time scales based largely on physical insight and past experiences with similar problems. The proposed method only requires a knowledge of the state equations. Considering a dynamical systems of the form

$$\dot{x} = f(x,u),\ u \in U, \qquad (16)$$

subject to suitable boundary conditions, where $x$ is an $n$-dimensional state vector, $u$ an $r$-dimensional control vector, and $U$ the set of admissible controls. It is assumed that bounds have been established on the components of the state vector, either by physical limitations or by a desire to restrict the state to a certain region of state space, $x_{i,m} \leq x_i \leq x_{i,M}$, with $x_{i,m}$ and $x_{i,M}$ representing the minimum and maximum values of the state variables. As noted in [19], most *ad hoc* assessments of time-scale separation are based on the concept of state variable speed [11], [22]. The speed of a state variable $x_i$ is defined as the inverse of the time it takes that variable to change across a specified range of values, which can be expressed as

$$S_i = \frac{\dot{x}_i}{\Delta x_i} = \frac{f_i(x,u)}{\Delta x_i}. \qquad (17)$$

where $\Delta x_i = x_{i,M} - x_{i,m}$. Two methods are proposed [19] to determine if two variables are candidates for time-scale separation, which is ultimately defined if the two variables have widely separated speeds. In this work, the method that considers a reference value of the state in the region of interest, $\bar{x}$, is adopted, hence

$$S_i = \frac{1}{\Delta x_i}\max_{u\in U} f_i(\bar{x},u). \qquad (18)$$

Since the maneuvers being considered for the UAV are those of climb performance, the reference states can be selected as those associated to the condition of maximum rate of climb $V_{v_{max}}$, condition that has been investigated in [14]. The bounds for the UAV model considered in this article [14] are therefore: $0 \leq h \leq 1000$, where the maximum altitude is defined by desired operation limits; $V_m \leq V \leq V_{max}$, with $V_m = 1.2 * V_{stall}$, and $V_{max}$ obtained in the performance analysis [14]; $-\gamma_{d_{max}} \leq \gamma \leq \gamma_{V_m}$, where $\gamma_{d_{max}}$ represents the maximum descent glide angle, and $\gamma_{V_m}$ represents the maximum flight path angle at $V_m$; $\alpha_{trim_{V_m}} \leq \alpha \leq \alpha_{trim_{V_{max}}}$, where $\alpha_{trim_{V_m}}$ and $\alpha_{trim_{V_{max}}}$ corresponds to the trim angle for $V_m$, and $V_{max}$, respectively; $\theta_m \leq \theta \leq \theta_M$ with $\theta_m = \gamma_m + \alpha_m$ and $\theta_M = \gamma_M + \alpha_M$; and finally, the bounds for the pitch rate $q_m \leq q \leq q_M$ are selected by desired operation limits.

From [14], the UAV being studied in this article has the following geometric properties, $S = 1.088\ m^2$, $c = 0.393\ m$, $m = 23.186\ kg$, $I_y = 7.447\ kgm^2$, $C_{D_0} = 0.0286$, $k = 0.0426$. From the stability analysis conducted in [14], the derivatives are: $C_{L_0} = 0.408$, $C_{M_0} = 0.0617$, $C_{L_\alpha} = 3.823$ per rad, $C_{M_\alpha} = -0.455$ per rad, $C_{L_\delta} = 0.284$ per rad, $C_{M_\delta} = -0.914$ per rad, $C_{M_q} = -13.590$ s/rad, $T_0 = 127.53N$, $T_1 = -2.9052 \times 10^{-1}Ns/m$ and $T_2 = -5.9616 \times 10^{-2}Ns^2/m^2$.

From the performance study in [14], it can be obtained that the maximum vertical climb speed is $V_{v_{max}} = 27.00$ m/s, the reference horizontal speed is $\bar{V} = 22.90\ m/s$, the stall velocity is given by $V_{stall} = 14.38\ m/s$, for $C_{L_{max}} = 1.65$, the maximum velocity $V_{max} = 38.47\ m/s$, the flight path angle for $\bar{V}$ is given $\gamma_{\bar{V}} = 19.06^\circ$, the gliding angle for minimum flight path angle is given by $\gamma_{V_d} = -7.63^\circ$, the trim angles for $\bar{V}$ are given by $\alpha_{trim_{\bar{V}}} = 3.47°$ and $\delta_{trim_{\bar{V}}} = 2.14°$ respectively. The final bounds and *speeds* of the state variables are resumed in Table I. It can be seen four clearly differentiated time-scales, since the *speeds* for altitude and velocity dynamics are not that separated, therefore, it will be assumed that they move in the same stretched time scale, and will be denoted by the augmented state vector $\chi = [h\ V]$.

TABLE I
BOUNDS AND *speeds* OF THE STATE VARIABLES.

| Variable | $x_{i,m}$ | $x_{i,M}$ | $\Delta x_i$ | $\bar{x}_i$ | $S_i$ |
|---|---|---|---|---|---|
| $h$ [m] | 0 | 1000 | 1000 | 200 | 0.0074 |
| $V$ [m/s] | 17.25 | 38.47 | 21.21 | 22.90 | 0.011 |
| $\theta$ [deg] | −7.58 | 32.84 | 40.42 | 22.53 | 0.41 |
| $\gamma$ [deg] | −4.53 | 23.31 | 27.83 | 19.06 | 2.21 |
| $\alpha$ [deg] | −3.06 | 9.53 | 12.59 | 3.47 | $N/A$ |
| $q$ [deg/s] | −264.44 | 264.44 | 528.88 | 0 | 82.43 |

With this in mind, Eqns. (11–15) are rewritten as a four-time-scale (4TS) singular perturbed model of the form

$$
\begin{aligned}
\dot{\chi} &= f_\chi(\chi, \theta, \gamma, \delta, \delta_T),\ \chi \in B_\chi, && (19)\\
\varepsilon_1 \dot{\theta} &= f_\theta(q),\ \theta \in B_\theta, && (20)\\
\varepsilon_1 \varepsilon_2 \dot{\gamma} &= f_\gamma(\chi, \theta, \gamma, \delta),\ \gamma \in B_\gamma, && (21)\\
\varepsilon_1 \varepsilon_2 \varepsilon_3 \dot{q} &= f_q(\chi, \theta, \gamma, q, \delta),\ q \in B_q, && (22)
\end{aligned}
$$

with $B_\chi, B_\theta, B_\gamma, B_q$ denoting closed sets of the variables $\chi$, $\theta$, $\gamma$ and $q$, respectively, being $\chi$ the slowest variable, $\theta$ the intermediate variable, $\gamma$ the fast variable, and $q$ the ultra-fast variable, and holding that $0 < \varepsilon_1\varepsilon_2\varepsilon_3 \ll \varepsilon_1\varepsilon_2 \ll \varepsilon_1 \ll 1$. In order to express the original set of differential Eqns. (11–15) in the standard 4TS singular perturbation formulation, a series of algebraic modifications using the *speeds* of the different variables are conducted. Let consider the different *speeds* as if they were the inverse of the inertias multiplying the time derivatives such $I_h = 1/S_h = 133.722$, $I_V = 1/S_V = 89.221$, $I_\theta = 1/S_\theta = 2.411$, $I_\gamma = 1/S_\gamma = 0.451$, $I_q = 1/S_q = 0.012$, where it can be easily identified that $I_V >> I_\theta >> I_\gamma >> I_q$, therefore, in order to express the equations of the 4TS in the correct multi-time singular perturbation formulation, all the perturbation parameters are normalized with respect to the slowest coefficient, that is $I_h$, yielding the parasitic constants selected for this problem given by $\varepsilon_1 = I_\theta/I_h = 1.803\times10^{-2}$, $\varepsilon_1\varepsilon_2 = I_\gamma/I_h = 3.375\times10^{-3}$, and $\varepsilon_1\varepsilon_2\varepsilon_3 = I_q/I_h = 9.0722 \times 10^{-5}$, resulting in

$$
\begin{aligned}
\dot{h} &= V \sin\gamma, && (23)\\
\dot{V} &= \delta_T \left(a_1 + a_2 V^2 + a_3 V\right) + V^2 \left[a_4 + a_5 + a_6 (\theta - \gamma)\right.\\
&+ \quad a_7 (\theta - \gamma)^2 + a_8 \delta + a_9 \delta^2\\
&+ \quad \left.a_{10} (\theta - \gamma) \delta\right] + a_{11} \sin\gamma, && (24)\\
\varepsilon_1 \dot{\theta} &= \varepsilon_1 q, && (25)\\
\varepsilon_1 \varepsilon_2 \dot{\gamma} &= V \left[\bar{a}_{12} + \bar{a}_{13} (\theta - \gamma) + \bar{a}_{14} \delta\right] + \frac{\bar{a}_{11}}{V} \cos\gamma, && (26)\\
\varepsilon_1 \varepsilon_2 \varepsilon_3 \dot{q} &= V^2 \left[\bar{a}_{15} + \bar{a}_{16} (\theta - \gamma) + \bar{a}_{17} \delta + \bar{a}_{18} q\right], && (27)
\end{aligned}
$$

with $\bar{a}_{11} = \varepsilon_1\varepsilon_2 a_{11}$, $\bar{a}_{12} = \varepsilon_1\varepsilon_2 a_{12}$, $\bar{a}_{13} = \varepsilon_1\varepsilon_2 a_{13}$, $\bar{a}_{14} = \varepsilon_1\varepsilon_2 a_{14}$, $\bar{a}_{15} = \varepsilon_1\varepsilon_2\varepsilon_3 a_{15}$, $\bar{a}_{16} = \varepsilon_1\varepsilon_2\varepsilon_3 a_{16}$, $\bar{a}_{17} = \varepsilon_1\varepsilon_2\varepsilon_3 a_{17}$, and $\bar{a}_{18} = \varepsilon_1\varepsilon_2\varepsilon_3 a_{18}$. In addition, the following approximations are considered in this article (which have been widely used in the literature for aircraft trajectory optimization using singular perturbation techniques [11])

$$
\sin\gamma \cong \gamma - \frac{\gamma^3}{6},\ \cos\gamma \cong 1 - \frac{\gamma^2}{2}, \qquad (28)
$$

The following section describes the four-time-scale analysis that will permit to derive the singular perturbation control strategy.

## IV. 4-Time-Scale Analysis

This section presents a sequential time-scale methodology that provides an approach in which, for a specific class of singularly perturbed nonlinear systems, a step-by-step procedure can be employed to design the proper control laws that guarantee a desired degree of stability of each of the time-scale subsystems. The approach is based on the sequential time-scale analysis similar to the one presented in [2], [13], which is an extension of the two-time-scale analysis presented in [1]. The approach consists in decomposing the original singularly perturbed system, Eqns. (23–27), denoted as $\Sigma_{SIFU}$ for simplicity, into a sequential set of two-time-scale (2TS) SPS. Each one of the letters in $\Sigma_{SIFU}$ denotes a time scale, Slow, Intermediate, Fast, and Ultrafast, and will be used as a reference to describe each time-scale subsystem or combination. The time-scale decomposition is achieved by applying, in a sequential manner, the associated stretched time scales for each of the subsystems, resulting in reduced order models. The time-scale decomposition is started by applying first the stretched time scale given by $\tau_3 = t/(\varepsilon_1\varepsilon_2\varepsilon_3)$, resulting in a 2TS SPS formed by the reduced order $\Sigma_{SIF}$-subsystem

$$
\begin{aligned}
\dot{h} &= V\left(\gamma - \frac{\gamma^3}{6}\right), && (29)\\
\dot{V} &= \delta_T \left(a_1 + a_2 V^2 + a_3 V\right) + V^2 \left[a_4 + a_5 + a_6 (\theta - \gamma)\right.\\
&+ \quad a_7 (\theta - \gamma)^2 + a_8 \delta + a_9 \delta^2\\
&+ \quad \left.a_{10} (\theta - \gamma) \delta\right] + a_{11}\left(\gamma - \frac{\gamma^3}{6}\right), && (30)\\
\varepsilon_1 \dot{\theta} &= \varepsilon_1 H_q(\theta, \gamma, \delta), && (31)\\
\varepsilon_1 \varepsilon_2 \dot{\gamma} &= V \left[\bar{a}_{12} + \bar{a}_{13} (\theta - \gamma) + \bar{a}_{14} \delta\right] + \frac{\bar{a}_{11}}{V}\left(1 - \frac{\gamma^2}{2}\right), && (32)
\end{aligned}
$$

and the boundary layer (fast), denoted as $\Sigma_U$-subsystem for simplicity, given by

$$
\frac{dq}{d\tau_3} = V^2 \left[\bar{a}_{15} + \bar{a}_{16} (\theta - \gamma) + \bar{a}_{17} \delta + \bar{a}_{18} q\right], \qquad (33)
$$

where $H_q(\theta, \gamma, \delta)$ represents the quasi-steady-state equilibrium (QSSE) of the boundary layer $\Sigma_U$-subsystem when setting $\varepsilon_3 = 0$, that is $0 = f_q(\chi, \theta, \gamma, q, \delta) \rightarrow \bar{q} = H_q(\theta, \gamma, \delta)$, resulting in

$$
\bar{q} = H_q(\theta, \gamma, \delta) = -\frac{\bar{a}_{15} + \bar{a}_{16}(\theta - \gamma) + \bar{a}_{17}\delta}{\bar{a}_{18}}. \qquad (34)
$$

Recall that in the space of configuration of the boundary layer $\Sigma_U$-subsystem, the variables, $\chi$, $\theta$, and $\gamma$, are treated like fixed parameters. The time scale analysis continues recognizing that the reduced order $\Sigma_{SIF}$-subsystem, Eqns. (29–32), can be decomposed again into a 2TS SPS by applying the stretched time scale $\tau_2 = t/(\varepsilon_1\varepsilon_2)$, resulting in a new reduced order (slow) subsystem, denoted as $\Sigma_{SI}$-subsystem for simplicity, defined as

$$\begin{aligned}
\dot{h} &= V\left(H_\gamma - \frac{H_\gamma^3}{6}\right), && (35)\\
\dot{V} &= \delta_T\left(a_1 + a_2V^2 + a_3V\right) + V^2\left[a_4 + a_5 + a_6\left(\theta - H_\gamma\right)\right.\\
&+ a_7\left(\theta - H_\gamma\right)^2 + a_8\delta + a_9\delta^2\\
&+ \left. a_{10}\left(\theta - H_\gamma\right)\delta\right] + a_{11}\left(H_\gamma - \frac{H_\gamma^3}{6}\right), && (36)\\
\varepsilon_1\dot{\theta} &= -\frac{\varepsilon_1\left(\bar{a}_{15} + \bar{a}_{16}(\theta - H_\gamma) + \bar{a}_{17}\delta\right)}{\bar{a}_{18}}, && (37)
\end{aligned}$$

and with a new boundary layer (fast) subsystem, denoted as $\Sigma_F$-subsystem for simplicity, given by

$$\frac{d\gamma}{d\tau_2} = V\left[\bar{a}_{12} + \bar{a}_{13}\left(\theta - \gamma\right) + \bar{a}_{14}\delta\right] + \frac{\bar{a}_{11}}{V}\left(1 - \frac{\gamma^2}{2}\right), \quad (38)$$

where $H_\gamma(\chi, \theta, \delta)$ represents the QSSE of the boundary layer $\Sigma_F$-subsystem when setting $\varepsilon_2 = 0$, that is $0 = f_\gamma(\chi, \theta, \gamma, \delta) \to \bar{\gamma} = H_\gamma(\chi, \theta, \delta)$, resulting in

$$\bar{\gamma} = H_\gamma(\chi, \theta, \delta) = A_1 \pm A_2\sqrt{A_3 + A_4\theta + A_5\delta}, \quad (39)$$

with $A_1 = -a_{13}V^2/a_{11}$, $A_2 = 1/a_{11}$, $A_3 = a_{13}^2V^4 + 2a_{11}a_{12}V^2 + 2a_{11}^2$, $A_4 = 2a_{11}a_{13}V^2$ and $A_5 = 2a_{11}a_{14}V^2$, where it can be shown that the positive solution is the valid one, and where $\chi$, and $\theta$ are treated like fixed parameters. Finally, it can be recognized that the $\Sigma_{SI}$-subsystem can be decomposed one more time into another 2TS SPS by considering the last stretched time scale $\tau_1 = t/\varepsilon_1$, resulting in a new reduced order (slow) subsystem, denoted as $\Sigma_S$-subsystem for simplicity, and given by

$$\begin{aligned}
\dot{h} &= V\left(\bar{H}_\gamma - \frac{\bar{H}_\gamma^3}{6}\right), && (40)\\
\dot{V} &= \delta_T\left(a_1 + a_2V^2 + a_3V\right) + V^2\left[a_4 + a_5 + a_6\left(H_\theta - \bar{H}_\gamma\right)\right.\\
&+ a_7\left(H_\theta - \bar{H}_\gamma\right)^2 + a_8\delta + a_9\delta^2\\
&+ \left. a_{10}\left(H_\theta - \bar{H}_\gamma\right)\delta\right] + a_{11}\left(\bar{H}_\gamma - \frac{\bar{H}_\gamma^3}{6}\right), && (41)
\end{aligned}$$

with the boundary layer $\Sigma_I$-subsystem given by

$$\frac{\mathrm{d}\theta}{\mathrm{d}\tau_1} = -\frac{\varepsilon_1\left(\bar{a}_{15} + \bar{a}_{16}(\theta - H_\gamma) + \bar{a}_{17}\delta\right)}{\bar{a}_{18}}, \quad (42)$$

where $H_\theta(\chi, \delta)$ represents the QSSE of the boundary layer $\Sigma_I$-subsystem when setting $\varepsilon_1 = 0$, that is $0 = f_\theta(\chi, \theta, \delta) \to \bar{\theta} = H_\theta(\chi, \delta)$, resulting in

$$H_\theta(\chi, \delta) = A_6 \pm A_7\sqrt{A_8 + A_9\delta} + A_{10}\delta, \quad (43)$$

with $A_6 = A_2^2A_4/2 - \bar{a}_{15}/\bar{a}_{16} + A_1$, $A_7 = A_2/(2\bar{a}_{16})$, $A_8 = \bar{a}_{16}^2A_2^2A_4^2 - 4\bar{a}_{15}\bar{a}_{16}A_4 + 4\bar{a}_{16}^2A_3 + 4\bar{a}_{16}^2A_1A_4$, $A_9 = 4\bar{a}_{16}^2A_5 - 4\bar{a}_{16}\bar{a}_{17}A_4$, and $A_{10} = \bar{a}_{17}/\bar{a}_{16}$. Recall also that $\bar{H}_\gamma(\chi, H_\theta, \delta)$ results from substituting the QSSE $H_\theta$ into Eq.(39), and given by

$$\bar{H}_\gamma(\chi, H_\theta, \delta) = A_1 \pm A_2\sqrt{A_3 + A_4H_\theta + A_5\delta}, \quad (44)$$

where $\chi$ is treated as a fixed parameter. The control strategy that will be presented in the following section uses this time-scale separation strategy to obtain a sequential control strategy that permits to stabilize each of the different subsystems ($\Sigma_S$, $\Sigma_I$, $\Sigma_F$ and $\Sigma_U$).

## V. Sequential Singular Perturbation Control Strategy

The control strategy goal consists in designing feedback control laws permits to follow known references in velocity ($V_{ref}$) and flight path angle ($\gamma_{ref}$). Following $\gamma_{ref}$ is attained by ensuring desired pitch rate, flight path angle and pitch angle dynamics with the use of the elevator deflection ($\delta$), while following $V_{ref}$ is achieved with the throttle position ($\delta_T$). The use of sequential time-scale decomposition permits to design control strategies for $\delta$ based on the sum of three components, $\delta = \delta_\theta + \delta_\gamma + \delta_q$, where each component is specifically designed to stabilize each one of the associated boundary layer subsystems, that is, $\delta_q = \Gamma_q(\chi, \theta, \gamma, q)$ for the ultrafast subsystem, Eq. (33), $\delta_\gamma = \Gamma_\gamma(\chi, \theta, \gamma)$ for the fast subsystem, Eq. (38), and $\delta_\theta = \Gamma_\theta(\chi, \theta)$ for the intermediate subsystem, Eq. (42).

In order to guarantee the validity of the sequential control strategy, a series of requirements on the control strategies need to be satisfied. The ultra-fast feedback control $\delta_q$ is designed to satisfy two crucial requirements, as seen in [1]: when the ultra-fast feedback function, $\delta_q$, is applied to the boundary layer Eq. (33), the closed-loop system should remain a standard SPS, which translates to that the equilibrium of the boundary layer

$$0 = f_q(\chi, \theta, \gamma, q, \Gamma_\theta + \Gamma_\gamma + \Gamma_q), \quad (45)$$

should have a unique root given by $\bar{q} = H_q(\theta, \gamma, \Gamma_\theta + \Gamma_\gamma)$ in $B_\chi \times B_\theta \times B_\gamma \times B_q$. This requirement assures that the choice of $\Gamma_q$ will not destroy this property of function $f_q$ in the open-loop system. The second requirement on $\Gamma_q(\chi, \theta, \gamma, q)$ is that it be *inactive* for $\bar{q} = H_q(\theta, \gamma, \Gamma_\theta + \Gamma_\gamma)$, that is

$$\Gamma_q\left[\chi, \theta, \gamma, H_q(\chi, \theta, \gamma, \Gamma_\theta + \Gamma_\gamma)\right] = 0. \quad (46)$$

Similarly, two requirements need to be satisfied by the (fast) control feedback $\delta_\gamma$ such that when applied to the boundary layer Eq. (38), the closed-loop system should remain a standard singularly perturbed system, which translates to that the equilibrium of the boundary layer

$$0 = f_\gamma(\chi, \theta, \gamma, \Gamma_\theta + \Gamma_\gamma), \quad (47)$$

should have a unique root given by $\bar{\gamma} = H_\gamma(\chi, \theta, \Gamma_\theta)$ in $B_\chi \times B_\theta \times B_\gamma$. This requirement assures that the choice of $\Gamma_\gamma$ will not destroy this property of function $f_\gamma$ in the open-loop system. The second requirement in $\delta_\gamma$, is that it be *inactive* for $\bar{\gamma} = H_\gamma(\chi, \theta, \Gamma_\theta)$, that is

$$\Gamma_\gamma\left[\chi, \theta, H_\gamma(\chi, \theta, \Gamma_\theta)\right] = 0. \quad (48)$$

With this in mind, the control strategy starts by applying the stretched time-scale $\tau_3$ resulting in the reduced order $\Sigma_{SIF}$-subsystem, Eqns. (29–32) with the boundary layer $\Sigma_U$-subsystem given by Eq. (33), and with the quasi-steady-state equilibrium given by Eq. (34). The reduced order order $\Sigma_{SIF}$-subsystem can be decomposed again by applying the stretched time-scale $\tau_2$ resulting in the reduced order $\Sigma_{SI}$-subsystem, Eqns. (35–37), and the boundary layer $\Sigma_F$-subsystem given by Eq. (38), with the equilibrium $H_\gamma(\chi, \theta, \delta_\theta)$ given by Eq. (39). The $\Sigma_{SI}$-subsystem, Eqns. (35–37), can be decomposed again by applying the last stretched time-scale $\tau_1$, resulting in the new reduced order $\Sigma_S$-subsystem, Eqns. (40–41), with the boundary layer $\Sigma_I$-subsystem given by Eq. (42).

Recall that according to Eqns. (46) and (48), $\Gamma_q$ and $\Gamma_\gamma$ become *innactive* when appearing in their respective equilibria in the $\Sigma_I$-subsystem, thus becoming $\delta = \delta_\theta$. The control signal $\delta_\theta$ is therefore selected as a feedback linearization signal for a target system of the form

$$\frac{\mathrm{d}\theta}{\mathrm{d}\tau_1} = -\tilde{b}_\theta \left(\theta - \theta_{ref}\right), \tag{49}$$

where $\tilde{b}_\theta = \varepsilon_1 b_\theta$ , with $b_\theta$ being the desired transient response for the $\Sigma_I$-subsystem, and $\theta_{ref}$ defined in terms of $V_{ref}$ and $\gamma_{ref}$ by the equilibrium analysis of the problem given by Eqns. (11–15). The control signal is therefore selected as

$$\begin{aligned} \delta_\theta &= B_1 \pm B_2 \sqrt{B_3 + B_4\theta + B_5\left(\theta - \theta_{ref}\right)} + B_6\theta \\ &+ B_7\left(\theta - \theta_{ref}\right), \end{aligned} \tag{50}$$

with $B_1 = \bar{a}_{16}^2 A_2^2 A_5/(2\bar{a}_{17}^2) - \bar{a}_{15}/\bar{a}_{17} + \bar{a}_{16}A_1/\bar{a}_{17}$, $B_2 = -\bar{a}_{16}A_2/(2\bar{a}_{17}^2)$, $B_3 = \left(\bar{a}_{16}A_2A_5\right)^2 - 4\bar{a}_{15}a_{17}A_5 + 4\bar{a}_{16}\bar{a}_{17}A_1A_5 + 4\bar{a}_{17}^2A_3$, $B_4 = 4\bar{a}_{17}^2A_4 - 4\bar{a}_{16}\bar{a}_{17}A_5$, $B_5 = 4\bar{a}_{17}\bar{a}_{18}A_5\tilde{b}_\theta$, $B_6 = -\bar{a}_{16}/\bar{a}_{17}$, and $B_7 = \bar{a}_{18}\tilde{b}_\theta/\bar{a}_{17}$, and where it can be shown that that positive solution of Eq. (50) is the right one. With the boundary layer $\Sigma_I$-subsystem stabilized and $\delta_\theta$ defined, the reduced order $\Sigma_S$-subsystem can be stabilized by selecting a feedback linearization control signal $\delta_T$ such that the velocity dynamics has a desired target dynamics of the form

$$\dot{V} = -b_V\left(V - V_{ref}\right), \tag{51}$$

with $b_V$ being the desired transient response for the $\Sigma_S$-subsystem, thus selecting

$$\begin{aligned} \delta_T &= -\frac{1}{a_1 + a_2V^2 + a_3V}\left[\left(a_4 + a_5\right)V^2 + V^2\left[\left(a_6\right.\right.\right. \\ &+ \left. a_7\left(H_\theta - \bar{H}_\gamma\right) + a_{10}\delta_\theta\right)\left(H_\theta - \bar{H}_\gamma\right) + \left(a_8 + a_9\delta_\theta\right)\delta_\theta\big] \\ &+ \left. a_{11}\left(\bar{H}_\gamma - \frac{\bar{H}_\gamma^3}{6}\right) + b_V\left(V - V_{ref}\right)\right]. \end{aligned} \tag{52}$$

The $\Sigma_F$-subsystem, Eq. (38) is stabilized by selecting the control signal $\delta_\gamma$, recalling that needs to satisfy Eqns. (47–48). This is achieved by substituting first $\delta_\theta$ into the $\Sigma_F$-subsystem, and rewriting the result in terms of its equilibrium $H_\gamma$, Eq. (39), by identifying that the original system is a function of the two possible solutions, that is

$$\frac{\mathrm{d}\gamma}{\mathrm{d}\tau_2} = -\frac{\bar{a}_{11}}{2V}\left[\gamma - H_\gamma(\chi,\theta,\delta_\theta)\right]\left[\gamma - \tilde{H}_\gamma(\chi,\theta,\delta_\theta)\right] + V\bar{a}_{14}\delta_\gamma, \tag{53}$$

Note also that according to Eq. (46), $\Gamma_q$ becomes *innactive* when appearing in its equilibrium in the $\Sigma_F$-subsystem, thus, $\delta = \delta_\theta + \delta_\gamma$, with $\delta_\theta$ being given by the control signal that stabilizes the $\Sigma_I$-subsystem, Eq. (50). Note that $H_\gamma(\chi,\theta,\delta_\theta)$ represents the positive solution of Eq. (39), while $\tilde{H}_\gamma(\chi,\theta,\delta_\theta)$ represents the disregarded negative solution, but both being necessary to complete the solution. The control signal $\delta_\gamma$ it is selected as a feedback linearization control signal for a selected target system of the form

$$\frac{\mathrm{d}\gamma}{\mathrm{d}\tau_2} = -\tilde{b}_\gamma\left[\gamma - H_\gamma(\chi,\theta,\delta_\theta)\right], \tag{54}$$

where $\tilde{b}_\gamma = \varepsilon_1\varepsilon_2 b_\gamma$, with $b_\gamma$ being the desired transient response for the $\Sigma_F$-subsystem. The choice of this target dynamics will satisfy that the choice of $\Gamma_\gamma$ will not destroy the property that the closed-loop system should have a unique equilibrium $\bar{\gamma} = H_\gamma(\chi,\theta,\Gamma_\gamma)$, and that $\Gamma_\gamma$ it be *inactive* for $\bar{\gamma} = H_\gamma(\chi,\theta)$, resulting in

$$\delta_\gamma = \frac{\left(\gamma - H_\gamma\right)\left[-\tilde{b}_\gamma + \frac{\bar{a}_{11}}{2V}\left(\gamma - \tilde{H}_\gamma\right)\right]}{\bar{a}_{14}V}. \tag{55}$$

With control signal $\delta_\gamma$ selected, the control signal $\delta_q$ that stabilizes the $\Sigma_U$-subsystem can be selected by ensuring requirements (45-46). Recall that for the $\Sigma_U$-subsystem, $\delta = \delta_\theta + \delta_\gamma + \delta_q$, therefore, by substituting $\delta_\theta$ and $\delta_\gamma$, Eqns. (50) and (55), respectively, into the $\Sigma_U$-subsystem, and rewriting it using the definition of the $H_q$ QSSE, Eq. (34), results in

$$\frac{\mathrm{d}q}{\mathrm{d}\tau_3} = V^2\left[\bar{a}_{18}\left[q - H_q(\chi,\theta,\gamma,\delta_\theta + \delta_\gamma)\right] + \bar{a}_{17}\delta_q\right]. \tag{56}$$

Similarly, in order to satisfy Eqns. (45-46) on the control signal $\delta_q$, lets choose a feedback linearization control signal for a selected target system of the form

$$\frac{\mathrm{d}q}{\mathrm{d}\tau_3} = -\tilde{b}_q\left[q - H_q(\chi,\theta,\gamma,\delta_\theta + \delta_\gamma)\right], \tag{57}$$

where $\tilde{b}_q = \varepsilon_1\varepsilon_2\varepsilon_3 b_q$ , with $b_q$ being the desired transient response for the $\Sigma_U$-subsystem, resulting in

$$\delta_q = -\frac{\left(\tilde{b}_q + \bar{a}_{18}V^2\right)\left[q - H_q(\chi,\theta,\gamma,\delta_\theta + \delta_\gamma)\right]}{V^2\bar{a}_{17}}. \tag{58}$$

This finalizes the control strategy with the control signals given by $\delta = \delta_\theta + \delta_\gamma + \delta_q$, Eqns. (50), (55), and (58), respectively, and $\delta_T$, Eq. (52). Following section provides some simulation results.

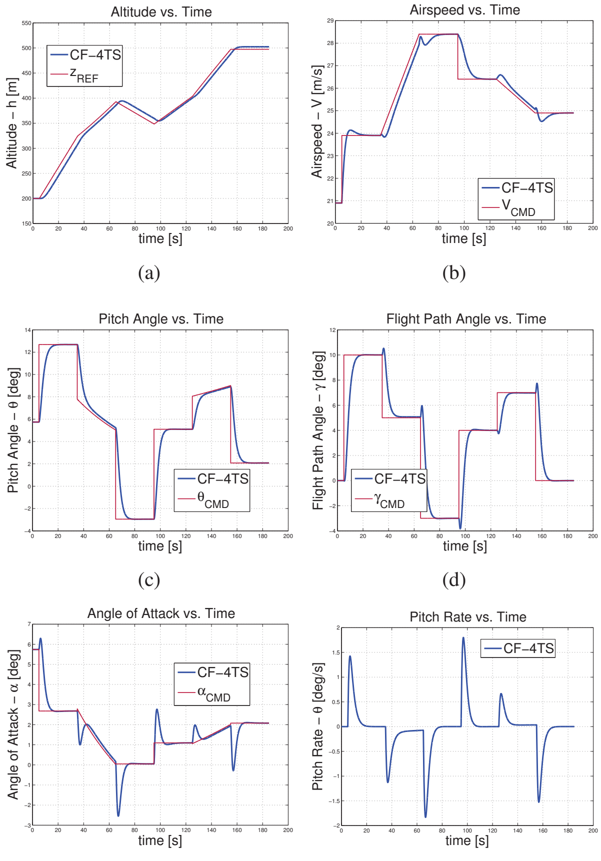
## VI. Numerical Results

This section presents some results corresponding to the simulations conducted to analyze the proposed control law. The numerical simulation uses a fourth-order Runge-Kutta fixed step integration method, with a time step of 0.001 seconds, written in the $MATLAB$ interface. The analysis includes variation in the references $V_{ref}$ and $\gamma_{ref}$. Actuator saturations are also included, namely $-40^\circ \leq \delta \leq 40^\circ$, and $0 \leq \delta_T \leq 1$.

Different cases will be considered: varying $V_{ref}$, while maintaining $\gamma_{ref} = 0$; varying $\gamma_{ref}$, while maintaining $V_{ref} = const$; and varying both $\gamma_{ref}$ and $V_{ref}$. Results for this last case are presented in Fig. 1 and 2, where constant acceleration and deceleration references are also generated, with 30 seconds per maneuver. Figure 1 shows the states: the altitude and aerodynamic airspeed, on the top row, pitch and flight path angle on the middle row, angle of attack and pitch rate on the bottom row; and Fig. 2 shows the control: elevator deflection and throttle position. The different variable reference set points are presented with a thinner red line. Note that it is also included a reference in altitude given by $h_{ref} = V_{ref}\sin\gamma_{ref}$, despite that the control strategy is focused on following both by separate, but serves to indicate that future control strategies will be derived so follow altitude profiles. Despite the complex reference profiles, the control strategy is able to follow them in both aerodynamic velocity and flight path angle. Also note that saturations are avoided by selecting the appropriate desired dynamic coefficients, $b_V = b_\theta = b_\gamma = b_q = 0.35$.

## VII. Conclusions

The presented control strategy permits to drive the UAV to follow variable references in both aerodynamic velocity and flight path angle by using a sequential strategy that permits to easily obtain appropriate feedback control laws that stabilize each of the subsystems. The control strategy provides a closed-form solution. The simulations were conducted using a realistic model of the Cefiro aircraft developed by the Dept. of Aerospace Engineering at the University of Seville [14], which will the platform where the future flight tests and validation of the control strategies will be conducted.

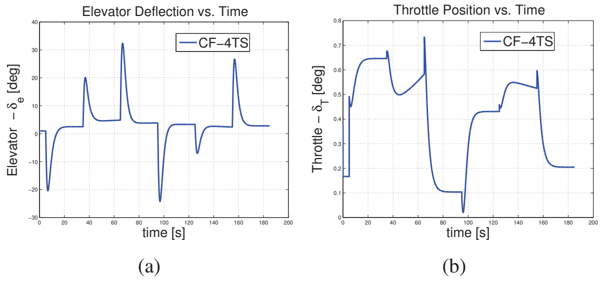


(a) (b)

(c) (d)

(e) (f)

Fig. 1. State history for simulations with variable $V$ and $\gamma$.



(a) (b)

Fig. 2. Control history for simulations with variable $V$ and $\gamma$.

## References

[1] P. Kokotović, H. Khalil, and J. O'reilly, *Singular perturbation methods in control: analysis and design*. Society for Industrial Mathematics, 1999, pp. 189–320.

[2] S. Esteban, F. Gordillo, and J. Aracil, "Three-Time Scale Singular Perturbation Control and Stability Analysis for an Autonomous Helicopter on a Platform," *International Journal of Robust and Nonlinear Control*, pp. 1–34, 2012, accepted for publication 11 March 2012.

[3] R. Brockett, "Feedback invariants for nonlinear systems," in *A link between science and applications of automatic control: proceedings of the seventh Triennial World Congress of the IFAC, Helsinki, Finland*. IFAC, June 1978, pp. 1115–1120.

[4] J. Buffington, A. Sparks, and S. Banda, "Full conventional envelope longitudinal axis flight control with thrust vectoring," in *American Control Conference*. IEEE, 1993, pp. 415–419.

[5] H. Sira-Ramírez, M. Zribi, and S. Ahmad, "Dynamical Sliding Mode Control Approach for Vertical Flight Regulation in Helicopters," *IEE Proceedings-Control Theory & Applications*, vol. 141, no. 1, pp. 19–24, 1994.

[6] H. Khalil, *Nonlinear Systems*. Prentice Hall, 1996.

[7] F. Gavilan, J. Acosta, and R. Vazquez, "Control of the longitudinal flight dynamics of an uav using adaptive backstepping," in *IFAC World Congress, Milan, Italy*, 2011.

[8] S. Balakrishnan and V. Biega, "Adaptive-critic-based neural networks for aircraft optimal control," *Journal of Guidance, Control, and Dynamics*, vol. 19, no. 4, pp. 893–898, 1996.

[9] S. Balakrishnan and S. Esteban, "Nonlinear flight control systems with neural networks," in *Proceedings of the AIAA Guidance, Navigation and Control Conference and Exhibit, Montreal*. AIAA, August 2001.

[10] D. Naidu and A. Calise, "Singular perturbations and time scales in guidance and control of aerospace systems: A survey," *Journal of Guidance, Control and Dynamics*, vol. 24, no. 6, pp. 1057–1078, 2001.

[11] R. Mehra, R. Washburn, S. Sajan, and J. Carrol, "A study of the application of singular perturbation theory," NASA CR-3167, 1979.

[12] S. Bertrand, N. Guénard, T. Hamel, H. Piet-Lahanier, and L. Eck, "A hierarchical controller for miniature vtol uavs: Design and stability analysis using singular perturbation theory," *Control Engineering Practice*, 2011.

[13] S. Esteban, "Three-time-scale Control of an Autonomous Helicopter on Platform," Automatics, Robotics and Telematic Ph.D., Universidad de Sevilla, Sevilla, Spain, July 2011.

[14] C. Bernal, A. Fernandez, P. Lopez, A. Martin, D. Perez, F. Samblas, S. Esteban, F. Gavilan, and D. Rivas, "Cefiro: an aircraft design project in the university of seville," in *9th European Workshop on Aircraft Design Education (EWADE 2009)*, 2009.

[15] B. Etkin and L. Reid, "Dynamics of flight: stability and control."

[16] J. Roskam, *Airplane flight dynamics and automatic flight controls*. DARcorporation, 2001.

[17] A. Calise, "A singular perturbation analysis of optimal thrust control with proportional navigation guidance," in *IEEE Conference on Decision and Control*. IEEE, 1977, pp. 1167–1176.

[18] ——, "Optimization of aircraft altitude and flight-path angle dynamics," *Journal of Guidance, Control, and Dynamics(ISSN 0731-5090)*, vol. 7, pp. 123–125, 1984.

[19] M. Ardema and N. Rajan, "Separation of Time-Scales in Aircraft Trajectory Optimization," *Journal of Guidance, Control, and Dynamics*, vol. 8, no. 2, pp. 275–278, 1985.

[20] ——, "Slow and fast state variables for three-dimensional flight dynamics," *Journal of Guidance, Control, and Dynamics*, vol. 8, no. 4, pp. 532–535, 1985.

[21] M. Heiges, P. Menon, and D. Schrager, "Synthesis of a helicopter full-authority controller," *Journal of Guidance, Control, and Dynamics*, vol. 15, no. 1, pp. 222–227, 1992.

[22] B. Sridhar and N. Gupta, "Missile guidance laws based on singular perturbation methodology," *Journal of Guidance and Control*, vol. 3, pp. 158–165, 1980.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Robust Stabilization of Networked Control Systems using the Markovian Jump System Approach

Ashraf Khalil
School of Electrical, Electronic and Computer Engineering
University of Birmingham
Birmingham, UK
ashrafkhalilg@gmail.com

Jihong Wang
School of Engineering
University of Warwick
Coventry, UK
jihong.wang@warwick.ac.uk

*Abstract*— **The key feature of Networked Control Systems (NCSs) is that the information is exchanged through a network among control system components. Transmitting control signals through shared networks induces time delays and data losses which may destabilize the system. This time delay may be constant periodic or random. The random time delay can be modeled using Markov Chains and the NCS can be modeled as Markovian jump system. The stochastic stability of the system has the form of Bilinear Matrix Inequality (BMI). The V-K iteration algorithm is used to solve the BMI and hence to design the stabilizing controller. A modified V-K iteration algorithm is presented in this paper where the decay rate is maximized in both the V- and K-loops. The V-K algorithm method is applied to the cart and inverted pendulum problem which shows that the decay rate is improved with the modified algorithm.**

***Keywords-component; networked control system, time delay, Markov , random time delay, jump, stability***

## I. Introduction

The advances in communication and network technology, and the availability of high speed computers have resulted in an increasing interest in NCSs. This type of control systems can be defined as a control system where the control loop is closed through a real-time communication network [1]. In NCSs the reference input, plant output and control input are exchanged through a real-time communication network as shown in Figure 1. The main advantages of NCSs are modularity, simplified wiring, low cost, reduced weight, decentralization of control, integrated diagnosis, simple installation, quick and easy for maintenance [2], flexible expandability with low cost. NCSs are able to easily fuse global information to make intelligent decisions over large physical spaces.

As the control loop is closed through a communication network the time delay and data dropout are unavoidable. This may degrade the performance of NCSs or even destabilize the system. In general, the control systems with time delays can be classified into time delay independent where the stability is not affected by the time delay and time delay dependent where the time delay affects the stability [3]. Time delay, no doubt, increases the complexity in the analysis and the design of NCSs. There are many methods in the literature for studying the stability of NCS, see for example [4]-[5]. Among these methods is the Markovian jump system approach which is mostly used to study the stability and stabilization of system with abrupt changes due to the variations in the system structure or partly system failure. In this way the system will have a number of models or modes and jumps from one mode to another in a random fashion and in many cases the jump parameter can be modeled using Markov Chains. In NCS the time delay can be random and because there is a correlation between the previous, current and next time delay, the time delay can be modeled as a Markov Chain.

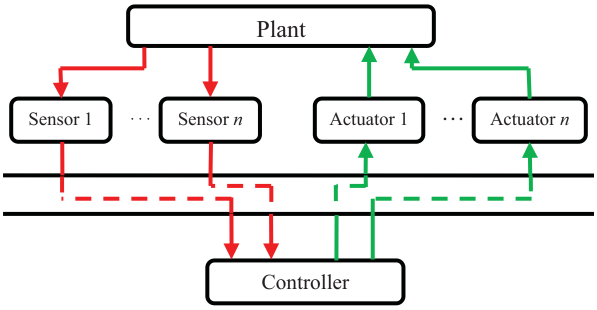


Figure 1. A Typical networked control system

The application of the discrete-time jump system in NCSs has been addressed in many papers, see for example [6-9]. In [6-8] the discrete-time model is augmented and the generated output feedback problem is formulated as BMI which is solved using the V-K iteration algorithm. In this paper we adopt the algorithm in [8] with some modification to the V-K iteration loop. The method in [8] is limited to time delays which are less than the sampling period and in [6][7] the method is extended to time delays larger than the sampling period. From control engineering point of view when the time delay larger than the sampling time the system performance is not acceptable. In [10] the authors use the discrete model for the plant and both the time delay between the sensor to controller and from the controller to the actuator are considered. The discrete mode dependent Lyapunov function has been used to derive a stabilizing switching controller. In [9] the authors concentrate on the problem of the random data drop outs and the sufficient conditions for the mean square stability are derived. The stability analysis and controller design with two random time delays are studied in [10-13]. In [12][13] the NCS is modeled where both the time delays are considered. The controller depends on the current sensor to controller time delay and the previous controller to actuator time delay and hence the controller depends on the three random variables, $\tau_k$ , $d_k$ , $d_{k-\tau_{k+1}}$ , which are interdependent. The resulting system cannot be regarded as the standard Discrete-Time Markovian Jump Linear System (DTMJLS). The derived theorem is in a set of LMI with nonlinear LMI constraint which are non-convex and can be solved by iterative algorithm such as the Cone Complementary Linearization (CCL). The optimal stochastic control is studied in [14] where the optimal stochastic controller is derived when the time delay is random. The use of the model predictive control in NCS has been studied in [15-17] where both the sensor to controller and controller to actuator time delays are considered.

The paper starts from the model of the NCS as DTMJLS is presented where the time delay is modeled as Markov Chain. Then the stability of the system is formulated as BMIs. The V-K algorithm is explained to solve the BMIs and it is tested on the cart and inverted pendulum problem.

## II. Mathematical Modelling of NCS with Time Delay

### A. NCS Systems with State Feedback Controller

The model of a single loop networked control system is shown in Figure 2. The measured plant signals are transmitted through the network and they will suffer random time delays and some of them may be lost. The random time delay makes the system to have the nature of a stochastic hybrid system. The discrete time-invariant plant model is given by:

$$\mathbf{x}(k+1) = \mathbf{A}_d\mathbf{x}(k) + \mathbf{B}_d\mathbf{u}(k) \qquad (1)$$

where $\mathbf{x}(k) \in \Re^n$ the system state vector, $\mathbf{u}(k) \in \Re^m$ the system control input, and the matrices $\mathbf{A}_d$ and $\mathbf{B}_d$ are given by:

$$\mathbf{A}_d = e^{\mathbf{A}h} \qquad \mathbf{B}_d = \int_0^h e^{\mathbf{A}(h-s)}\mathbf{B}ds \qquad (2)$$

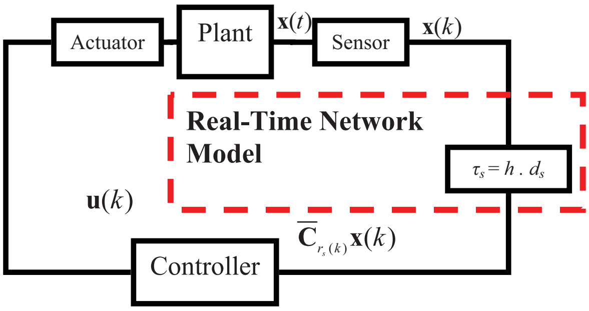


Figure 2. The networked control system.

In the model shown in Figure 2 the time delays are lumped together between the sensor and the controller. In many of the published work in the literature the time delay between the controller and the actuator is neglected. For the following analysis the following assumptions are required and are made:

Assumption 1:

- The sensors are clock driven. The actuator and the controller are event driven; which means that the sensors sample the plant states periodically and the actuators and the controllers use the data as soon as they arrive.
- The data are sent as a single packet.
- The data are received in chronological order which means that old data are disregarded.

The mode-dependent switching state feedback control law is given by:

$$\mathbf{u}(k) = \mathbf{K}(r_s(k))\mathbf{x}(k - r_s(k)) \qquad (3)$$

where $\tau(k) = r_s(k) \cdot h$ , $h$ is the sampling period and $r_s(k)$ is a bounded random integer sequence governed by Markov Chains with $0 \le r_s(k) \le d_s < \infty$, and $d_s$ is the finite delay bound. By augmenting the state variable:

$$\overline{\mathbf{x}}(k) = \begin{bmatrix}\mathbf{x}(k)^T & \mathbf{x}(k-1)^T & \cdots & \mathbf{x}(k-d_s)^T\end{bmatrix}^T$$

where $\overline{\mathbf{x}}(k) \in R^{(d_s+1)n}$ , applying the controller (3) into (1) the closed-loop system becomes:

$$\overline{\mathbf{x}}(k+1) = (\overline{\mathbf{A}} + \overline{\mathbf{B}}\mathbf{K}(r_s(k))\overline{\mathbf{C}}(r_s(k)))\overline{\mathbf{x}}(k) \qquad (4)$$

where;

$$\overline{\mathbf{A}} = \begin{bmatrix} \mathbf{A} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} & \mathbf{0} \end{bmatrix} \overline{\mathbf{B}} = \begin{bmatrix} \mathbf{B} \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}$$

$$\overline{\mathbf{C}}(r_s(k)) = \begin{bmatrix}\mathbf{0} & \cdots & \mathbf{0} & \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0}\end{bmatrix}$$

$\overline{\mathbf{C}}(r_s(k))$ incorporates the time delay into the model and has all elements being zero except for the $r_s(k)$th block being an identity matrix. The closed-loop system (4) can be rewritten as;

$$\overline{\mathbf{x}}(k+1) = \mathbf{A}_{cl}(r_s(k))\overline{\mathbf{x}}(k) \qquad (5)$$

### B. NCS with Dynamic Output Feedback Controller

Stabilizing the plant (1) with a dynamic controller as shown in Figure 3, the dynamic controller model is given by:

$$\begin{aligned}\mathbf{z}(k+1) &= \mathbf{F}\mathbf{z}(k) + \mathbf{G}\mathbf{y}(k) \\ \mathbf{v}(k) &= \mathbf{H}\mathbf{z}(k) + \mathbf{J}\mathbf{y}(k)\end{aligned} \qquad (6)$$

In the case of the dynamic controller, both the time delay from the sensors to the controller and from the controller to the actuators are considered. Augmenting the controller states as;

$$\overline{\mathbf{z}}(k) = \left[\mathbf{z}(k)^T \quad \mathbf{v}(k)^T \quad \cdots \quad \mathbf{v}(k-d_{ca})^T\right]^T$$

The controller model with the augmenting states is then given by:

$$\begin{aligned} \overline{\mathbf{z}}(k+1) &= \overline{\mathbf{F}}\overline{\mathbf{z}}(k) + \overline{\mathbf{G}}\mathbf{y}(k) \\ \mathbf{u}(k) &= \overline{\mathbf{H}}(r_{ca}(k))\overline{\mathbf{z}}(k) + \overline{\mathbf{K}}(r_{ca}(k))\mathbf{y}(k) \end{aligned} \qquad (7)$$

where;

$$\overline{\mathbf{F}} = \begin{bmatrix} \mathbf{F} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{H} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} & \mathbf{0} \end{bmatrix} \overline{\mathbf{G}} = \begin{bmatrix} \mathbf{G} \\ \mathbf{J} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}$$

$$\overline{\mathbf{H}}(r_{ca}(k)) = \begin{cases} [\mathbf{H} \quad \mathbf{0} \quad \cdots \quad \mathbf{0} \quad \mathbf{0} \quad \cdots \quad \mathbf{0}] & \text{if} \quad r_{ca}(k) = 0 \\ [\mathbf{0} \quad \cdots \quad \mathbf{0} \quad \mathbf{I} \quad \mathbf{0} \quad \cdots \quad \mathbf{0}] & \text{if} \quad r_{ca}(k) \neq 0 \end{cases}$$

$$\overline{\mathbf{K}}(r_{ca}(k)) = \begin{pmatrix} \mathbf{J} & \text{if} \quad r_{ca}(k) = 0 \\ \mathbf{0} & \text{if} \quad r_{ca}(k) \neq 0 \end{pmatrix}$$

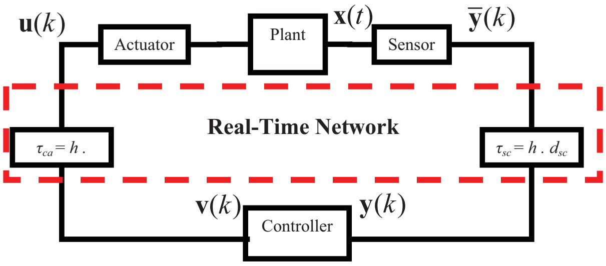


Figure 3. Networked control system with both time delays from sensor to controller and from controller to actuator are taking into account

When the time stamping is used, $\mathbf{F}$, $\mathbf{G}$, $\mathbf{H}$ and $\mathbf{J}$ are replaced by $\mathbf{F}(\tau_{sc})$, $\mathbf{G}(\tau_{sc})$, $\mathbf{H}(\tau_{sc})$, and $\mathbf{J}(\tau_{sc})$. The augmented plant model with output feedback can be described by:

$$\begin{aligned} \tilde{\mathbf{x}}(k+1) &= \tilde{\mathbf{A}}\tilde{\mathbf{x}}(k) + \tilde{\mathbf{B}}\mathbf{u}(k) \\ \mathbf{y}(k) &= \mathbf{C}\tilde{\mathbf{C}}(r_{sc}(k))\tilde{\mathbf{x}}(k) \end{aligned} \qquad (8)$$

Augmenting both the plant states and controller states as: $\overline{\mathbf{x}}(k) = \left[\tilde{\mathbf{x}}(k)^T \quad \overline{\mathbf{z}}(k)^T\right]^T$. The closed loop system with the plant (8) and the controller (7) becomes;

$$\overline{\mathbf{x}}(k+1) = (\overline{\mathbf{A}} + \overline{\mathbf{B}}\mathbf{K}(r_{ca}(k))\overline{\mathbf{C}}(r_{sc}(k)))\overline{\mathbf{x}}(k) \qquad (9)$$

where;

$$\overline{\mathbf{A}} = \begin{bmatrix} \tilde{\mathbf{A}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \overline{\mathbf{B}} = \begin{bmatrix} \tilde{\mathbf{B}} & \mathbf{0} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \overline{\mathbf{C}}(r_{sc}(k)) = \begin{bmatrix} \mathbf{0} & I \\ \mathbf{C}\tilde{\mathbf{C}}(r_{sc}(k)) & \mathbf{0} \end{bmatrix}$$

$$\mathbf{K}(r_{ca}(k)) = \begin{bmatrix} \tilde{\mathbf{F}} & \tilde{\mathbf{G}} \\ \tilde{\mathbf{H}}(r_{ca}(k)) & \tilde{\mathbf{K}}(r_{ca}(k)) \end{bmatrix}$$

Equation (9) can be written as;

$$\overline{\mathbf{x}}(k+1) = \mathbf{A}_{cl}(r_{sc}(k), r_{ca}(k))\overline{\mathbf{x}}(k) = \mathbf{A}_{cl}(r_s(k))\overline{\mathbf{x}}(k) \qquad (10)$$

The two time delays are random and bounded, $\tau_{scm} \geq \tau_{sc} \geq 0$ and $\tau_{cam} \geq \tau_{ca} \geq 0$. These can be modelled as two homogeneous Markov Chains and they jump from mode to mode according to their transition probabilities $P_{sc}$ and $P_{ca}$ respectively. The random variable $\tau_{sc}$ and $\tau_{ca}$ can be converted to single random variable, $r(k)$ where the transition probability, $P$, is given by Kronecker product of the $P_{sc}$ and $P_{ca}$ as;

$$P = P_{sc} \otimes P_{ca} \qquad (11)$$

For simplicity (10) can be written as:

$$\mathbf{x}(k+1) = \mathbf{A}(r(k))\mathbf{x}(k) \qquad (12)$$

Equations (5) and (10) are standard DTMJLS. Equation (5) is a jump system with one mode which is the sensor to the controller time delay while the system in (10) has two modes which are the sensor to the controller and the controller to the actuator time delays. The system matrix will be $\mathbf{A}_{cl}(r(k))$□$\{\mathbf{A}_{cl}(0),\ldots,\mathbf{A}_{cl}(d)\}$ according to the jump parameter $r(k)$□$\theta = \{0,\ldots,d\}$. In order to stabilize the system with mode-independent or mode-dependent controller the mean square stability must be established.

### *C. The model of the random time delay as Markov Chain*

The random time delay is modelled as a finite state Markov process with the following properties:

$$\begin{aligned} &P\{r_s(k+1) = j \;\;|\;\; r_s(k) = i\} = p_{ij} && 0 \leq i, j \leq d_s \\ &0 \leq p_{ij} \leq 1 && \sum_{j=0}^{d} p_{ij} = 1 \end{aligned} \qquad (13)$$

where $d_s$ is the number of modes and $r_s(k)$ is the Markovian process. The general transition probability matrix is given by:

$$P = \begin{bmatrix} p_{00} & p_{01} & 0 & 0 & \cdots & 0 \\ p_{10} & p_{11} & p_{12} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & 0 \\ p_{d0} & p_{d1} & p_{d2} & p_{d3} & \cdots & p_{d_s d_s} \end{bmatrix} \qquad (14)$$

The constraint (13) means the summation of the probabilities in every row is one. The assumption made is that the old data are discarded. Suppose that at instant $k$ we received $\mathbf{x}(k)$, at $k$+1 if there is no new data then the old data will be used by the controller, but if we receive $\mathbf{x}(k$ -1) at $k$ +1 then it will be older than $\mathbf{x}(k)$ and hence $\mathbf{x}(k\ -1)$ must be discarded, this can be interpreted as;

$$p\{r_s(k+1) > r_s(k)+1\} = 0 \qquad (15)$$

From (15) the time delay can increase only at one step but it can decrease as many steps as can be seen from (14). The diagonal elements in (14) represent the probability of successive equal time delays or in other words the probability that the network remain in the same state. The upper diagonal elements represent the possibility of receiving longer delays or increasing the network load. The zero elements represent the discard of the old data.

## III. The Stability of the Discrete-Time Markovian Jump Linear System (DTMJLS)

The Mean Square stability of the Markovian Jump Systems is equivalent to the Asymptotic Wide Sense Stationary Stability (AWSS) [18]. For the jump system the stochastic stability, mean square stability and the exponential mean square stability are all equivalent and every condition implies the almost sure (asymptotic) stability.

Definition 1: [6]

The system (12) is mean square stable if for every initial condition state, $(\mathbf{x}_0, r_0)$,

$$\lim_{k\to\infty} \mathrm{E}\left[\|\mathbf{x}(k)\|^2\right] = 0 \qquad (16)$$

Definition 2: [6]

The system (12) is mean square stable with decay rate $\beta$ [19] if for every initial condition state, $(\mathbf{x}_0, r_0)$,

$$\lim_{k\to\infty} \beta^k \mathrm{E}\left[\|\mathbf{x}(k)\|^2\right] = 0 \qquad \beta > 1 \qquad (17)$$

The necessary and sufficient conditions for mean square stability for jump system are given in the following theorem.

Theorem 1 [18]: The mean square stability of system (12) is equivalent to the existence of symmetric positive definite matrices $\mathbf{Q}_0, \ldots, \mathbf{Q}_d$ satisfying any one of the following 4 conditions:

$$\mathbf{A}_i\left(\sum_{j=0}^{d} p_{ji}\mathbf{Q}_\mathbf{j}\right)\mathbf{A}_i^T < \mathbf{Q}_i, \qquad i = 0,\cdots,d$$

$$\mathbf{A}_j^T\left(\sum_{i=0}^{d} p_{ji}\mathbf{Q}_i\right)\mathbf{A}_j < \mathbf{Q}_j, \qquad j = 0,\cdots,d$$

$$\sum_{j=0}^{d} p_{ji}\mathbf{A}_j\mathbf{Q}_j\mathbf{A}_i^T < \mathbf{Q}_i, \qquad i = 0,\cdots,d$$

$$\sum_{j=0}^{d} p_{ji}\mathbf{A}_i^T\mathbf{Q}_j\mathbf{A}_i < \mathbf{Q}_j, \qquad i = 0,\cdots,d$$

Where $i$=0,…,$d$ represents the number of the modes. The conditions 1-4 are equivalent for studying the stability of the DTMJLS but for the controller design each condition will lead to a different controller. Choosing condition (4) in Theorem 1 and replacing $\mathbf{Q}_i$ by $\alpha\mathbf{Q}_i$ (where the decay rate or Lyapunov Exponent, $\beta = 1/\alpha$ and $\lim_{k\to\infty}\beta^k M(k) = 0$) on the right hand side, the closed-loop system becomes:

$$\sum_{j=0}^{d} p_{ji}(\mathbf{A}_i + \mathbf{B}_i\mathbf{K}_i\mathbf{C}_i)^T\mathbf{Q}_j(\mathbf{A}_i + \mathbf{B}_i\mathbf{K}_i\mathbf{C}_i) < \alpha\mathbf{Q}_j, \ i=0,\ldots,d \ (18)$$

The coupled equations (18) are BMIS which are non convex and finding a global optimal solution is very difficult. However many control problems are formulated as BMIs, there are a few methods for solving the BMIs. For example the path-following linearization method reported in [20] can be used where each matrix is perturbed and the higher order terms are neglected. The most widely used techniques for the solution is by iteration methods such the D-K, G-K and V-K iteration algorithms [21]. If we fix $\mathbf{K}_i$ ($i$=0,…,d) then we have a Generalized Eigenvalue Problem (GEVP) and if we fix $\mathbf{Q}_i$ ( i=0,…,d) then we have Eigenvalue Problem (EVP) [8]. Both of these problems can be solved very efficiently using the Matlab LMI toolbox. Using Schur complement to (18) then we have:

$$\begin{bmatrix} \alpha\mathbf{Q}_j & (\mathbf{A}_0 + \mathbf{B}_0\mathbf{K}_0\mathbf{C}_0)^T & \cdots & (\mathbf{A}_d + \mathbf{B}_d\mathbf{K}_d\mathbf{C}_d)^T \\ (\mathbf{A}_0 + \mathbf{B}_0\mathbf{K}_0\mathbf{C}_0) & p_{j0}^{-1}\mathbf{Q}_0^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ (\mathbf{A}_d + \mathbf{B}_d\mathbf{K}_d\mathbf{C}_d) & 0 & \cdots & p_{jd}^{-1}\mathbf{Q}_d^{-1} \end{bmatrix} > 0 \qquad (19)$$

## IV. The V-K Iteration Algorithm

In the V-K algorithm the BMI is divided into two LMI's and by solving these two LMI's a local optimal solution can be found. The problem solution process is divided to three basic problems which are: FP (Feasibility Problem), EVP, and GEVP that can be solved using the Matlab LMI toolbox. In the V-K algorithm, the problem is iterated between the EVP and the GEVP. The proof of the algorithm convergence is given in [21]. The detailed algorithm is shown in the flowchart in Figure 4. The algorithm starts with the initialization, then if the solution is feasible the EVP and GEVP are iterated until the desired transition matrix is reached. In this improved algorithm the decay rate is maximized in both the EVP and GEVP. The initial transition probability matrix is:

$$P_0 = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \ldots & 0 \end{bmatrix} \approx \begin{bmatrix} 1-n\cdot\in & \in_1 & \ldots & \in_n \\ 1-n\cdot\in & \in_1 & \ldots & \in_n \\ \vdots & \vdots & \ddots & \vdots \\ 1-n\cdot\in & \in_1 & \ldots & \in_n \end{bmatrix}$$

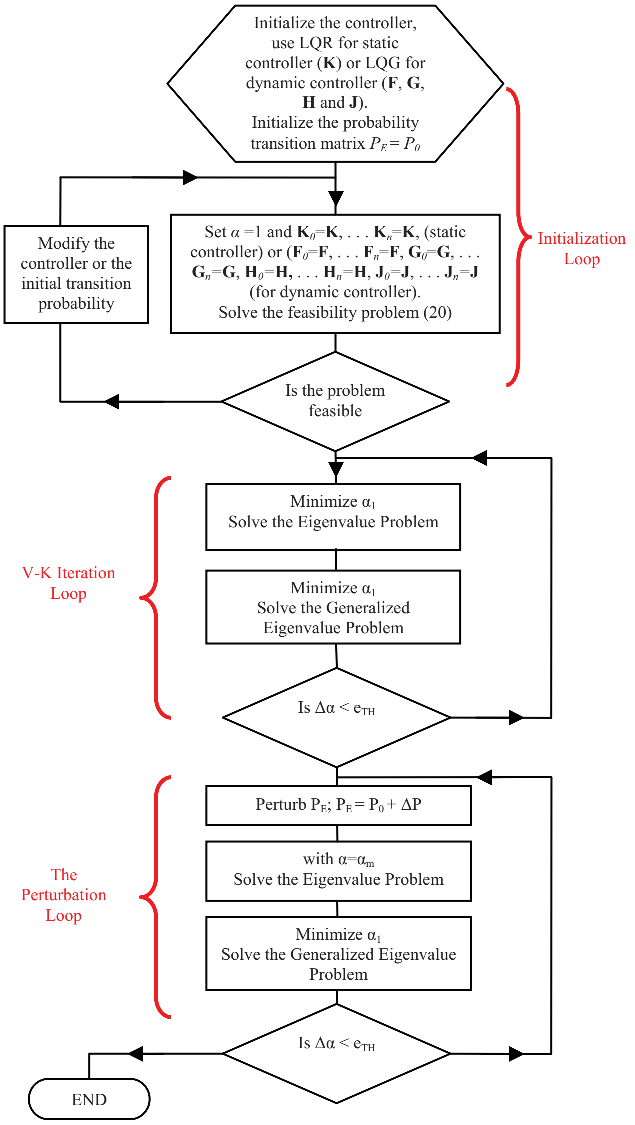

Initialize the controller, use LQR for static controller (**K**) or LQG for dynamic controller (**F**, **G**, **H** and **J**). Initialize the probability transition matrix $P_E = P_0$
Modify the controller or the initial transition probability
Set $\alpha$ =1 and $\mathbf{K}_0$=**K**, . . . $\mathbf{K}_n$=**K**, (static controller) or ($\mathbf{F}_0$=**F**, . . . $\mathbf{F}_n$=**F**, $\mathbf{G}_0$=**G**, . . . $\mathbf{G}_n$=**G**, $\mathbf{H}_0$=**H**, . . . $\mathbf{H}_n$=**H**, $\mathbf{J}_0$=**J**, . . . $\mathbf{J}_n$=**J** (for dynamic controller). Solve the feasibility problem (20)
Initialization Loop
Is the problem feasible
V-K Iteration Loop
Minimize $\alpha_1$ Solve the Eigenvalue Problem
Minimize $\alpha_1$ Solve the Generalized Eigenvalue Problem
Is $\Delta\alpha < e_{TH}$
The Perturbation Loop
Perturb $P_E$; $P_E = P_0 + \Delta P$
with $\alpha=\alpha_m$ Solve the Eigenvalue Problem
Minimize $\alpha_1$ Solve the Generalized Eigenvalue Problem
Is $\Delta\alpha < e_{TH}$
END

Figure 4. The V-K iteration algorithm

It should be noted that the initial controller is designed for the free delay system. To get an initial feasible solution we have to start from small time delays and perturb the transition probability matrix toward higher time delays. The perturbation $\in$ should be very small positive number in the order of 0.005. An example of the perturbation matrix is:

$$\Delta P_0 = \begin{bmatrix} -s & s & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix} \quad \Delta P_1 = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ -s & s & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

As can be seen the sum of the perturbation through any row is zero. More aggressive initial transition probability matrix can be used. In [7][8] the perturbation is around 0.01 but even with this small perturbation sometimes the problem is divergent and we need to use smaller perturbation, for example around 0.005. Also for the two modes, the two probability matrices are perturbed at the same time while in our algorithm they are perturbed separately.

Example 1

The pendulum mass is denoted by $m$ and the cart mass is $M$, the length of the pendulum rod is $L$. The open loop system is unstable. The states are defined as $x_1 = x$, $x_2 = \dot{x}$, $x_3 = \theta$, $x_4 = \dot{\theta}$. The linearized model can be given as:

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & \dfrac{-mg}{M} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & \dfrac{(M+m)g}{ML} & 0 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ \dfrac{1}{M} \\ 0 \\ \dfrac{-1}{ML} \end{bmatrix} u = \mathbf{A}\mathbf{x} + \mathbf{B}u$$

$$\mathbf{y} = \begin{bmatrix} x \\ \theta \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \mathbf{x} = h(x,u)$$

The parameters used are: $M$=1 kg, $m$ = 0.4 kg, L = 0.7 m. The sampling time is $h = 0.1$ s. The time delay is bounded by 2: $r_s(k) \in \{0,1,2\}$. The initial condition is $x$ = 0 and $\theta$ = 0.1. After sampling the system with 0.1 s sampling rate, the system matrices are given by:

$$\mathbf{A} = \begin{bmatrix} 1 & 0.1 & -0.0199 & -0.0007 \\ 0 & 1 & -0.4049 & -0.0199 \\ 0 & 0 & 1.0996 & 0.1033 \\ 0 & 0 & 2.0247 & 1.0996 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 0.0050 \\ 0.1009 \\ -0.0073 \\ -0.1476 \end{bmatrix}$$

The required transition probability is given by:

$$P = \begin{bmatrix} 0.5 & 0.5 & 0 \\ 0.3 & 0.6 & 0.1 \\ 0.3 & 0.6 & 0.1 \end{bmatrix}$$

Using the LQR matlab function with **Q**=**I** and **R**=1. The controller is given by:

$$\mathbf{K}_{LQR} = \begin{bmatrix} 0.5943 & 1.4745 & 28.7321 & 6.7849 \end{bmatrix}$$

with the required transition probability and the LQR controller does not stabilize the system with the time delay because the solution is infeasible, the initial transition probability and the perturbation matrix are chosen as:

$$P_0 = \begin{bmatrix} 0.499 & 0.499 & 0.002 \\ 0.4 & 0.5 & 0.1 \\ 0.4 & 0.5 & 0.1 \end{bmatrix} \quad \Delta P_i = \begin{bmatrix} 0 & 0 & 0 \\ -0.005 & 0.005 & 0 \\ -0.005 & 0.005 & 0 \end{bmatrix}$$

After 20 iterations the desired transition matrix is reached and the stabilizing controller is given as:

$$\mathbf{K} = \begin{bmatrix} 0.3181 & 0.7972 & 21.2058 & 5.4654 \end{bmatrix}$$

Note that the process can be started with any $P$ and $\mathbf{K}_{LQR}$ as long as they give feasible solution. Using Theorem 1 [22] the MADB using the LQR controller is 0.1210 s. With the stabilizing controller that takes the random time delay into consideration, Theorem 1 [22] gives 0.1420 s which shows an improvement in the stability margin with the new controller. The V-K iteration loop took 4-iterations and the perturbation loop took 20 iterations, the minimum decay rate is 0.8837. By changing the EVP loop by making an inner loop for minimizing α, the minimum attained decay rate is 0.8645 and the delay margin increased to 0.1563 s. The system response is shown in Figure 5. In the simulation the nonlinear dynamics is used. The stabilizing controller with the improved algorithm is:

$$\mathbf{K} = \begin{bmatrix} 0.2823 & 0.7050 & 20.5227 & 4.9714 \end{bmatrix}$$

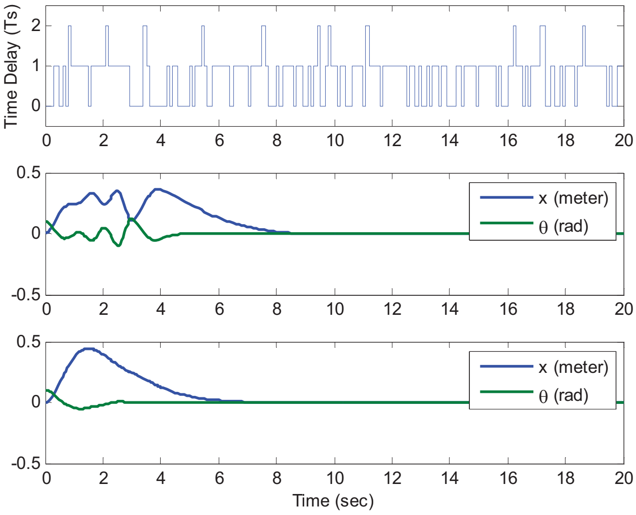


Figure 5. (a) The random time delay, (b) The response with the LQR controller, (c) The response with the controller generated by the improved V-K algorithm

## V. Conclusion

In this paper, the NCS is modeled as discrete-time Markovian linear jump system where the time delay is modeled as Markov chain. Using the mean square stability the system stability as formulated as BMIs. The V-K iteration algorithm is used to solve the BMIs. We used an improved V-K iteration algorithm where the decay rate is improved in both the EVP and the GEVP loops. The method is tested on the cart and the inverted pendulum and we found that the decay rate is improved.

## References

[1] Xiefu Jiang, Qing-Long Han, Shirong Liu, and Anke Xue. "A New *H∞* Stabilization Criterion for Networked Control Systems", *IEEE Transactions On Automatic Control*, Vol.53, No.4, 1025-1032. May 2008.

[2] Gregory C. Walsh, Hong Ye, and Linda G. Bushnell. “Stability analysis of networked control systems,” *IEEE Transactions on Control System Technology,* Vol. 10, No. 3, 438–446, May 2002.

[3] Magdi. S. Mahmoud, "Robust Control and Filtering For Time Delay systems", New York: Marcel Dekker, 2000.

[4] J. P. Hespanha, P. Naghshtabrizi, and Y. Xu, "A survey of recent results in networked control systems," Proceedings of the IEEE, vol. 95, no. 1, pp. 138-172, 2007.

[5] C. Mo-Yuen and T. Yodyium, "Network-based control systems: a tutorial," in 27th Annual Conference of the IEEE Industrial Electronics Society, vol.3 ed Piscataway, NJ, USA, 2001, pp. 1593-1602.

[6] L. Lei-Ming, T. Chao-Nan, and Z. Hai-Jun, "Analysis and design of networked control systems with long delays based on Markovian jump model," in Proceedings of 2005 International Conference on Machine Learning and Cybernetics, Vol. 2 ed Piscataway, NJ, USA: IEEE, 2005, pp. 953-959.

[7] L. Lei-Ming and T. Chao-Nan, "Stabilization design of networked control systems," in 2008 International Conference on Machine Learning and Cybernetics, vol.4 ed Piscataway, NJ, USA: IEEE, 2008, pp. 2137-2142.

[8] X. Lin, A. Hassibi, and J. P. How, "Control with random communication delays via a discrete-time jump system approach," in Proceedings of the 2000 American Control Conference. ACC, vol.3 ed Danvers, MA, USA: American Autom. Control Council, 2000, pp. 2199-2204.

[9] M. Yu, L. Wang, T. Chu, and G. Xie, "Modelling and control of networked systems via jump system approach," IET Control Theory &amp; Applications, vol. 2, no. 6, pp. 535-541, June2008.

[10] Z. Liqian, S. Yang, C. Tongwen, and H. Biao, "A new method for stabilization of networked control systems with random delays," in Proceedings of the 2005 American Control Conference, vol. 1 ed Piscataway, NJ, USA: IEEE, 2005, pp. 633-637.

[11] Bo Yu and Yang Shi, "State Feedback Stabilization of Networked Control Systems With Random Time Delays and Packet Dropout," Michigan, USA: 2008, pp. 639-645.

[12] S. Yang and Y. Bo, "Output Feedback Stabilization of Networked Control Systems With Random Delays Modeled by Markov Chains," Automatic Control, IEEE Transactions on, vol. 54, no. 7, pp. 1668-1674, July2009.

[13] S. Yang, Y. Bo, and H. Ji, "Mixed H2/H∞ control of networked control systems with random delays modeled by Markov chains," 2009, pp. 4038-4043.

[14] Johan Nilsson, "Real-time Control Systems with Delays." PhD Lund Institute of Technology , 1998.

[15] Z. Guofeng, C. Xiang, and C. Tongwen, "A model predictive control approach to networked systems," in Proceedings of the 46th IEEE Conference on Decision and Control Piscataway, NJ, USA: IEEE, 2008, pp. 3339-3344.

[16] W. Jing, Z. Liqian, and C. Tongwen, "Model predictive control for networked control systems," International Journal of Robust and Nonlinear Control, vol. 19, no. 9, pp. 1016-1035, June2009.

[17] J. Wu, L. Zhang, and T. Chen, "An MPC Approach to Networked Control Design," 2007, pp. 10-14.

[18] Oswaldo L.V.Costa, "Stability Results for Discrete-Time Linear Systems with Markovian Jumping Parameters," Journal of Mathematical Analysis and Applications, vol. 197, pp. 154-178, 1993.

[19] L. El Ghaoui and M. A. Rami, "Robust state-feedback stabilization of jump linear systems via LMIs," International Journal of Robust and Nonlinear Control, vol. 6, no. 9-10, pp. 1015-1022, Nov.1996.

[20] A. Hassibi, J. How, and S. Boyd, "A path-following method for solving BMI problems in control," in Proceedings of the 1999 American Control Conference, vol.2 ed Piscataway, NJ, USA: IEEE, 1999, pp. 1385-1389.

[21] David Banjerdpongchai, "Parametric Robust Controller Synthesis Using Linear Matrix Inequalities." University of Stanford, 1997.

[22] A. F. Khalil and W. Jihong, "A new stability and time-delay tolerance analysis approach for Networked Control Systems," 2010, pp. 4753-4758.

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# Developments in Dependability Modeling of Networked Control Systems

Manoj Kumar
Control Instrumentation Division
Bhabha Atomic Research Centre
Trombay, Mumbai 400085, India
Email: kmanoj@barc.gov.in

A.K. Verma
Stord/Haugesund University College
Bj$\phi$rsonsgate 45
Haugesund 5528, Norway
Email: akvmanas@gmail.com

A. Srividya
Department of Civil Engineering
Indian Institute of Technology Bombay
Mumbai 400076, India
Email: asvidya@civil.iitb.ac.in

P.P. Marathe
Control Instrumentation Division
Bhabha Atomic Research Centre
Trombay, Mumbai 400085, India
Email: cnidppm@barc.gov.in

*Abstract*—**Networked control systems (NCS) are in existence for quite some time. The network-induced delays and packet drops; and their effect on control system's performance has been a burning topic for last one decade. Dependability in context of NCS has not got much attention. While it is an important aspect of NCS intended to be used in critical application. The paper reviews the present state of developments in NCS and presents open problems pertaining to dependability of NCS. Solutions to these problems may increase the presence of NCS by many folds and in many fields.**

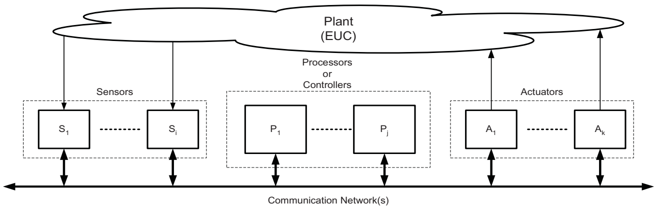


Fig. 1. Schmatic of a typical NCS

## I. Introduction

Networked Control System (NCS) contains a number of interconnected devices that exchange data through shared communication networks. Examples of such systems are found in industrial automation, building automation, office and home automation, intelligent vehicle systems and advanced aircraft and spacecraft. These systems are driven by an important feature, instead of point-to-point connections, all system elements (sensors, actuators and controllers) are connected to the network as nodes. The advantages of this implementation include [1], [2]: reduced system wiring, plug and play devices, increased system agility, ease of system diagnosis and maintenance. These features result in modular and flexible system design, simple and fast implementation and powerful system diagnosis and maintenance utilities [3].

Schematically a typical NCS is shown as in Figure 1. Sensor node samples the process parameters with a given sampling period, convert physical parameters to digital and pack the message to send to controller. The controller node unpacks the message(s) from sensor node and use control algorithm to calculate control signals to be sent to actuator node. Actuator node according to control signal takes the corrective action in process. All these messages are sent over shared network. In NCS, time delay has one more factor in addition to node processing delay, it is network-induced delay, i.e. from sensor to controller and controller to actuator.

Control systems are reactive systems; they need to interact with their environment (i.e. process/plant) constantly in a timely manner. Due to this property, control systems come under the category of real-time systems. They can be hard or soft real-time systems based on the consequences of failure. In NCS, the feedback loop is closed by shared communication network(s) where information is sent by means of packets. These shared communication media are prone to random delay and loss of packets. This leads to two challenging problems in analysis and design of networked control systems (NCS), network induced delays and packet dropouts. Both problems can significantly degrade the NCS's performance and dependability. It has long been realized that network induced communication delay is time-varying and nondeterministic, suggesting that the delay behavior is unpredictable. Packet dropout occurs when communication networks are unreliable or the communication latency is so big that the packet has to be purposely dropped.

Thess inherent issues of NCS - network-induced delay and packet loss - are well acknowledged [1], [2], [4], [5], [6], [7], [8], [9], [3], [10], [11], [12]. The delay and packet loss can degrade the quality of performance (QoP) of control system or even de-stabilize the system, if not properly designed [1], [2], [13], [14], [15], [7], [16], [6], [5], [17], [8], [18]. The two main directions to approach this problem are, i) design a communication protocol that guarantee delays, ii) design control strategies that *a priori* compensate for network-induced delay and packet loss.

When NCS are used for control, then their main objective is

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

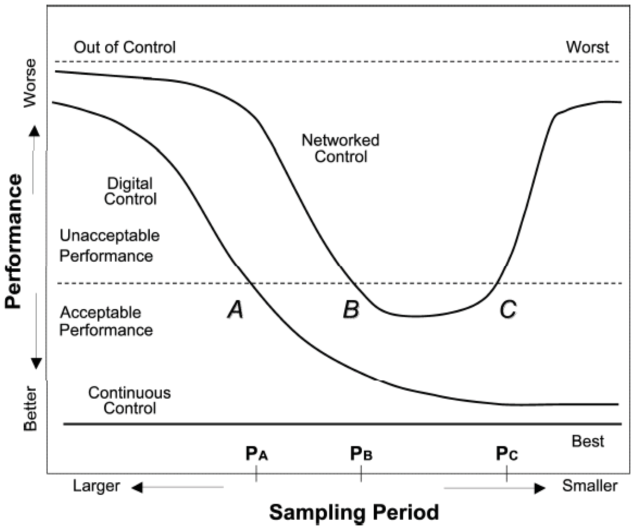


Fig. 2. Performance comparison [2], [9]

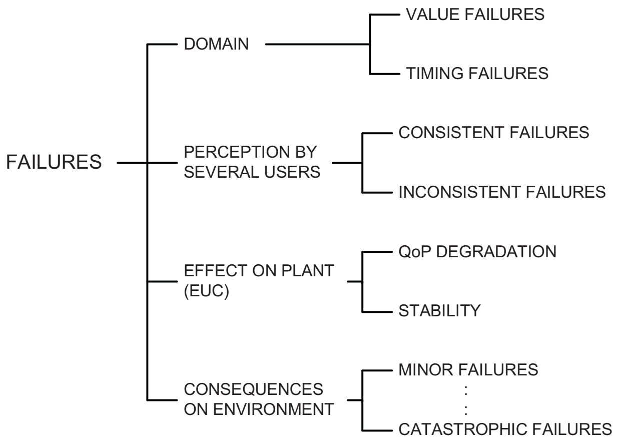


Fig. 3. Failure classification [19], [21]

to guarantee the stability and performance of plant/equipment under control (EUC), i.e. meet the control system specifications. These specifications include phase margin, gain margin, overshoot, steady state error, response-time, and tracking error etc.

The comparison of control performance versus sampling period for continuous control (analog), digital control, and network control is given in Fig. 2. During the NCS design stage, a performance chart can be derived as shown in Fig. 2. This performance chart provides a clear way to choose the proper sampling periods for an NCS. For a fixed control law, the worst, acceptable, and the best regions can be defined based on control specifications. The performance axis in Fig. 2 could be chosen to reflect a subset of the control system specifications. Since the performance of continuous control is not a function of sampling period, the performance index is constant for a fixed control law. For digital control case, the performance only depends on the sampling period assuming no other uncertainties. The performance degradation point A (sampling period $P_A$) in digital control could be estimated based on the relationship between control system bandwidth and sampling rate. For the networked control case, point B can be determined by further investigating the characteristics and statistics of network-induced delays and device processing time delays. As the sampling period gets smaller, the network traffic load becomes heavier, the possibility of more contention time or data loss increases, and longer time delays result. This situation causes the existence of point C in networked control.

With the advances in technologies related to NCS and advantages offered by them over conventional systems, NCS are penetrating into almost every aspect of our life. When these systems are used in critical applications, such as, nuclear power plant, avionics, process plants and automobiles etc., failure of these systems could result in loss of huge investment, effort, life or damage to environment. In such cases, dependability analysis becomes an important tool for decision making at all stages of system life cycle - design, deployment, operation and phase-out. In fact for systems concerning safety of people, demonstration of dependability through testing/analysis is a mandatory requirement before system can be deployed.

The paper reviews the related developments in the area of NCS and introduces dependability. Brief overview of dependability aspects are given in section II. Section III reviews the related developments. Dependability models are discussed in section IV. The open issues pertaining to NCS dependability modeling are outlined in section V, followed by conclusion in section VI.

## II. Dependability: Basic Fundamentals

Dependability of a system is defined by Algirdas Avizienis et al. [19], [20] as “ability to deliver service/function that can justifiably be trusted”. The service delivered by a system is its behavior as it is perceived by its user(s). User could be another system (physical, human) that interacts with the former. Service is delivered when the service implements the system function, where function is the behavior of the system described by its specification.

A system failure is an event that occurs when the delivered service deviates from correct service. A failure is a transition from correct service to incorrect service. Failure is manifestation of error, which in turn is caused by fault [19], [20], [21]. An error is that part of system state that may cause a subsequent failure: a failure occurs when an error reaches the service interface and alters the service. A fault is the adjudged or hypothesized cause of an error. A fault is active when it produces an error, otherwise it is dormant. Fig. 3, shows the modes characterizing incorrect service. Effect on plant (EUC) has been added considering NCS.

Fig. 4 shows the dependability tree consisting of attributes, means and threats.

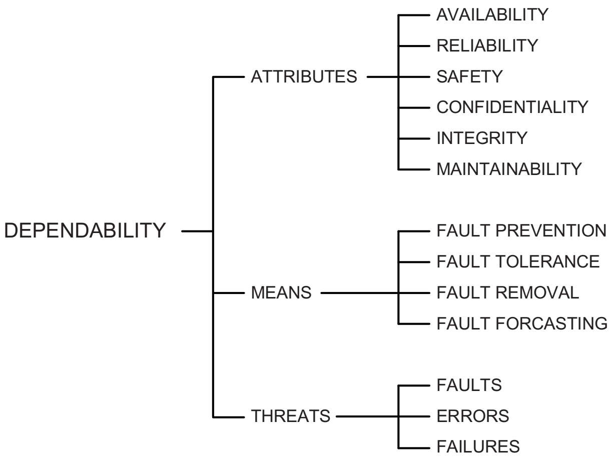


Fig. 4. Dependability tree [19], [21]

The means to attain dependability are a set of four techniques [19], [21]:

1) **Fault prevention:** to prevent the occurrence or introduction of faults
2) **Fault tolerance:** capability to deliver correct service in the presence of faults
3) **Fault removal:** to reduce the number and severity of faults
4) **Fault forecasting:** to estimate the present number, the future incidence, and the likely consequence of faults

The dependability attributes for electronic systems depend on their application or function. An electronic system when used in critical applications can be categorized as safety-critical, mission-critical and economically critical. The definitions of these are given below:

**Safety-critical systems:** systems required to ensure safety of equipment under control (EUC), people and environment.

**Mission-critical systems:** systems whose failure result in failure/loss of mission.

**Economically-critical systems:** systems whose failure result in availability of EUC, causing massive loss of revenue.

Dependability attributes applicable to these systems are safety, reliability and availability, respectively. Fig. 5 shows the applicable dependability attributes pictorially.

## III. Review of NCS related research

NCS related research developments can be categorized as, i) network related, ii) control related, and iii) dependability related. Some of the work is reviewed below.

Network related research is in development and performance evaluation of Fieldbus technology. Ethernet related technologies have also been an area of research because of the market share and cost benefit of Ethernet. For Ethernet new development includes traffic smoothing and switched Ethernet.

The real-time industrial network, often referred to as fieldbus, is an important element for building automated manufacturing systems. Thus, in order to satisfy the real-time requirements of field devices such as sensors, actuators and controllers, a number of fieldbus protocols have been developed. These fieldbus protocols have an important advantage over the widely used Ethernet in terms of their deterministic behavior. However, the application of fieldbuses has been limited because of the high cost of hardware and the difficulty in interfacing them with multi-vendor products. In order to solve these problems, computer network technology, especially Ethernet, is being adopted in industrial automation field. The key technical obstacle for Ethernet for industrial applications is that of its non-deterministic behavior [2], [3]. Non-deterministic behavior makes it inadequate for real-time applications, where packets containing real-time information (control command or sensor/actuator signal) have to be delivered within a certain time limit. To overcome the limitations of standard Ethernet, recent development has lead to switched Ethernet, EtherCAT, EPL and PROFInet [3], [22], [23], [24], [25], [26], [27], [28], [29], [30]. These have been adopted in industrial applications because of the elimination of uncertainties in network operation, which leads to improved performance.

Control system related developments can be classified in three categories: i) dealing with only delay, ii) dealing only with packet drops and iii) dealing with both [31], [32], [33].

For quite a long, it has been realized that conventional reliability models are not sufficient to analyze hard real-time systems, since they do not adequately model the temporal properties of such system [34]. A number of methods and techniques have been proposed since then to model and analyze temporal behavior and correctness [34], [2], [33], [30], [31], [26], [24], [12], [1], [29], [35], [36], [37]. A brief review of these is tabulated in Table 1.

## IV. Dependability models

To address dependability in context of NCS, available literature has the following:

- techniques to design robust controller for a specified bound on time-delay and packet drop rate.
- a number of networks have been characterized for their time delay and packet drop behavior
- a number of control networks have been proposed with guaranteed behavior
- techniques are being developed to characterize a given network to specify its time-delay and packet drop behavior for specified traffic pattern and network load

Conventional dependability models do not incorporate failures due to not meeting timing constraint and dynamic properties. From literature, it is evident that for NCS functional dependability attributes are more appropriate [37]. For programmable electronic systems functional safety has already been standardized and gaining importance [38].

With respect to dependability modeling of NCSs, there seems to be two different categories of systems - i) failure to meet time deadline leads to system failure and, ii) failure due to not keeping the performance objective.

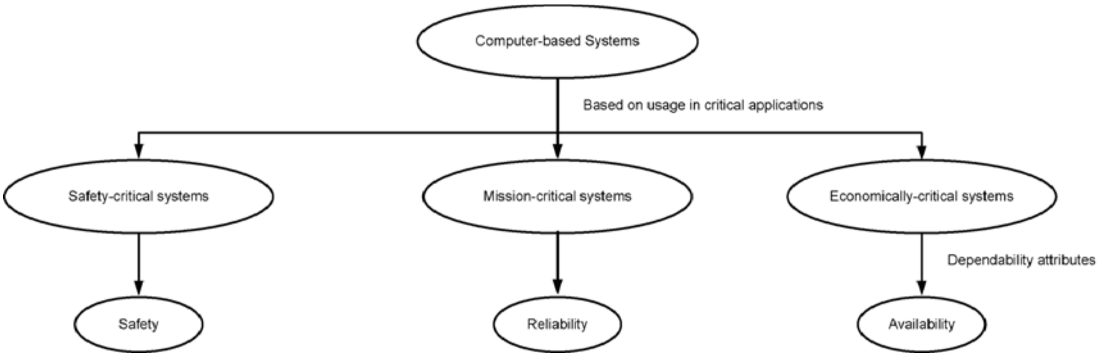
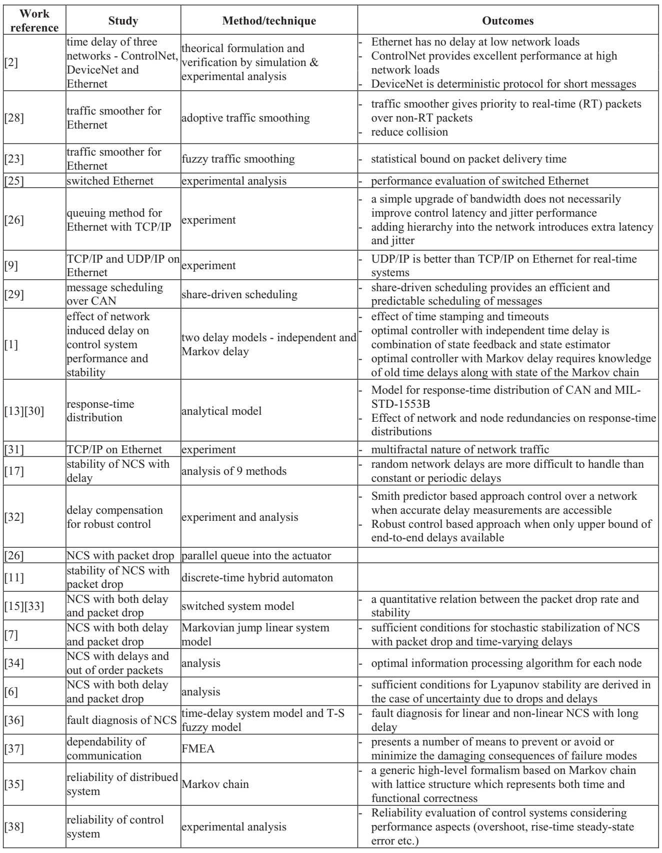


Fig. 5. Failure domains and dependability attributes

Table 1 Review of related literature



| Work reference | Study | Method/technique | Outcomes |
|---|---|---|---|
| [2] | time delay of three networks - ControlNet, DeviceNet and Ethernet | theorical formulation and verification by simulation & experimental analysis | - Ethernet has no delay at low network loads<br>- ControlNet provides excellent performance at high network loads<br>- DeviceNet is deterministic protocol for short messages |
| [28] | traffic smoother for Ethernet | adoptive traffic smoothing | - traffic smoother gives priority to real-time (RT) packets over non-RT packets<br>- reduce collision |
| [23] | traffic smoother for Ethernet | fuzzy traffic smoothing | - statistical bound on packet delivery time |
| [25] | switched Ethernet | experimental analysis | - performance evaluation of switched Ethernet |
| [26] | queuing method for Ethernet with TCP/IP | experiment | - a simple upgrade of bandwidth does not necessarily improve control latency and jitter performance<br>- adding hierarchy into the network introduces extra latency and jitter |
| [9] | TCP/IP and UDP/IP on Ethernet | experiment | - UDP/IP is better than TCP/IP on Ethernet for real-time systems |
| [29] | message scheduling over CAN | share-driven scheduling | - share-driven scheduling provides an efficient and predictable scheduling of messages |
| [1] | effect of network induced delay on control system performance and stability | two delay models - independent and Markov delay | - effect of time stamping and timeouts<br>- optimal controller with independent time delay is combination of state feedback and state estimator<br>- optimal controller with Markov delay requires knowledge of old time delays along with state of the Markov chain |
| [13][30] | response-time distribution | analytical model | - Model for response-time distribution of CAN and MIL-STD-1553B<br>- Effect of network and node redundancies on response-time distributions |
| [31] | TCP/IP on Ethernet | experiment | - multifractal nature of network traffic |
| [17] | stability of NCS with delay | analysis of 9 methods | - random network delays are more difficult to handle than constant or periodic delays |
| [32] | delay compensation for robust control | experiment and analysis | - Smith predictor based approach control over a network when accurate delay measurements are accessible<br>- Robust control based approach when only upper bound of end-to-end delays available |
| [26] | NCS with packet drop | parallel queue into the actuator | |
| [11] | stability of NCS with packet drop | discrete-time hybrid automaton | |
| [15][33] | NCS with both delay and packet drop | switched system model | - a quantitative relation between the packet drop rate and stability |
| [7] | NCS with both delay and packet drop | Markovian jump linear system model | - sufficient conditions for stochastic stabilization of NCS with packet drop and time-varying delays |
| [34] | NCS with delays and out of order packets | analysis | - optimal information processing algorithm for each node |
| [6] | NCS with both delay and packet drop | analysis | - sufficient conditions for Lyapunov stability are derived in the case of uncertainty due to drops and delays |
| [36] | fault diagnosis of NCS | time-delay system model and T-S fuzzy model | - fault diagnosis for linear and non-linear NCS with long delay |
| [37] | dependability of communication | FMEA | - presents a number of means to prevent or avoid or minimize the damaging consequences of failure modes |
| [35] | reliability of distribued system | Markov chain | - a generic high-level formalism based on Markov chain with lattice structure which represents both time and functional correctness |
| [38] | reliability of control system | experimental analysis | - Reliability evaluation of control systems considering performance aspects (overshoot, rise-time steady-state error etc.) |

The models of these failures which can utilize the available techniques of dependability analysis are the need of the hour. One important failure model is hazard rate. If hazard rate of timing failure and performance-related failure can be obtained, then using available techniques dependability attribute of interest can be estimated.

For systems belonging to first category, response-time distribution can be used to estimate probability of missing the deadline [39], [29], [12].

$$\begin{aligned}
&p = R\left(t > t_d\right)\\
&\text{where:}\\
&p: \text{ probability of timing failure}\\
&t_d: \text{ time deadline}\\
&R\left(\cdot\right): \text{response - time distribution (CDF)}
\end{aligned}\tag{1}$$

Let system failure criteria is $n$ consecutive deadline violation (or timeliness failures). When $n = 1$, number of cycles at which timeliness failure will occur follows geometric distribution [40].

$$\begin{aligned}
&P\left(Z = n\right) = p^{n-1}q\\
&\text{where}\\
&Z: \text{random variable}\\
&q: \text{ probability of occurrence of timeliness failure}\\
&p: 1 - q
\end{aligned}\tag{2}$$

Geometric distribution is a memoryless distribution in discrete time and is counterpart of exponential distribution in continuous time [40]. At gross level (larger time scale), it can be easily converted to exponential distribution. In exponential distribution characterizing parameter is hazard rate, which in this case is referred to as "*timeliness hazard rate*".

$$\begin{aligned}
&\lambda^T = \tfrac{1}{t}\ln\left(\frac{1}{P\left(Z > \left\lceil \frac{t}{t_C}\right\rceil\right)}\right)\\
&where\\
&\lambda^T: \text{Timeliness hazard rate}\\
&t: \text{Operating time}\\
&t_C: \text{Cycle time or period}
\end{aligned}\tag{3}$$

When $n > 1$, number of cycles for timeliness failure will not follow geometric distribution. This process (number of cycles for timeliness failure) is a memory process and directly cannot be modeled as Markov. Using the technique of additional states [41], Markov model can be used to model this process. This is expalined in [29].

For systems belonging to second category, techniques need to be evolved to derive hazard rates of failure due to performance degradation (overshoot, stability, etc.). The other open issues in dependability modeling of NCS are discussed in next section.

## V. Dependability issues in NCS

When systems are used in critical applications, all the means to achieve dependability are employed. Some of them are listed below:

*Fault Prevention:* Methods to avoid the occurrence of failure. For example,

1) selection and use of reliable hardware components,
2) use of fault-free algorithm and software,
3) use of communication network with predictable delay and preferably no drop during transmission

*Fault Tolerance:* Means to tolerate the failure. For example,

1) Redundancy for hardware components, including transmission media,
2) software fault tolerance, recovery and roll-back,
3) control algorithm to cope with random packet delay and drop

*Fault Removal:* Detection and restoration of failed component/unit.

1) diagnosis of hardware component failures,
2) diagnosis of software failures,
3) monitoring of traffic and detection of anomaly

Although some work has been done to improve NCS dependability, but there exist some research gap for dependability modeling and analysis. These are listed as follows:

1) Functional dependability considering *value* as well as *timing* failures
2) Methods to deal with redundancy at node level (mainly actuator node)
3) Methods to deal with redundancy at communication level (mainly controller node)
4) Effect of fault occurrence and removal on system performance
5) Effect of on line repair

The above means try to ensure the uninterrupted operation of system. Like any other critical system, NCS is also expected to have graceful degradation or go a defined state on failure, e.g. *fail-safe* or *fail-silent*. In addition to this, it shall give opportunity for alternate means of control. E.g. in case control system has failed, it shall allow the remote manual control.

Some other issues are:

- Timing of faults: faults affecting the network may have various effects on the control system, depending upon the state of the system, when the failure occurs. E.g. a lost message in transient state does not have the same effect as in steady state.
- Timing model: for safety systems, choice of timing model and indulgent protocol also plays an important role [42].

## VI. Conclusion

NCS have been attracting significant interest in the past few years and will continue to do so for the years to come. With the advent of cheap, small, and low-power processors with communication capabilities, it has become possible to endow sensors and actuators with processing power and the ability to communicate with remote controllers through shared networks. In view of these developments, it is expected that NCS will become the more predominant; replacing the current centralized digital control systems that rely on dedicated connections between system elements.

Along with all advantages that NCS offer, they have some inherent limitations. The main issue being the network induced delays, packet drop, out-of-sequence packet and multiple packets. The network quality of service (QoS) related issues affect the quality of performance (QoP) of control systems and sometime lead to in stability. A lot of work has been done in last decade to address these problems.

Dependability analysis is important when NCS is expected to be used for a critical application. To improve dependability, systems are designed with redundancy, maintenance features and diagnosis. These features make the dependability analysis a complex process. The research gap in this direction has been outlined. As can be seen that scope of dependability of NCS is not limited to 'control algorithms' embedded in the controller node or to the traffic behavior on network. It considers the complete system and analyzes all kind of probable failures.

## Acknowledgment

The authors would like to thank Shri Y.K. Taly for his continual support and guidance. We also thank Shri U. Mahapatra for his kind guidance and critical comments to improve the work. Continual support and encouragement from Shri G.P. Srivastava is highly appreciated.

## References

[1] J. Nilsson, "Real-time control systems with delays," Ph.D. dissertation, Lund Institute of Technology, Sweden, 1998.

[2] F.-L. Lian, "Analysis, design, modeling, and control of networked control systems," Ph.D. dissertation, University of Michigan, 2001.

[3] T. Yang, "Networked control systems: a brief survey," *IEE Proc.-Control Theory Applications*, vol. 153, no. 4, pp. 403–412, July 2006.

[4] J. Hespanha, P. Naghshtabrizi, and Y. Xu, "A survey of recent results in networked control systems," *Proceedings of the IEEE*, vol. 95, pp. 138–162, 2007.

[5] M. Garcia-Rivera and A. Barreiro, "Analysis of networked control systems with drops and variable delays," *Automatica*, vol. 43, no. 12, pp. 2054 – 2059, 2007.

[6] X. Ye, S. Liu, and P. Liu, "Brief paper modelling and stabilisation of networked control system with packet loss and time-varying delays," *Control Theory Applications, IET*, vol. 4, no. 6, pp. 1094 –1100, june 2010.

[7] Y.-C. Tian and D. Levy, "Compensation for control packet dropout in networked control systems," *Information Sciences*, vol. 178, no. 5, pp. 1263–1278, 2008.

[8] M. Das, R. Ghosh, B. Goswami, A. Gupta, A. Tiwari, R. Balasubrmanian, and A. Chandra, "Network control system applied to a large pressurized heavy water reactor," *Nuclear Science, IEEE Transactions on*, vol. 53, no. 5, pp. 2948 –2956, oct. 2006.

[9] F.-L. Lian, J. Moyne, and D. Tilbury, "Network design consideration for distributed control systems," *IEEE Transaction on Control System Technology*, vol. 10, pp. 297–307, 2002.

[10] J. van Schendel, M. Donkers, W. Heemels, and N. van de Wouw, "On dropout modelling for stability analysis of networked control systems," in *American Control Conference (ACC), 2010*, 30 2010-july 2 2010, pp. 555 –561.

[11] B. Lu, "Probabilistic design of networked control systems with uncertain time delay," in *ASME International Mechanical Engineering Congress and Exposition, Proceedings, 9 Part A*, 2008, pp. 355–362.

[12] M. Kumar, A. K. Verma, and A. Srividya, “Probabilistic modeling of network-induced delays in networked control systems,” *Int. J. Applied Mathematics and Computer Sciences*, vol. 5, no. 1, pp. 43–54, 2009.

[13] L. Dritsas and A. Tzes, “Robust stability bounds for networked controlled systems with unknown, bounded and varying delays,” *Control Theory Applications, IET*, vol. 3, no. 3, pp. 270 –280, march 2009.

[14] W.-A. Zhang and L. Yu, “Modelling and control of networked control systems with both network-induced delay and packet-dropout,” *Automatica*, vol. 44, no. 12, pp. 3206 – 3210, 2008.

[15] Y.-B. Zhao, G.-P. Liu, and D. Rees, “Modeling and stabilization of continuous-time packet-based networked control systems,” *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 39, no. 6, pp. 1646 –1652, dec. 2009.

[16] Y. Tipsuwan and M.-Y. Chow, “Control methodologies in networked control systems,” *Control Engineering Practice*, vol. 11, no. 10, pp. 1099 – 1111, 2003.

[17] F. Hao and X. Zhao, “Linear matrix inequality approach to static output-feedback stabilisation of discrete-time networked control systems,” *Control Theory Applications, IET*, vol. 4, no. 7, pp. 1211 –1221, july 2010.

[18] I. Pan, S. Das, and A. Gupta, “Tuning of an optimal fuzzy pid controller with stochastic algorithms for networked control systems with random time delay,” *ISA Transactions*, vol. 50, no. 1, pp. 28 – 36, 2011.

[19] A. Avizienis, J.-C. Laprie, and B. Randell, “Fundamental concepts of dependability,” in *Proc. of 3rd Information Survivability Workshop*, Oct. 2000, pp. 7–11.

[20] J. Dugan and K. Trivedi, “Coverage modeling for dependability analysis of fault-tolerant systems,” *IEEE Transactions on Computers*, vol. 38, no. 6, pp. 775–787, 1989.

[21] A. Avizienis, J.-C. Laprie, B. Randell, and C. Landwehr, “Basic concepts and taxonomy of dependable and secure computing,” *IEEE Transaction Dependable and Secure Computing*, vol. 1, no. 1, pp. 11–33, 2004.

[22] L. L. Bello, G. Kaczynski, and O. Mirabella, “Improving the real-time behavior of ethernet networks using traffic smoothing,” *IEEE Transaction on Industrial Informatics*, vol. 1, no. 3, pp. 151–161, 2005.

[23] P. Pedreiras, L. Almeida, and P. Gai, “The ftt-ethernet protocol: Merging flexibility,timeliness and efficiency,” *Real-Time Systems, Euromicro Conference on*, vol. 0, p. 152, 2002.

[24] K. C. Lee and S. Lee, “Performance evaluation of switched ethernet for real-time industrial communications,” *Computer Standards &amp; Interfaces*, vol. 24, no. 5, pp. 411 – 423, 2002.

[25] Y.-C. Tian, Q.-L. Han, C. Fidge, M. O. Tade, and T. Gu, “Communication architecture design for real-time networked control systems,” in *Proceedings of the IEEE Conference on Communications, Circuits and Systems*, 2006, pp. 1840–1845.

[26] F.-L. Lian, J. Moyne, and D. Tilbury, “Performance evaluation of control networks: Ethernet, controlnet, and devicenet,” *IEEE Control System Magazine*, vol. 21, pp. 66–83, 2001.

[27] S.-K. Kweon, K. G. Shin, and G. Workman, “Achieving real-time communication over ethernet with adaptive traffic smoothing,” *Real-Time and Embedded Technology and Applications Symposium, IEEE*, vol. 0, p. 90, 2000.

[28] T. Nolte, “Share-driven scheduling of embedded networks,” Ph.D. dissertation, Malardalen University, Sweden, 2006.

[29] M. Kumar, A. K. Verma, and A. Srividya, “Response-time modeling of controller area network (CAN),” in *Int. Conf. on Distributed Computing and Networking (ICDCN09), Lecture Notes in Computer Science, LNCS 5408*. Springer, 2009, pp. 163–174.

[30] Y.-C. Tian, Z.-G. Yu, and C. Fidge, “Multifractal nature of network induced time delay in networked control systems,” *Physics Letters A*, vol. 361, pp. 103 – 107, 2007.

[31] N. Vatanski, J.-P. Georges, C. Aubrun, E. Rondeau, and S.-L. Jms-Jounela, “Networked control with delay measurement and estimation,” *Control Engineering Practice*, vol. 17, no. 2, pp. 231 – 244, 2009.

[32] W.-A. Zhang and L. Yu, “A robust control approach to stabilization of networked control systems with time-varying delays,” *Automatica*, vol. 45, no. 10, pp. 2440 – 2445, 2009.

[33] V. Gupta, A. Dana, J. Hespanha, R. Murray, and B. Hassibi, “Data transmission over networks for estimation and control,” *Automatic Control, IEEE Transactions on*, vol. 54, no. 8, pp. 1807 –1819, aug. 2009.

[34] C. Perkins and A. Tyrrell, “A new markov model for dependability and temporal evaluation of hard real-time systems,” in *Proceedings of ESS95, The 7th European Simulation Symposium*, 1995, pp. 26–28.

[35] H. Fang, H. Ye, and M. Zhong, “Fault diagnosis of networked control systems,” *Annual Reviews in Control*, vol. 31, no. 1, pp. 55 – 68, 2007.

[36] L. Cauffriez, J. Ciccotelli, B. Conrard, and M. Bayart, “Design of intelligent distributed control systems: a dependability point of view,” *Reliability Engineering and System Safety*, vol. 84, no. 1, pp. 19–32, 2004.

[37] R. Ghostine, J.-M. Thiriet, and J.-F. Aubry, “Variable delays and message losses: Influence on the reliability of a control loop,” *Reliability Engineering and System Safety*, vol. 96, pp. 160–171, 2011.

[38] *IEC 61508: Functional Safety of electric/electronic/programmable electronic safety-related systems*, IEC Std. 0-8, 2000.

[39] K. S. Trivedi, S. Ramani, and R. Fricks, “Recent advances in modeling response-time distributions in real-time systems,” *Proceedings of the IEEE*, vol. 91, pp. 1023–1037, 2003.

[40] K. S. Trivedi, *Probability & Statistics with Reliability, Queueing, and Computer Science Applications*. Englewood Cliffs, New Jersey: Prentice-Hall, 1982.

[41] D. R. Cox and H. D. Miller, *The Theory of Stochastic Processes*. London: Methuen & Co., 1970.

[42] I. Keidar and A. Shraer, “How to choose a timing model?” in *Proc. $37^{th}$ Annual IEEE/IFIP Int. Conf. on Dependable Systems and Networks (DSN’07)*, 2007.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# A DNA Sequencing Based Approach to Fault Diagnosis in Automotive Electronic Networks

Roozbeh Bonyadi, R. Peter Jones
School of Engineering
The University of Warwick
Coventry, United Kingdom

James Taylor, Mark Amor-Segan
WMG
The University of Warwick
WMG is a member of the High Value Manufacturing Catapult

***Abstract*—The increasing number of Electronic Control Units within the network of a vehicle is increasing the level of complexity of these networks. Thus, fault diagnosis of these sophisticated systems becomes more complex. The aim of this paper is to provide a technique to enable CAN-based fault detection in a premium vehicle network or any other network which uses the Controller Area Network (CAN) protocol for communication. The fault detection technique described here is based on a sequential behaviour of the CAN network of a vehicle and by using signal processing methods commonly used in DNA sequencing analysis, fault detection was achieved and data were classified in clusters of normal scenarios and fault scenario.**

***Keywords-Fault Diagnosis;Electronic Control Unit; Controller Area Network Protocol; DNA Sequencing; Density Power Spectrum; Cross Correlation; Classification; Clustering***

## I. INTRODUCTION

Over the past two decades, numerous mechanical and hydraulic systems have been gradually replaced by electronics due to the development of embedded systems in automotive networks of vehicles [1]. These systems precisely assist driver to take control of the vehicle through the steering, engine control, suspension, braking, stability and traction functions [2]. Each function enabled by these electronic systems has an embedded electronic control unit (ECU) [3], or is distributed among a group of ECUs. The automotive networks of vehicle became more complex as the number of ECUs increased. ECUs need to exchange information among each other. Consequently, for managing data and granting read and write access to ECUs, protocols such as Controller Area Networks (CAN), Local Interconnect Network (LIN), Media Oriented Systems Transport Network (MOST) and more recently the future technology, FlexRay, were defined. One of the most widely used protocols in automotive systems is CAN protocol defined by a Robert Bosch GmbH in early 1990's [4].

In premium vehicles there are over 50 ECUs for both infotainment and control systems which are connected to each other. This increases the level of complexity of the network. Thus, the need of fault diagnosis within these networks becomes more important. To the best of our knowledge system wide diagnostic is not yet available. Fig.1 shows a typical premium vehicle network system architecture that supports both infotainment systems and control systems.

Here the fault diagnosis of the CAN network of a premium vehicle is studied. This vehicle has two different bandwidths: Low speed CAN which takes control of body functions and high speed CAN which takes control of the critical functions of the vehicle which needs real time control. The fault diagnosis within modern vehicle network systems so that faults within the complex network of the vehicle can be identified as belonging to a certain category is investigated here. The introduced method can also be applied in new areas, adopting the real-time fault detection and on-board fault diagnosis, utilising the data rich environment of the CAN network. To verify the system's behaviour fault injection technique is used. This is inserting of an artificial fault into the system and monitoring the response of the whole system [5].

The main purpose was to find a specification in different tests to find a way to distinguish fault scenarios from normal scenarios. After investigating different aspects of the CAN network, based on DNA sequencing approach, a method for distinguishing fault and normal scenarios is presented. Classification and data mining were used to classify scenarios.

It is essential to understand the level of complexity of CAN networks. In the next section different applications of this method is provided and data types and data gathering is discussed in section 3. In section 4, a DNA sequencing based method is proposed as a method to distinguish fault scenarios from normal scenarios. Section 5 provides the results of data analysis. The conclusion of the paper is provided in section 6.

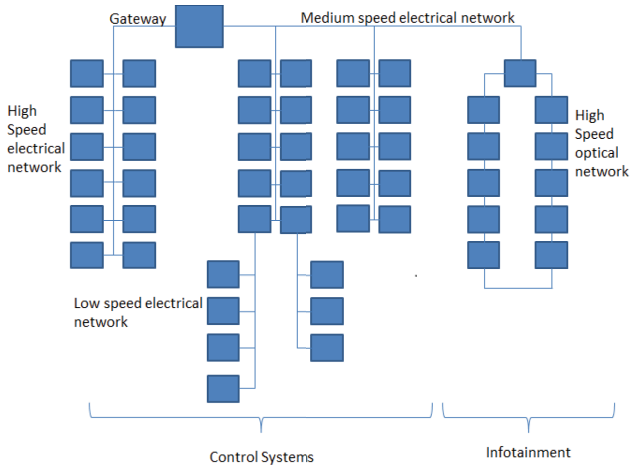

Gateway
Medium speed electrical network
High Speed electrical network
High Speed optical network
Low speed electrical network
Control Systems
Infotainment

Fig. 1 A typical premium vehicle network system architecture

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

## II. Application

The CAN protocol enables robust serial communication and was introduced by German automotive system supplier Robert Bosch. It was chosen for embedded systems networked applications in various markets such as medical equipment, test equipment, industrial automation and mobile machines. More recently, CAN networks have been popular in automation and control applications too [4, 8]. The techniques used in this paper were developed on the CAN network of a vehicle, but they can also be used in any other applications apart from vehicle industry in which CAN protocols are used.

## III. Data Collection

For the data collection, specific experiments were carried out on a representative vehicle network. The aim of these experiments was to collect data relating to illegal vehicle wake-up. The vehicle network is monitored and saved on computer, utilising a CANcase device to interface to the high speed and low speed networks via USB.

### *A. Message Types*

Two types of messages are published through the CAN network of car, periodic messages and non-periodic messages. Periodic messages are sent regularly through the network for long periods of time. Non-periodic messages are a classification of messages which have an external actuator and are the main focus of this paper. In this network, there are 14 message IDs in the high speed CAN and 47 message IDs in the low speed CAN which are involved in the wake-up process.

### *B. Scenarios and Tests*

The network of the vehicle enters sleep mode when there are no interactions between ECUs to reduce the battery consumption. The sleep mode is a very low-current standby mode with bus wake-up capability [6].

When a non-periodic message is sent through the network, ECUs start to communicate by sending information about their task. After completion of the task, ECUs start to notify that they are ready to enter sleep mode, and upon receiving a confirmation message of “go to sleep mode” ECUs gradually start entering sleep mode [7]. The period of time which network goes from sleep mode to wake-up mode and then again back to sleep mode is called a test. The external action causing data to flow, which classifies the type of the test, is called a scenario.

If the non-periodic message which produced a test is normal function defined by manufacturer, the scenario is considered as a normal scenario. A fault scenario is caused by artificially injecting a non-periodic message into the network.

### *C. Data Visualisation*

Each test is stored as a text file into the computer. As each message passing through the CAN is logged into the computer, it is stored in a new line. Fig.2 shows a small part of data collected from one normal scenario. The first column is time which shows the time that messages were logged in. The second column shows in which CAN the message has been transmitted. If it is 1 it shows that message belongs to the High speed CAN and if it is 2 the message belongs to the Low speed CAN. The third row is the identification number or ID of message in hexadecimal code. Each message has individual number which shows specific data. The rest of 8 columns show data within the message in hexadecimal code. Each message has individual ID which shows specific data. In the analysis stage, each ID is converted into decimal numbers. A normal scenarios occurs when a valid wake-up message is sent, such as pressing the lock button on the car key, and a fault scenario is produced by sending a message not normally involved in wake-up, such as message 1C8, into the data. As all normal scenarios belong to the body functions of the vehicle, analysis will be focused on the low speed CAN.

## IV. Diagnostic Approaches

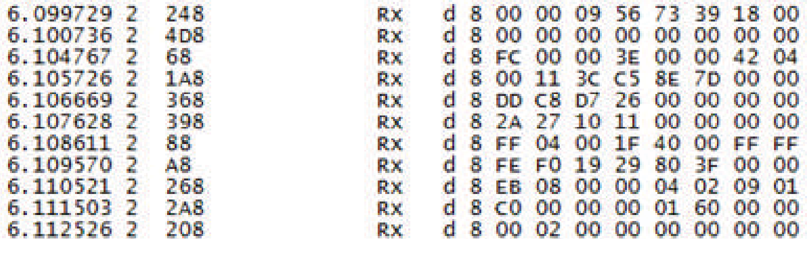
### *A. DNA Sequencing*

The digital genomic information are characterised in a form of sequence of finite numbers of entries coming after each other [9]. There are some similarities between the collected data from the real vehicle’s CAN network and DNA strands. As CAN messages enter the network through competitive arbitration, the sequence of messages becomes a feature of the network, similar to the sequence of bases in a DNA strand. In this section signal processing tools are applied on these sequences. Using classification methods described in section D fault diagnosis in the CAN network of the vehicle is achieved.

### *B. Binary Indicators*

47 different messages passing through the low speed CAN network were identified as significant. Thus 47 numbers are multiplied in its corresponding binary indicator sequences and then all of them are summed in order to build the sequence. Each sequence in the binary indicators ($u_{ID_i}$) is valued “1” if $ID_i$ exists in that sequence or a value of “0” if $ID_i$ is not present in that sequence:

$$x[n]=ID_1.u_{ID_1}[n]+ID_2.u_{ID_2}[n]+\ldots+ID_{47}.u_{ID_{47}}[n]$$
$$n=0, 1, 2, 3, \ldots, N-1 \qquad (1)$$



```
6.098786 2  1C8            Rx   d 8 80 00 02 08 18 02 20 11
6.099729 2  248            Rx   d 8 00 00 09 56 73 39 18 00
6.100736 2  4D8            Rx   d 8 00 00 00 00 00 00 00 00
6.104767 2  68             Rx   d 8 FC 00 00 3E 00 00 42 04
6.105726 2  1A8            Rx   d 8 00 11 3C C5 8E 7D 00 00
6.106669 2  368            Rx   d 8 DD C8 D7 26 00 00 00 00
6.107628 2  398            Rx   d 8 2A 27 10 11 00 00 00 00
6.108611 2  88             Rx   d 8 FF 04 00 1F 40 00 FF FF
6.109570 2  A8             Rx   d 8 FE F0 19 29 80 3F 00 00
6.110521 2  268            Rx   d 8 EB 08 00 00 04 02 09 01
6.111503 2  2A8            Rx   d 8 C0 00 00 00 01 60 00 00
6.112526 2  208            Rx   d 8 00 02 00 00 00 00 00 00
```

Fig. 2 An example of data collected from vehicle in text file format opened in Windows Notepad.

Frequency domain analysis using DFT is performed, using a sequence of length of N to provide the frequency content, X[k], at a frequency of k.

$$X[k]=ID_1.U_{ID_1}[k]+ID_2.U_{ID_2}[k]+\ldots+ID_{47}.U_{ID_{47}}[k]$$
$$k=0,\ 1,\ 2,\ 3,\ \ldots,\ N\text{-}1 \qquad (2)$$

In which $U_{ID_1}[k]$, $U_{ID_2}[k]$ … $U_{ID_{47}}[k]$ are the DFT of each of the binary indicator sequences x[n], producing a 47 dimensional representation of a frequency spectrum of a sequence of the messages passing through the CAN. The total spectral content, S[k], of a sequence of messages in the CAN at frequency of k is:

$$S[k]=\left|U_{ID_1}[k]\right|^2+\left|U_{ID_2}[k]\right|^2+\ldots+\left|U_{ID_{47}}[k]\right|^2 \qquad (3)$$

Three normal scenarios and two fault scenarios were established to perform fault diagnosis. Fig. 3 illustrates the Density Power Spectrum of the sequence of a test from one of the normal scenarios. The peaks of the power spectrum alter between scenarios, although all scenarios occur in a similar frequency range. In the case of DNA sequences, there are only four binary indicators and as a result there is only one peak that occurs at the frequency of N/3 where N is the length of sequence of x[n]. In the message sequences of the CAN network, there are 47 binary indicators. Hence, there is more than just one peak in the sequence of IDs.

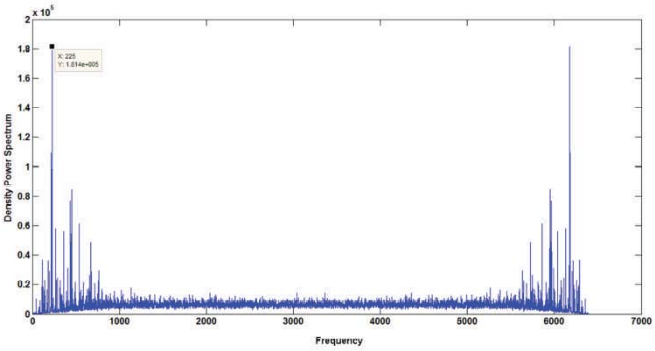


Fig. 3 Density power spectrum of a test from scenario of pressing open boot button on car key.

### C. Clustering Data

Clustering is assigning a set of data into subsets or groups so that data in the same clusters have similarity in some cases [10]. The frequency peaks in the DPS cluster the tests from these five scenarios into 3 different groups: normal scenario 1, normal scenario 2 and faulty scenarios.

### D. Classifier and Classification Method

Classification is a data mining process which aims to accurately assign target classes to data sets. The classifier which is used here is a hybrid system which effectively combines hand-built classifiers and empirical learning methods together [11]. This has the advantage of being able to utilise the characteristics of entire power spectrum in the classification.

One of simplest classification problems is binary classification between two states. This means that target group has only two possible states and classifier predicts state of the data set on a basis of whether they have the property or not. This type of classification is used in this paper.

```
$for\ i,j = 1:5$
$if\ |AXCTS(1,j) - AMVXC(i,j)| \le STD(i,j)$
        $\rightarrow scenario\ group\ of\ i$
$if\ \begin{cases} 1 \le i \le 2 & \rightarrow normal\ scenario\ group\ 1 \\ i = 3 & \rightarrow normal\ scenario\ group\ 2 \\ i > 3 & \rightarrow \qquad fault\ scenario \end{cases}$
```

Fig. 4 Algorithm of comparing a test set with training sets for classification.

#### 1) *Training Set*

Classification methods require training to be effective, which means a training data set is needed. This training set used to establish the relationship between predictors and targets. Different features can be defined in order to extract useful information from training set, and this information can be used to determine classes of future test sets.

Since the length of sequences in different scenarios was different, the test with the longest length, test l, was considered as a basis length of the sequences (length *l*). The length of test l according to the 100 randomly picked tests is 24592 sequences. For the rest of data sets in the test sets, if the length of the sequence of the messages in a test is less than *l,* then sequences of zeros are added to all of the 47 binary indicator sequences of $u_{ID_1}[k]$, $u_{ID_2}[k]$, …, $u_{ID_{47}}[k]$; and if the length of the sequence is greater than *l*, the first *l* sequences were considered, and the rest discarded. This is done because the values produced by this method are dependent on the length of the data sets. By using a standard base length, this effect will be normalised.

Next, the DFT of all the binary indicators are calculated and the density power spectrum of the whole sequence is achieved using equation 3. The density power spectrum of a discrete signal of the messages passing sequentially through the CAN network of the vehicle is a discrete signal. For classifying different scenarios, these discrete signals need to be compared to each other and form the basis of the classification scheme.

#### 2) *Comparison Tool*

The method chosen to compare these discrete signals is the cross correlation. It is a common method for estimating the correlation rate between two series. For a discrete signal, cross correlation is:

$$R_{xy}(d)=\begin{cases}\sum_{n=0}^{l\text{-}d\text{-}1} x[n+d]y^*[n] & d\geq 0 \\ R_{xy}^{\ *}(d) & d<0\end{cases} \qquad (4)$$

In which, *l* is the base length of the data sets and d is the signal delay. With the delay of d in x[n], the degree of correlation between x[n] and y[n] is calculated by dot-product of these two signals. The signals are considered to be periodic, so the cross correlation has twice of the length of its original signals. This operation was performed in MATLAB using the cross correlation command, which by default, computes the raw correlation between two signals without normalising it. The lengths of all data sets are converted to *l* so the normalised cross correlation which is the cross correlation divided by the length of the signals is achieved in advance (biased normalisation).

$$R_{xy,\ \mathrm{biased}}(\mathrm{d})=\frac{1}{l}R_{xy}(\mathrm{d}) \tag{5}$$

20 tests are chosen from each of the five scenarios, giving a total of 100 tests. The density power spectrum of all the 100 tests is calculated. Hence, there are 100 discrete time signals available. The cross correlation of each of the two of these power spectrums are calculated. The result of these calculations are stored in a massive cell with a dimension of 100×100 and in each of these cells, the cross correlation results with the length of 2×*l* is stored.

The peak of these cross correlations shows the maximum correlation of the two tests. The cross correlation of the density power spectrum of each test set and each test in the training sets are calculated and stored in a massive cell. For the calculated cell of the cross correlation with the dimension of 100×100 maximum values of the cross correlation is stored in a matrix with the same dimensionality to create the training sets.

Next, the matrix is divided into 5 matrices with dimension of 20×20 which are the maximum values of cross correlation between the scenario 1 and all the 5 scenarios (including auto-correlation with scenario 1). This is the procedure of making the training set for the first scenario. This should be completed for the second scenario, the third scenario and so on.

The maximum value of the cross correlation, shows rate of correlation between two scenarios. Average of maximum values of the cross correlation (AMVXC) of the 20 tests of each matrix are calculated and saved for each scenarios (Table 1). Also the Standard Deviation (STD) of these maximum values of the cross correlation of these 20 tests are calculated and stored for each scenario (Table 2).

The average value and the standard deviation of each matrix create the training sets for the classification. Fig. 5 shows the steps and the procedure of creating the training sets for the classification of first scenario. So the result of all these calculation will be a 5×5 matrix of average of the maximum values of the cross correlation between the scenarios and another 5×5 matrix is the standard deviation of the maximum values of the cross correlation between the scenarios.

### 3) *Classifying a Test Set*

For each given test set the cross correlation between the density power spectrum of that test set and the density power spectrum of the 100 tests used to build the training set is calculated. This results in 100 cross correlations.

The maximum value of the cross correlation between the test set and the 20 tests of the scenario 1 will be stored in a separate matrix; the maximum value of the cross correlation between the test set and the 20 tests of the second scenario will also be stored in another matrix and so on. Eventually, there will be 5 matrices with dimension of 1×20. The average of these values in each matrix will be calculated and stored in a matrix called Average of XCorrelation of Test Set or AXCTS (i ,j) (Table 3).

The AXCTS is then subtracted from each of the scenario's AMVXC (Table 4). Comparing each row with its counterpart in Table 2, it can be seen that this example test belongs to scenario 2 (normal scenario group 1) as the values in this row are less than those in Table 2. This algorithm is illustrated in Fig.4.

## V. RESULTS

Six test sets from each of the 5 scenarios (total of 30 tests) are randomly picked and considered as the test sets. The classification procedure is carried out on each test set. The confusion matrix in Table 5 shows 100% accuracy for assigning scenarios to the 3 classification groups. Fig.6 shows these 30 test sets in AXCTS matrix classified in the three clusters. In this plot each row of the AXCTS matrix is shown with 5 markers representing AXCTS value for its corresponding scenario. As it can be seen for all tests, the clusters are distinguished.

TABLE I. THE AVERAGE OF THE MAXIMUM VALUES OF THE CROSS CORRELATION FOR THE TRAINING SET

| AMVXC | Scenario1 | Scenario2 | Scenario3 | Scenario4 | Scenario5 |
|---|---|---|---|---|---|
| **Scenario 1** | 68.5296E+11 | 64.8896E+11 | 90.7273E+11 | 32.8471E+11 | 34.0116E+11 |
| **Scenario 2** | 64.8896E+11 | 61.6390E+11 | 85.9672E+11 | 31.5055E+11 | 32.6112E+11 |
| **Scenario 3** | 90.7273E+11 | 85.9672E+11 | 623.5030E+11 | 45.9586E+11 | 47.3855E+11 |
| **Scenario 4** | 32.8471E+11 | 31.5055E+11 | 45.9586E+11 | 18.4082E+11 | 18.9465E+11 |
| **Scenario 5** | 34.0116E+11 | 47.3855E+11 | 18.9465E+11 | 19.5366E+11 | 24.0919E+11 |

TABLE II. THE STD OF THE MAXIMUM VALUES OF THE CROSS CORRELATION FOR THE TRAINING SET

| STD | Scenario1 | Scenario2 | Scenario3 | Scenario4 | Scenario5 |
|---|---|---|---|---|---|
| **Scenario 1** | 2.8279E+11 | 3.5061E+11 | 5.5539E+11 | 2.6415E+11 | 2.4432E+11 |
| **Scenario 2** | 3.5061E+11 | 3.9585E+11 | 6.5875E+11 | 2.6612E+11 | 2.5055E+11 |
| **Scenario 3** | 5.5539E+11 | 6.5875E+11 | 32.5400E+11 | 3.1451E+11 | 3.1038E+11 |
| **Scenario 4** | 2.6415E+11 | 2.6612E+11 | 3.1451E+11 | 1.6417E+11 | 1.5966E+11 |
| **Scenario 5** | 2.4432E+11 | 2.5055E+11 | 3.1038E+11 | 1.5966E+11 | 1.5493E+11 |

TABLE III. THE AVERAGE VALUE OF THE MAXIMUM VALUES OF THE CROSS CORRELATION BETWEEN GIVEN TEST SET AND THE TRAINING SETS

| AXCTS | Scenario 1 | Scenario 2 | Scenario 3 | Scenario 4 | Scenario 5 |
|---|---|---|---|---|---|
| Test Set | 65.4680E+11 | 62.0810E+11 | 85.616E+11 | 31.673E+11 | 32.792E+11 |

TABLE IV. DIFFERENCE BETWEEN THE AMVXC OF THE TRAINING SETS AND AXCTS OF THE TEST SET

| \|AXCTS – AMVXC\| | Scenario 1 | Scenario 2 | Scenario 3 | Scenario 4 | Scenario 5 |
|---|---|---|---|---|---|
| Scenario 1 | 3.06E+11 | 2.81E+11 | 5.11E+11 | 1.17E+11 | 1.22E+11 |
| Scenario 2 | 0.58E+11 | 0.44E+11 | 0.35E+11 | 0.17E+11 | 0.18E+11 |
| Scenario 3 | 25.30E+11 | 23.90E+11 | 538.00E+11 | 14.30E+11 | 14.60E+11 |
| Scenario 4 | 32.60E+11 | 30.60E+11 | 39.70E+11 | 13.30E+11 | 13.80E+11 |
| Scenario 5 | 99.50E+11 | 109.00E+11 | 105.00E+11 | 51.20E+11 | 56.90E+11 |

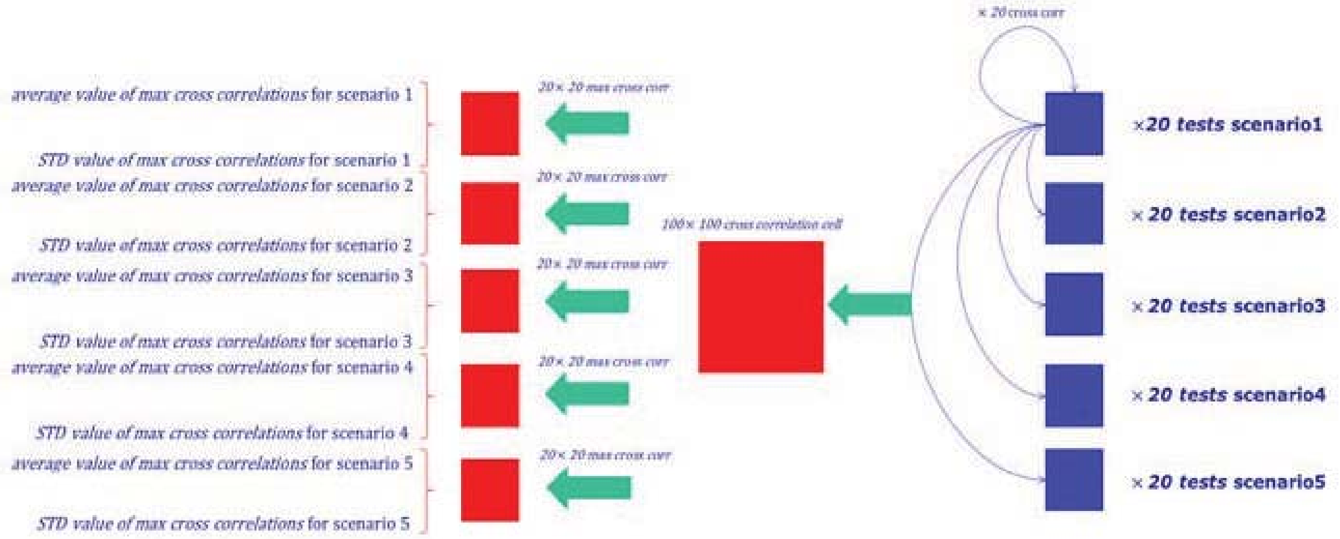

average value of max cross correlations for scenario 1
STD value of max cross correlations for scenario 1
average value of max cross correlations for scenario 2
STD value of max cross correlations for scenario 2
average value of max cross correlations for scenario 3
STD value of max cross correlations for scenario 3
average value of max cross correlations for scenario 4
STD value of max cross correlations for scenario 4
average value of max cross correlations for scenario 5
STD value of max cross correlations for scenario 5
20× 20 max cross corr
20× 20 max cross corr
20× 20 max cross corr
20× 20 max cross corr
20× 20 max cross corr
100× 100 cross correlation cell
× 20 cross corr
×20 tests scenario1
× 20 tests scenario2
× 20 tests scenario3
× 20 tests scenario4
× 20 tests scenario5

Fig. 5 Procedure of creating training set to be used for DPS method for classification for first scenario

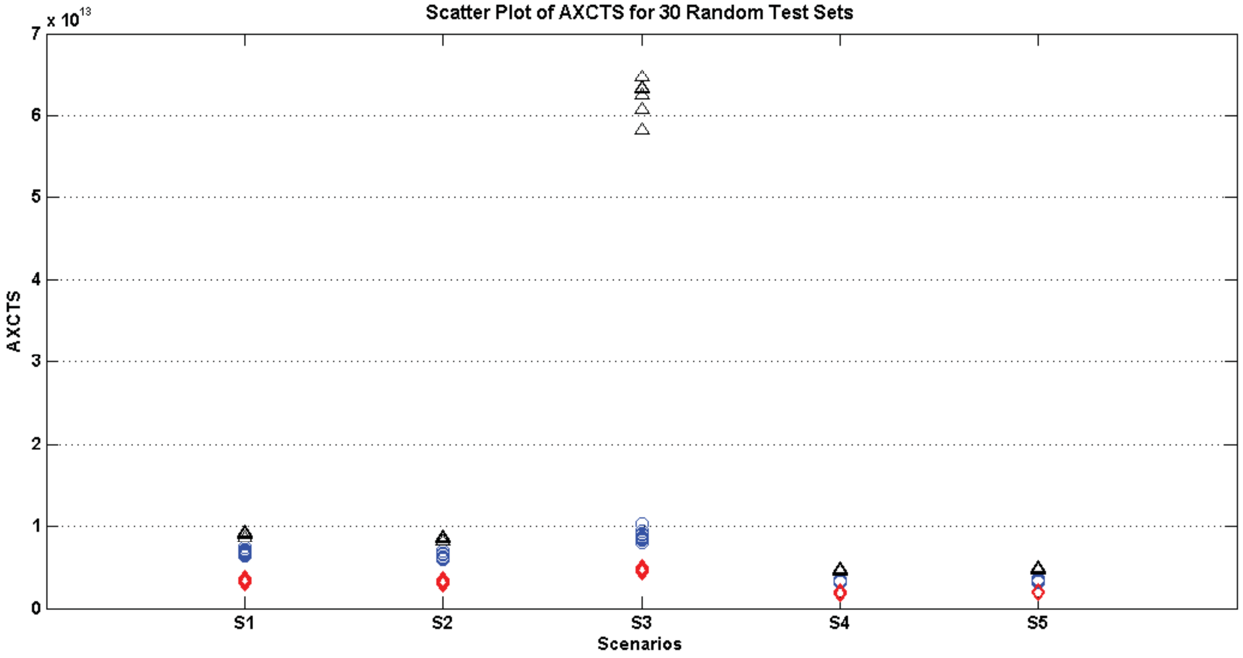


Fig. 6 Results, showing the matrix of AXCTS classified the 30 test sets in 3 clusters; Diamond: Fault scenario, Circle: Normal Scenario 1, Triangle: Normal Scenario 2

TABLE V. CONFUSION MATRIX FOR CLASSIFYING TEST DATA

| | Scn.1 | Scn.2 | Scn.3 | Scn.4 | Scn.5 |
|---|---|---|---|---|---|
| **Group 1** | 6 | 6 | 0 | 0 | 0 |
| **Group 2** | 0 | 0 | 6 | 0 | 0 |
| **Faults** | 0 | 0 | 0 | 6 | 6 |

## VI. CONCLUSION

In this paper, development and diversity of networks in vehicles was studied. The main focus was the Controller Area Network (CAN) protocol as a communication tool. Moreover, it was discussed that fault diagnosis in electronic systems and networks of a vehicle is becoming an increasingly important factor.

For the fault detection, a data mining technique was applied and, from comparison to DNA, the sequencing nature of the data was considered as a feature that could be used as a classifier. The algorithm used here as a diagnosis tool was developed and coded using MathWorks MATLAB R2009a.

Signal processing methods where used to derive the density power spectrum of the binary indicator sequences from the messages sequences in each test. It can be concluded that the place of occurrence of the peaks in the density power spectrum are different among the scenarios and it can be used as the classification feature. According to this feature, three clusters were identified: the normal scenario 1 cluster, the normal scenario 2 cluster and the fault scenario cluster.

Furthermore, instead of just considering the peaks the whole sequence of the power spectrum density was considered. The density power spectrum is a discrete signal in frequency domain and as comparison tool for the classification. The available data was split into training sets and the test sets. The cross correlation of the training sets and the test sets was utilised as a classification feature. This hybrid classifier was able to distinguish the fault scenarios from the normal scenarios in 100% of cases, showing that this method is an effective classifier for these data sets.

This technique was developed on the CAN network of a vehicle, but application of it is not limited to vehicle industry. It can also be used on other backgrounds which use the CAN protocol or even on other networks in a vehicle such as LIN and MOST. Also new methods which adopt real time fault detection and on-board fault diagnosis and use the large amount of information available from the system network to pinpoint the cause of detected faults can use the technique introduced here. The real time fault detection system requires a generic electronic control unit (generic ECU) to monitor data at fast speeds in order to find faults. This requires algorithms with a small numbers of variables in order to respond quickly to faults. This extracted feature may be a useful variable for this purpose.

This method is most effective when a sequence of messages is expected in response to an event (such as network wake-up and shutdown), with the advantage of not needing to understand the underlying network functions. However, other types of faults may not be as clearly defined as this. Network fault detection and diagnosis is a complex issue with no single answer, requiring multiple approaches to categorise faults. This method can provide an additional tool to further research in this area.

## REFERENCES

[1] Navet, N., Song, Y., Simonot-Lion, F., Wilwert, C., "Trends in Automotive Communication Systems," Proceedings of the IEEE, vol.93, no.6, June 2005, pp.1204-1223.

[2] Suwatthikul J., and McMurran, R., Peter Jones, R., "Automotive Network Diagnostic Systems," Industrial embedded systems, IES '06 international symposium, IEEE, 2006, pp.1-4, 1-4244-0777-X.

[3] Corno, F., Tosato, S., Gabrielli, P., "System-level analysis of fault effects in an automotive environment," *Defect and Fault Tolerance in VLSI Systems, 2003. Proceedings, 18th IEEE International Symposium on*, vol.3, no.5, Nov. 2003, pp.529-536.

[4] Johansson, K., Törngren, M., and Nielsen, L., "Handbook of Networked and Embedded Control Systems," D. Hristu-Varsakelis and W. S. Levine, Eds. Boston, MA: Birkhäuser, 2005.

[5] Arlat, J., Costes, A., Crouzet, Y., Laprie, J., and Powell, D., "Fault injection and dependability evaluation of fault-tolerant systems," IEEE Transactions on Computers, vol. 42, no.8, 1993, pp.913-923.

[6] NXP Semiconductors. "TJA1042 High-speed CAN transceiver with Standby mode". [Online]. (URL: http://www.nxp.com/documents/data_sheet/TJA1042.pdf) 2011. (Accessed 4th Aug 2011).

[7] stillerb infineon Company. "TLE6251-2G High Speed CAN-Transceiver with Wake and Failure Detection," [Online]. (URL: http://www.infineon.com/dgdl/TLE6251-2G_DS_rev10.pdf?folderId=db3a3043163797a6011666d32a0c0de1&fileId=db3a304320d39d5901215451b99c05dc) 2009. (Accessed 4th Au 2011)

[8] Simonot-Lion, F., "In-car embedded electronic architectures: how to ensure their safety," presented at the 5th IFAC Int. Conf. Fieldbus Systems and Their Applications, Aveiro, Portugal, 2003.

[9] Anastassiou, D., "Genomic signal processing," *Signal Processing Magazine, IEEE* , vol.18, no.4, Jul 2001, pp.8-20.

[10] MacQueen, J. B. "Some Methods for classification and Analysis of Multivariate Observations", Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, vol.1, 1967, pp.281-297.

[11] Towell, G. G., and Shavlik, J. W,. "Knowledge-based artificial neural networks, Artificial Intelligence," vol.70, nos. 1-2, October 1994, pp.119-165. ISSN 0004-3702, DOI: 10.1016/0004-3702(94)90105-8.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# VSW-SOAP: A SOAP Message Transmission Mechanism Based On Variable Sliding Window

Xiaoxuan Ma, Zhixin Chen
School of Electricity and Information Engineering
Beijing University of Civil Engineering and Architecture
Beijing, China
mxxuan@sohu.com
chenzhixin@bucea.edu.cn

***Abstract*—As an extensible lightweight message processing framework, SOAP has become the lightweight communication protocol for large scale data exchange in the distributed environment. However, one of the most important problems of SOAP is its relatively poor performance which leads to its limit in high performance applications. The paper proposes a SOAP message transmission mechanism based on variable sliding window, which can improve the performance of SOAP and support transferring data of large, flexible or unknown size via SOAP. In addition, data segmentation is designed to transfer large data and variable-size data and variable sliding window is used to control and improve the transmission efficiency of SOAP messages. The algorithm of application level congestion control was proposed and the transmission reliability of SOAP message was discussed with respect to four exception cases. Finally, the two main impact factors for the transmission efficiency based on variable sliding window such as the size of the sliding window and the length of data slice are analyzed and tested. Simulation results show that transmission efficiency is effectively improved by a reasonable parameters set of the variable sliding window.**

***Keywords-SOAP;sliding window;congestion control; transmission reliability***

## I. INTRODUCTION

Data exchange is the common, fundamental and critical requirement of distributed application in network environment. Data exchange is mainly used to solve the interoperability of different data resources in the infomationization, i.e. to solve the integration and sharing of different data resources between different heterogeneous systems. The essence of data exchange is the sharing and exchange of data. The data exchange platform is completely based on Web service technology using SOAP [1] as its communication protocol. SOAP is an extensive lightweight message processing framework based on XML [2], and has become a de facto standard for data exchange in the open network environment.

With the development of distributed computing technology, more different systems and networks at different speed are connected to the Internet, which results in more serious network congestion. Therefore, one of the most important part of Internet QoS[3] is how to deploy the peer communications at different speeds in different network to match capacity of the network. Large scale communications require effective transmission efficiency. Though SOAP provides a powerful interoperability, its own characteristics cause its lower transmission efficiency, e.g. when packaging a SOAP message, the transcoding and parsing will consume a large amount of time; generally, after data are package into a SOAP message, the data package will become more times than the original size; In addition, the time consumption of network transmission will result in the transmission of SOAP messages to take a long time. To the applications which require the higher efficiency of data transmission or large scale data transmission, SOAP can’t be effective to meet practical requirements.

According to the lower transmission efficiency of SOAP message, there are many improvements to improve the transmission efficiency as follows: to improve the transmission efficiency through improved SOAP binding [4]; according to different applications, using the appropriate SOAP message encoding rules [5]; to reduce construction and parsing of SOAP messages [6]; using compression technology to compress the SOAP message [7].

Sliding window technology is mainly used for information flow control, to coordinate the work pace of the sender and the receiver, to avoid the loss of data because the sender sends data too fast to the receiver too late to deal with [8]. This technology is not only widely used in the communication between Internet and the other network, but also used in information exchange of the different layers within the network system. In theory, sliding window can be used in anywhere which requires the flow control.

The paper proposes the VSW-SOAP(Variable Sliding Window-SOAP) to effectively improve the efficiency of data transmission. VSW-SOAP uses variable sliding window mechanism in SOAP message transmission and takes data slice as the SOAP message attachments. The basic data unit in the variable sliding window is the SOAP message with slice attachment.

The paper is organized as follows: Section II analyzes the principle of VSW-SOAP, defines the variable sliding window, discusses SOAP attachment segmentation, the structure of SOAP message and the transmission workflow of SOAP messages; Section III proposes an application level congestion control algorithm based on variable sliding window; Section IV

This work was funded by Beijing municipal outstanding talents Fund(Grant No. 2011D005017000002) and Dr. Start Fund of BUCEA (Grant No.101101815).

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

discusses the reliable transmission mechanism for SOAP message; Section V analyzes the main factors of VSW-SOAP and gives the simulation results; the conclusion and outlook for further work is given in Section VI.

## II. PRINCIPLES OF VSR-SOAP

### A. Variable sliding window

For data transmission, in extreme cases, unbounded input stream requires unbounded memory, which is clearly not applicable. Therefore, a window strategy is used to limit the number of tuples for each input stream stored in the processing of unbounded flow. The sliding window is set an interval on the data stream which only includes part of the new data of the data stream. With the arrival of new data, the window proceeds to replace the old data with new data. Therefore, the sliding window can be seen as a historical snapshot of a limited part of the data stream, while taking advantage of the characteristics of the sliding window to meet the real time and unbounded requirements of data flow.

As a historical snapshot of a limited part of the data stream, sliding window can be divided into time-based sliding window and tuple-based sliding window. The two sliding window usually assume that the window size is unchanged. In previous data transmission algorithms, many algorithms are based on the unchanged sliding window mechanism which achieves considerable results in terms of time and space complexity. However, these technologies have usually ignored the time-varying characteristics of specific issues in data stream. Therefore it is difficult to implement adaptive adjustment of data model when the data distribution characteristics are changeable [9, 10].

Variable sliding window can adaptively adjust the window size according to the change of data stream flow and data distribution in order to achieve a minimum consumption of memory space and processing time. The principle of variable sliding window is as follows: the sender maintains a continuous data set to be sent, called the sending window; at the same time, the receiver also maintains a continuous data set to be received, called the receiving window. The window is used for store data during message transmission. When updating data, the window will slide. In accordance with the principle of the sliding window, the sender has a sending window and the receiver has a receiving window. The sending window is used to implement flow control. The size of sending window represents the maximum number of data sent by the sender when the acknowledgment message is not received. The receiving window is used to control whether data can be received. At the receiving node, only when the serial number of the data is in the receiving window, the data can be accepted. The window size of variable sliding window is determined according to the data flow change. When the flow rate is quickly, reduce the size of the sliding window; otherwise, when the flow rate is very slow, increase the size of the sliding window in real time.

The differences between the SOAP sliding window and the other sliding window lies in: the data stored in SOAP sliding window are the SOAP messages, the function of the sliding window is to control the speed of the SOAP messages, to implement the network congestion control and to ensure a reliable orderly transmission of SOAP message attachment data. After the transmission is finished, the sliding window of the receiver will reassemble the slice data in the received SOAP message attachments into the original data.

### B. SOAP attachment segmentation

If the data of the SOAP message with attachments is large, it is difficult to read the whole data into memory only once. Even if the data can read only once, it will be very difficult and time consuming for large data encryption and decryption operations. Especially when the network quality is poor, if there are transmission data errors, retransmitting the entire data is costly. Thus, for the large scale data attachment in the SOAP message, data can be segmented into multiple slices. The data slices can be encapsulated as a new SOAP message attachment using MIME, DIME or XOP. Each time the sender can send a SOAP message attachment containing one or more data slices.

The process of data segmentation and encapsulation as a SOAP attachment is shown in Figure 1.

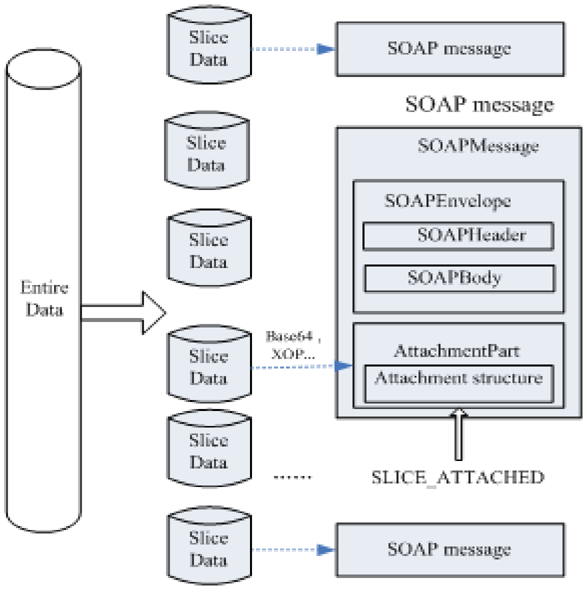


Figure 1. The process of data segmentation and encapsulation as a SOAP attachment

Because the data encapsulated in the SOAP message attachment is not complete data, but one or more data slices, the body of the SOAP message is required to at least include the following information: SOAP message number(SOAPID), the data number to be sent(DataNumber), the slice number (SliceNumber), whether it's the last data (isLastData), whether it is the last slice (isLastSlice). SOAPID is used to uniquely identify the SOAP message number that the data slice belongs to. DataNumber is used to represent the data that the data slice belongs to. SliceNumber is data slice number which represents the current transmission data slice. isLastData represents whether it's the last data. isLastSlice represents whether it's the last data slice. In this way, if there's requirement to transmit large scale attachments data, SOAP message data can be broken down into multiple small SOAP messages with a slice attachment.

By SOAP attachment segmentation, it avoids to read the entire data into memory at one time which is useful to encrypt and decrypt. Besides, the approach also supports flexible or unknown size SOAP attachment.

### *C. The structure of the SOAP message*

In the variable sliding window, the SOAP message is divided into data message and acknowledgment message.

The data message is sent by the sender and each data message contains one or more data slices. The length of slice is determined through consultation between the sender and the receiver. The slices are encapsulated in the SOAP message in accordance with the agreed specification. The basic information of data message is indicated in the body of SOAP message, which mainly includes: SOAPID, DataNumber, SliceNumber, IsLastData, IsLastSlice, the length of each data slice(size), safety information(Encryption). The structure of data message is shown in Figure 2.

```
 <TransportRequest>
      <SOAPID/>
      <DataNumber/>
      <SliceNumber/>
      <IsLastSlice>yes/no</IsLastSlice>
      <IsLastData>yes/no</IsLastData>
      <Encryption/>
       <Data> </Data>
</TransportRequest>
```

Figure 2. The structure of data message

The acknowledgment message is sent by the receiver to confirm the data message issued by the sender. The acknowledgment message of the SOAP message body contains the following information: SOAPID, the type of acknowledgment message(ACKType), the data number confirmed(ACKDataNumber), the slice number confirmed(ACKSliceNumber) , IsLastData, IsLastSlice, Specific response information(ResponseInfo).The structure of acknowledgment message is shown in Figure 3.

```
<TransportResponse>
        <SOAPID/>
        <ACKType/>
        <ACKDataNumber/>
        <ACKSliceNumber/>
        <IsLastData>yes/no</IsLastData>
        <IsLastSlice>yes/no</IsLastSlice>
        <ResponseInfo></ResponseInfo>
</TransportResponse>
```

Figure 3. The structure of acknowledgment message

### *D. The workflow of VSW-SOAP*

Based on variable sliding window, the workflow of the SOAP message transmission mechanism can be described as follows:

1) The connection between the sender and the receiver can be established through the three-way handshake. The three-way handshake can be implemented based on predefined transmission protocols and ports. In the three-way handshake, the following information are required to provide: SOAP attachment package specifications, transport protocols, sliding window size, the length of slice, timeout interval, the maximum number of retransmission once timeout, the data slice number in a single SOAP message;

2) According to the conventions built by three-way handshake, the sender and the receiver are initialized, including initialization of the sliding window, starting the timers of the sender and receiver;

3) The sender reads data in accordance with the agreed length of slice, then packages the data slices into data message, places data message in a sliding window to be sent. After the transmission process is finished, the sender will wait for acknowledgment messages from the receiver;

4) The receiver receives and parses the data message. According to the parsing results, the receiver will determine whether to place the data into the sliding window or to discard the data. Once all data have been received, the sliding window will slide and the receiver will send the acknowledge message to the sender;

5) Once the sender receives the acknowledgment message, it will parse the SOAP message and determine whether all data have been sent according to the parsing results. If not all data have been sent, the sender will determine whether to retransmit data message or to slide the window in order to read into the follow-up data.

The sender and the receiver will repeat Step 4 and Step 5 until all data are sent.

## III. THE ALGORITHM OF APPLICATION LEVEL CONGESTION CONTROL

Though the SOAP message transmission mechanism based on variable sliding window can effectively improve the transmission efficiency, it also has some disadvantages, e.g. it can cause network congestion when the data are injected into the network too fast. TCP protocol itself has a congestion control mechanism. TCP congestion control usually takes the following three technologies: slowly start, accelerate the decreasing and congestion avoidance [11]. However, because TCP is the transport layer protocol, if the application layer still uses TCP's three congestion control technologies and frequently changes the size of sliding window, it will lead to inefficiency.

This section proposes an application layer congestion control algorithm based on variable sliding window in order to ensure the smooth transmission of the SOAP message. The algorithm of application layer congestion control is shown in Figure 4.

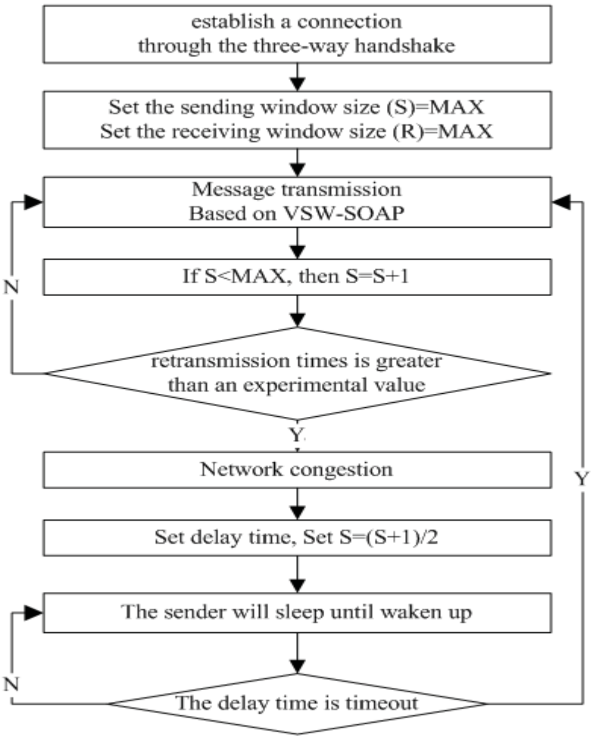

establish a connection
through the three-way handshake
Set the sending window size (S)=MAX
Set the receiving window size (R)=MAX
Message transmission
Based on VSW-SOAP
If S<MAX, then S=S+1
N
retransmission times is greater
than an experimental value
Y
Network congestion
Y
Set delay time, Set S=(S+1)/2
The sender will sleep until waken up
N
The delay time is timeout

Figure 4. The algorithm of application layer congestion control

1) When the sender and the receiver establish a connection through the three-way handshake, set the maximum(MAX) of the sending window size acceptable to both sides, then set the sending window size (S)=MAX and the receiving window size (R)=MAX;

2) In accordance with the sending window size, a sliding window is used to transmit the SOAP message;

3) Once the sender receives an acknowledgment message, it will determine if the sending window size (S) is less than the maximum (MAX) of sending window. If S is less than the maximum of sending window, then S = S +1; Otherwise, keep the sending window size unchanged;

4) During transmission, if the retransmission times of a SOAP message is less than or equal to an experimental value (e.g. 3 times), the system will consider that network transmission is smooth, then will repeat Step2 to Step 4.

If the retransmission times of a SOAP message is greater than an experimental value (e.g. 3 times), the system will consider network congestion, suspend the data transmission for DELAY_T, set S=(S+1)/2. At this time if the receiver did not receive the message for a long time, the receiver will sleep until waken up by the sender.

5) Determine the delay time is timeout. If the delay time is timeout, then continues to suspend data transmission, or proceeds to repeat Step 2 to step 5 until the data transmission is completed.

During the algorithm of congestion control, DELAY_T means data transmission pause time which depends on the congestion time of current network. If the receiver doesn't receive the sender's data for a long time, it will go to sleep state or termination state, and finally will be wakened up by the sender.

## IV. THE TRANSMISSION RELIABILITY OF SOAP MESSAGE

In order to ensure the transmission reliability of the SOAP message based on variable sliding window, the following four cases will be considered during the specific analysis and design: loss of data message, loss of acknowledgment message, duplicate message and data timeout.

Once the data message is lost, if the sender still do not receive the acknowledgment message after waiting for a certain period of time, it will resend the data message. Once the acknowledgment message is lost, if the sender still do not receive the acknowledgment message after waiting for a certain period of time, it will also resend the data message. When the receiver receives the duplicate data message, the receiver will discard the duplicate message, and resend the confirmation of the maximum data slice. Once the receiver waits for data messages timeout, the receiver will resend the confirmation of the maximum data slice, and then wait for the next timeout. If the receiver waits more than timeout maximum value, it can be seen as failure of the transmission.

## V. SIMULATION

In VSW-SOAP, the main factors affecting the transmission efficiency are as follows: sliding window size S (the sending window size Ssender, the receiving window size Sreceiver), the data slice length L and the network transport protocol. The paper won't discuss the network transport protocol for it depends on its realization.

Ssender is negotiated between the sender and the receiver when establishing a connection. Besides, during the transmission, the sender and the receiver can also negotiate to adjust the sliding window size and the length of data slice based on network bandwidth, packet round trip delay, or their own hardware and software environment. Sreceiver is decided by the receiver. If Sreceiver is less than Ssender, the received data is easy to exceed the receiving window range, resulting in data frequently discarded; on the contrary, if Sreceiver is greater than Ssender, it can cause excessive memory overhead although it can cache more data messages. Thus Ssender is best equal or similar to Sreceiver.

A number of test cases are designed to test that S and L how to affect the transmission efficiency. The tests are conducted using the sliding window size from 1 to 50, and the sliding window size is equal to the receiving window size. Because the length of slice is restricted by memory size, the actual test range of length of slice gradually increases from 20K to 10M.

The relationship between the transmitting time and the sliding window size is shown in Figure 5. During the test, HTTP is used as the transport protocol and the length of slice is set 800K.

It can be seen from Figure 5 that the larger the sliding window size, the shorter the transmitting time. When the sliding window size increases to a certain extent, the transmitting time will reach an extreme and almost cease to decrease. The larger the sliding window size, the more the consumption of resources. When the window size increases to a certain extent, the memory and CPU processing speed will become a performance bottleneck. Under the current test environment, when the window size increases from 5 to 10, the transmission efficiency has been increased to a better state. In addition, the sliding window can eliminate the transmitting time difference due to the RTT. When window size increases to a certain extent, the transmitting time is almost the same under the conditions of different RTT.

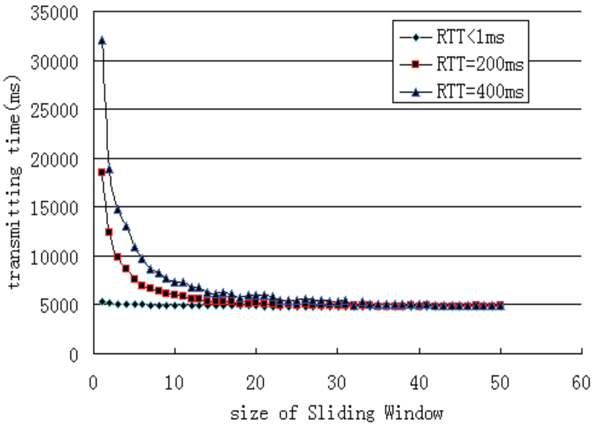


Figure 5. The relationship between the transmitting time and the sliding window size

The relationship between the transmitting time and the length of slice is shown in Figure 6. For the test, RTT is equal to 100ms and the window size is 6. The length of slice gradually increases from 20Kbytes to 10Mbytes.

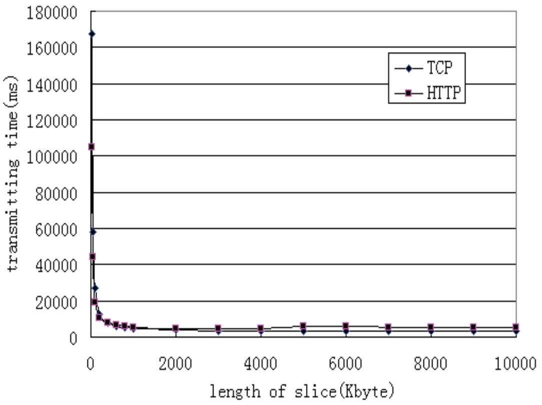


Figure 6. The relationship between the transmitting time and the length of slice

It can be seen from figure 6 that the greater the length of slice, the shorter the transmitting time. When the length of slice increases to a certain extent, the transmitting time will reach an extreme and almost ceases to decrease. Obviously, the length of slice is restricted by the memory, transmission concurrency and CPU processing capacity. Therefore, the length of slice must be set an appropriate value according to practical requirement. Under the current test environment, when the length of slice increases to 400K, the transmission efficiency has been increased to a better state.

## VI. Conclusion

As an extensible lightweight message processing framework, SOAP has become a de facto standard for information exchange in the open network environment. As a SOAP message transmission mechanism based on variable sliding window, VSW-SOAP can effectively improve the transmission efficiency. On the basis of the analysis of the problems faced by data exchange based on SOAP, this paper puts forward the SOAP message transmission mechanism based on variable sliding window, designs an application level congestion control algorithm and discusses the transmission reliability of SOAP message. Simulation results show that transmission efficiency is effectively improved by a reasonable set of parameters of the variable sliding window. With the further development of Web applications such as e-commerce, the VSW-SOAP will also play a greater role.

## References

[1] Martin Gudgin, Marc Hadley et al. SOAP Version 1.2 Part 2: Adjuncts. W3C Recommendation. W3C Group, 24 June 2003, http://www.w3.org/TR/2003/REC-soap12-part2-20030624

[2] Paul V. Biron, Kaiser Permanente. XML Schema Part 2: Datatypes Second Edition W3C Recommendation. W3C Group. http://www.w3.org/TR/2004/REC-xmlschema-2-20041028. 2004

[3] Lin Chuang, Wang Yuan Zhuo, Ren Feng-Yuan. Research on QoS in next generation network, Jisuanji Xuebao/Chinese Journal of Computers, vol.31, pp.1525-1535, September 2008

[4] Christian Werner, Carsten Buschmann, Tobias Jacker, Stefan Fischer. Enhanced Transport Bindings for Efficient SOAP Messaging. In Proceedings of the IEEE International Conference on Web Services (ICWS), 2005

[5] Dan Davis, Manish P arashar, Rutgers. Latency Performance of SOAP Implementations. In Proceedings of the 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid, 2002

[6] Hui Liu, Xin Lin, Minglu Li. Modeling Response Time of SOAP over Http. Proceedings of the IEEE International Conference on Web Services (ICWS) , 2005

[7] Kenneth Chiu, Madhusudhan Govindaraju, Randall Bramley. Investigating the Limits of SOAP Performance for Scientific Computing. In proceedings of the 11 th IEEE International Symposium on High Performance Distributed Computing HPDC-11, 2002

[8] Wang BaoJun, Zhan Ying. A survey and performance evaluation on sliding window for data stream. 2011 IEEE 3rd International Conference on Communication Software and Networks, 2011

[9] Chang J H , Lee W S. A sliding window method for finding recently frequent itemset s over online data streams. J of Information Science and Engineering, 20 (4),2004

[10] Zhu X D , Huang Z Q. Conceptual modeling rules extracting for data streams . Knowledge-based System , 21 (8), 2008

[11] Zhou Jianghe, White Langford B. H ∞ controller design for TCP congestion control. 12th Annual Australian Communications Theory Workshop,2011

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Bézier Curve based Trajectory Planning for an Intelligent Wheelchair to Pass a Doorway

Ling Chen, Sen Wang, Huosheng Hu and Klaus McDonald-Maier
School of Computer Science and Electronic Engineering
University of Essex, Colchester CO4 3SQ, United Kingdom
E-mail: {lcheno, swangi, hhu, kdm}@essex.ac.uk

***Abstract*—Door passing is the basic capability of an intelligent wheelchair. This paper presents a novel approach to address the door passing issue using Bézier curve based trajectory planning. The planed path consists of two segments: one from corridor to door and the other from door to the goal position. For each segment, an optimal Bézier curve is generated as a reference trajectory for an intelligent wheelchair to travel smoothly and accurately subject to corridor constraints, curvature limitation and obstacles. The simulation is conducted to verify the feasibility of the proposed approach, and the results show a good performance in terms of tracking accuracy and good maneuverability.**

***Keywords*-Door passing; Bézier curve; optimization; trajectory planning; obstacle avoidance**

## I. Introduction

Door passing is considered as the fundamental capability of intelligent wheelchairs that are operated in an indoor environment. The wheelchair users are normally elderly and disabled, who may suffer some types of disabilities such as Parkinson's disease. They have difficulty in operating wheelchair using a traditional joystick, not to mention to pass through the constrained doorway. Therefore, there is great demand that the intelligent wheelchair could autonomously travel through confined and narrow doorways without user intervention or carer supervision.

In general, this problem can be addressed by two typical methods - obstacle avoidance and trajectory planning. Among all the methodologies solving obstacle avoidance, potential field based method gains the most popularity [1]–[3]. Although it provides an elegant solution to the obstacle avoidance problem, it has the disadvantage of falling into a local minimum as well as being undulating in some singularity point which is unacceptable for the wheelchair, as human being sitting on the wheelchair will feel uncomfortable in this circumstance. Meanwhile, some existing wheelchairs provide the door passage mode, such as the TAO one [4]. However, to the best of our knowledge, they all use the range finder sensor, e.g., infrared, to detect the door frame and avoid it at a low moving speed and the curvature of the traveling trajectory is not smooth and continous [4].

Recently, a trajectory planning based method drives the wheelchair to follow a desired path. Due to the small ratio of the door width to the wheelchair width (only about 300 mm access space for the wheelchair [5], see Fig. 1), it is difficult for the wheelchair to pass through the door at every heading direction. A frontier point method integrated with simultaneous localization and mapping was used for door passing of wheelchair in [6]. However, it needs to continuously generate the mean frontier points, which makes the curves not smooth and time consuming. A Case-Based Reasoning method for door passing was also presented in [7]. But it neither considered the smoothness of the trajectory nor the narrow structure of the doorway.

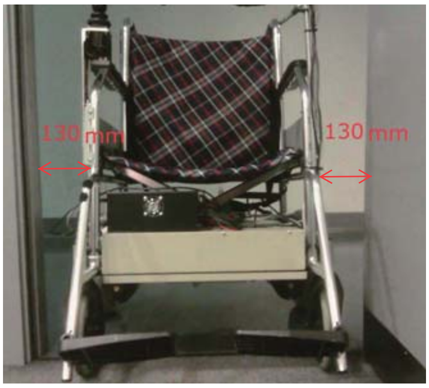


Fig. 1. Wheelchair in the doorway

In order to produce a smooth trajectory, trajectory planning based on circular arc [8], spline [9] and Bézier curves [10], [11] can be adopted. Jolly $et.\ al$ [10] proposed an efficient, Bézier curve based approach for the trajectory planning of a mobile robot, considering the boundary conditions, velocity limitation, etc. Choi $et.\ al$ in [11] presented two trajectory planning algorithms based on Bézier curve for autonomous vehicles with constraints produced by waypoints and corridor width. However, it regards the vehicle as the particle, which is infeasible for the door passing of the wheelchair, in addition the obstacle avoidance has not been taken into account.

This paper intends to propose an efficient strategy for an intelligent wheelchair to pass through the narrow doorway in real time and with the greatest possible success. It introduces the Bézier curve based trajectory planning and the constrained optimization to generate smooth trajectories for the wheelchair to follow without significantly reducing the speed. Moreover, the proposed scheme can avoid the static obstacles while moving to the destination.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

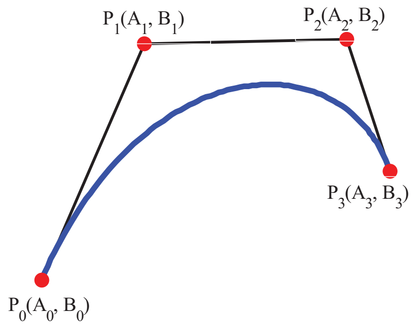


Fig. 2. A cubic Bézier curve defined by four control points

The rest of this paper is organized as follows. Section II gives a brief introduction about the Bézier curve. The door passing strategy is elaborated in Section III, where the problem description, trajectory planning based on Bézier curve as well as the wheelchair control are presented. Section IV presents the simulation results which verify the feasibility of the proposed door passing strategy based on Bézier curves. Finally, a brief conclusion and future work are presented in Section V.

## II. Bézier Curve

Nowadays, Bézier curve as a powerful and efficient tool has been widely used in computer graphics, animation and fonts because of its capacity of smoothing [12]. The Bézier curves can be formulated as

$$B(t) = \sum_{i=0}^{n} b_{i,n}(t)P_i, \quad t \in [0,1] \tag{1}$$

where the polynomials

$$b_{i,n}(t) = \binom{n}{i} t^i (1-t)^{n-i}, \quad i = 0, \ldots n \tag{2}$$

are known as Bernstein basis polynomials of degree $n$, and the binomial coefficient, $\binom{n}{i}$ is expressed as $C_i^n$. The points $P_i$ are called control points for the Bézier curve. The polygon formed by connecting the Bézier points with lines, starting with $P_0$ and finishing with $P_n$, is called the Bézier polygon (or control polygon). The convex hull of the Bézier polygon contains the Bézier curve. The Bézier curves of degree 1, 2 and 3 are called linear, quadratic and cubic Bézier curves respectively, and quadratic and cubic Bézier curves are most common.

The reasons for why Bézier curves can be ideally used for trajectory planning in robotics rest with the following properties of Bézier curve:

- The curve begins at $P_0$ and ends at $P_n$. This is the so called *endpoint interpolation property*.
- The start (end) of the curve is tangent to the first (last) section of the Bézier polygon.

These two properties are exactly the requirements of trajectory planning given two points(start and goal points). The quadratic curve and other curves of higher degree can satisfy with the requirements. However, the quadratic curve is less flexible than cubic curves, and higher degree (more than 3) curves are more computationally expensive to evaluate. Therefore, in this paper, to make a balance between flexibility and computational expense of the proposed curve, cubic curves are chosen as the Bézier curves for trajectory planning for the wheelchair to pass the doorway.

According to (1) and (2), the cubic Bézier curve defined by four control points $P_0(A_0, B_0), P_1(A_1, B_1), P_2(A_2, B_2)$ and $P_3(A_3, B_3)$ in Fig. 2 can be expressed as

$$\begin{aligned} B(t) &= \sum_{i=0}^{3} b_{i,n}(t)P_i \ , \ t \in [0,1] \\ &= (1-t)^3 P_0 + 3(1-t)^2 t P_1 + 3(1-t)t^2 P_2 + t^3 P_3 \end{aligned} \tag{3}$$

As we only consider 2-D environment, the cubic Bézier curve can be specified in the form of $B(t) = (x(t), y(t))$ where

$$x(t) = A_0(1-t)^3 + 3A_1(1-t)^2 t + 3A_2(1-t)t^2 + A_3 t^3 \tag{4a}$$

$$y(t) = B_0(1-t)^3 + 3B_1(1-t)^2 t + 3B_2(1-t)t^2 + B_3 t^3 \tag{4b}$$

## III. Door Passing Strategy

### A. Problem Description

We consider the door passing problem of a wheelchair as a trajectory planning problem. As long as the wheelchair is able to follow the designed path which passes through the doorway without colliding with the door wall as well as other obstacles, then the door passing problem will be well solved. As can be seen in Fig. 3, the wheelchair in the corridor intends to traverse the door on its left side and reach the target in the room. To successfully achieve this goal, the door passing strategy should comply with the following criteria:

- The actual trajectory of the wheelchair has to be graceful and smooth without oscillating or ambiguity, as shown in Fig.3, where the red line represents the desired trajectory in which no obstacles are in the way, and the blue dash line represents the desired trajectory avoiding obstacles.
- Due to the mechanical constraints of the wheelchair, the curvature of the trajectory needs to be limited to certain ranges so that the turning rate will not be too high.
- It is easy for the wheelchair to bump into the door frame if the heading of the wheelchair is not perpendicular to the door plane because of the large volume of the wheelchair. Therefore, it is demanded that the heading of the wheelchair is perpendicular to the door plane when the wheelchair arrives at the $P_d$ point.

We propose the general methods for the wheelchair to pass the doorway as: first, dividing the whole trajectory $\overset{\frown}{P_sP_dP_t}$ into two parts - $\overset{\frown}{P_sP_d}$ and $\overset{\frown}{P_dP_t}$; second, designing the desired trajectory $\overset{\frown}{P_sP_d}$ and $\overset{\frown}{P_dP_t}$ independently based on Bézier curve, to make sure each trajectory meet with the criteria stated above; finally developing an algorithm to control the wheelchair to follow the desired trajectory.

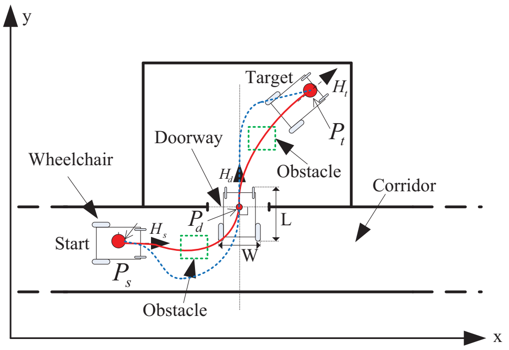


Fig. 3. The schematic description of door passing process of a wheelchair. The red line represents the desired trajectory where no obstacles are in the way, and the blue dash line represents the desired trajectory avoiding obstacles. $P_s$ represents the current position of the wheelchair, $P_d$ is the middle position of the door, and $P_t$ is the target position that the wheelchair is supposed to reach.

### *B. Trajectory Planning based on Bézier Curve*

In this subsection, the trajectory planning for door passing is presented. Its key idea is that the control points which totally define the shape of the Bézier curve are determined using optimization method. The merit of this method lies in the minimum curvature changes of the curve, satisfying the various constraints of the wheelchair and environment and providing the smoothest Bézier curve.

The trajectory planning strategies are divided into two categories depending on whether there is an obstacle. In Fig.3, let $H_s$, $H_d$ and $H_t$ denote the headings at the positions $P_s$, $P_d$ and $P_t$ respectively, and denote the control points of $\widehat{P_sP_d}$ and $\widehat{P_dP_t}$ to be $P_{ci}$ and $P_{ri}$, $i \in 0, 1, 2, 3$, respectively. Then, the trajectory planning is required to generate the cubic Bézier curves which connect the points $P_s$, $P_d$ and $P_t$ with the orientations $H_s$, $H_d$ and $H_t$. Because the target in the room is obstructed by the wall, the wheelchair cannot decide the exact position of the target until it closes to the door. Therefore, the whole trajectory $P_s\widehat{P_d}P_t$ has to be divided into two segments - $\widehat{P_sP_d}$ and $\widehat{P_dP_t}$, for which the Bézier curve based trajectories are independently designed.

*1) Trajectory Planning without Obstacle:* The trajectory planning without obstacle only needs to consider the positions and headings of the beginning and end points of each Bézier curve. There is no demand for obstacle avoidance.

*a) Segment from Corridor to Door:* When the desired door is detected, the wheelchair uses its current position $P_s$, the middle of the door $P_d$ and their corresponding orientations to perform the optimization based control point estimation. The door detection can be achieved by some established techniques, such as computer vision. Since $P_s$ and $P_d$ correspond to the control points $P_{c0}$ and $P_{c3}$ respectively, which are known, the parameter estimation only needs to calculate $P_{c1}$ and $P_{c2}$, which completely determine the shape of the curve once given $P_{c0}$ and $P_{c3}$.

In order to meet the aforementioned requirements, several constraints are devised for the optimization according to the environment and the reality. They are summarized into three types.

i) Corridor structure. In reality, the wheelchair should always be within the corridor while it is turning to the doorway. Therefore, the planned trajectory has to be constrained by the physical structure of the corridor, such as corridor width. Owing to the convex property of Bézier curve, the course can completely lie in the corridor once all the control points are in it. Therefore, the boundary constraint is that $P_{c1}$ and $P_{c2}$ can only be chosen from the corridor area. This can be formulated as

$$P_{c1} \in C \quad P_{c2} \in C \tag{5}$$

where $C$ is the region of the corridor. With this requirement, the produced path can be constrained in the corridor.

ii) Orientations. Since the trajectory planning begins with the heading $H_s$, the slope of tangent line of the designed Bézier curve at position $P_{c0}$ should equal $H_s$. Similarly, the tangent line at $P_{c3}$ should be the normal of the door at the middle in order to overcome the small ratio of door width to the wheelchair's. Therefore, the possible ranges of $P_{c1}$ and $P_{c2}$ can be further reduced to one dimension by the means of this idea. In other words, the $P_{c1}$ and $P_{c2}$ must respectively locate on the tangent lines at positions $P_{c0}$ and $P_{c3}$ due to the aforementioned second property of Bézier curve in Section II. Moreover, $P_{c1}$ has to be searched along the direction of $H_s$, while $P_{c2}$ only can lie on the opposite direction of $H_d$. Hence, the constraints are

$$\theta_0 = \arccos\left(\frac{\overrightarrow{P_{c0}P_{c1}} \cdot H_s}{\|\overrightarrow{P_{c0}P_{c1}}\| \cdot \|H_s\|}\right) = 0 \tag{6a}$$

$$\theta_3 = \arccos\left(\frac{\overrightarrow{P_{c2}P_{c3}} \cdot H_d}{\|\overrightarrow{P_{c2}P_{c3}}\| \cdot \|H_d\|}\right) = 0 \tag{6b}$$

These restrictions ensure that the Bézier curve determined by $P_{c1}$ and $P_{c2}$ can join the starting and end points $P_{c0}$ and $P_{c3}$ with the desired orientations $H_s$ and $H_d$.

iii) Complete convexity. The curvature limitation described in III-A requires that there is no sharp bend or sudden change of curvature in the curve. Thus, the Bézier curve is better to be completely convex. It has been known that a Bézier curve is completely convex if its control polygon is convex. Therefore, $P_{c1}$ and $P_{c2}$ have to be constrained to guarantee the complete convexity of the curve. The cross product method is introduced in this proposed strategy to fulfill this convex requirement. The cross products $\omega_1$ and $\omega_2$ of vectors $\overrightarrow{P_{c0}P_{c1}}$, $\overrightarrow{P_{c1}P_{c2}}$ and $\overrightarrow{P_{c2}P_{c3}}$ are

$$\omega_1 = \overrightarrow{P_{c0}P_{c1}} \times \overrightarrow{P_{c1}P_{c2}} \tag{7a}$$

$$\omega_2 = \overrightarrow{P_{c1}P_{c2}} \times \overrightarrow{P_{c2}P_{c3}} \tag{7b}$$

Therefore, the constraint of the complete convexity is

$$sign(Z_{\omega_1}) \cdot sign(Z_{\omega_2}) > 0 \tag{8}$$

where $sign(Z_{\omega_i})$, $i = 1, 2$, are the rotation direction of the $\omega_i$.

The constrained optimization problem is to find the control points $P_{c1}$ and $P_{c2}$ which make the curve smooth. Then, the curvature and the rate of its change should be as small as possible. According to (4), the curvature of a Bézier curve with respect to $t$ is

$$\kappa(t) = \frac{1}{\rho(t)} = \frac{\dot{x}(t)\ddot{y}(t) - \dot{y}(t)\ddot{x}(t)}{(\dot{x}^2(t) + \dot{y}^2(t))^{3/2}} \tag{9}$$

where $\rho(t)$ is the radius of curve. Therefore, $P_{c1}$ and $P_{c2}$ can be computed by the following constrained optimization problem, which is subjected to (5), (6) and (8):

$$\min_{P_{c1},P_{c2}} \quad \int_0^1 [(\kappa(t))^2 + (\dot{\kappa}(t))^2]dt \tag{10a}$$

$$s.t. \qquad (5) \quad (6) \quad (8) \tag{10b}$$

where $\dot{\kappa}(t)$ is the first derivative of curvature. The interior point method is introduced in this study to solve all the optimization problems. Thus, the calculated $P_{c1}$ and $P_{c2}$ can meet the various requirements and provide an optimal trajectory planning for the first part $\widehat{P_sP_d}$.

*b) Segment from Door to Target:* When the wheelchair arrives in the doorway, it is expected to be at the center of the door with heading being perpendicular to the door. However, because of the disturbances and the control errors, the final position and orientation of last step may not coincide with the desired ones even though the errors may be small under an efficient control strategy. In order to satisfy the third requirement in III-A, the true heading $H_d^a$ needs to be adjusted to the perpendicular direction of the door before addressing the trajectory planning $\widehat{P_dP_t}$.

This trajectory planning from door to target should start from the true position $P_d^T$ with the heading $H_d^a$ in order to maintain the smoothness of the intersection of the two Bézier curves. Moreover, the $P_d^T$ and $P_t$ correspond to the control points $P_{r0}$ and $P_{r3}$ of the Bézier curve respectively. The main idea of this trajectory planning from door to target is similar to the segment from corridor to door except some specified constraints. The constrained optimization problem of this part is

$$\min_{P_{r1},P_{r2}} \quad \int_0^1 [(\kappa(t))^2 + (\dot{\kappa}(t))^2]dt \tag{11a}$$

$$s.t. \qquad P_{r1} \in R \quad P_{r2} \in R \tag{11b}$$

$$\arccos\left(\frac{\overrightarrow{P_{r0}P_{r1}} \cdot H_d^a}{\|\overrightarrow{P_{r0}P_{r1}}\| \cdot \|H_d^a\|}\right) = 0 \tag{11c}$$

$$\arccos\left(\frac{\overrightarrow{P_{r2}P_{r3}} \cdot H_t}{\|\overrightarrow{P_{r2}P_{r3}}\| \cdot \|H_t\|}\right) = 0 \tag{11d}$$

$$sign(Z_{\omega_3}) \cdot sign(Z_{\omega_4}) > 0 \tag{11e}$$

where $R$ is room area, and $Z_{\omega_i}$, $i = 3, 4$, are the cross products of vectors $\overrightarrow{P_{r0}P_{r1}}$, $\overrightarrow{P_{r1}P_{r2}}$ and $\overrightarrow{P_{r2}P_{r3}}$. (11b) is the room boundary constraint, which bounds the planned trajectory by setting the the control points $P_{r1}$ and $P_{r2}$ in the room area. As the orientation and complete convexity constraints, (11c), (11d) and (11e) are respectively very similar to (6a), (6b) and (8) in terms of format and functions. Since the initial value of the optimization problem highly decides if the global minimum can be reached, the beginning point of the period $P_d^T$ is used as the initialization for $P_{r1}$ and $P_{r2}$.

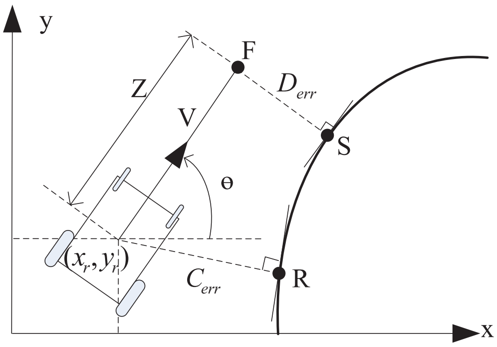


Fig. 4. The position error of the wheelchair is calculated from a point $F$, which is projected in front of the vehicle with a distance of $t$, and unto the desired trajectory to point $S$ [11].

*2) Trajectory Planning with Obstacle:* It is well-known that the door crossing is more difficult for the wheelchair with an obstacle around and this scene is common in practice. In general, the typical techniques divide the door crossing problem into two phases: first, the wheelchair is driven to the area near the door, avoiding the obstacle; then, the door crossing strategy is designed to pass the doorway successfully. However, the special obstacle avoidance step is redundant, especially for the static obstacle, when a one step method is adopted to efficiently address the door passing with obstacle.

This approach also transforms the door crossing with obstacle into the optimization problem, and it is mainly based on the aforementioned trajectory planning. However, the planning of the Bézier curve has to be adjusted according to the obstacle. Assume the static obstacles are distributed as Fig.3. Then, the problem can be addressed by adding an additional constraint to the optimization problem (10) and (11):

$$\begin{aligned} (y(t_k) - y_o)^2 + (x(t_k) - x_o)^2 > d_r^2 \\ t_k \in [0, 1], k \in [1, 1000] \end{aligned} \tag{12}$$

where $x(t_k)$ and $y(t_k)$ are the discrete points sampled from (4), and $d_r$, which depends on the dimensions of the wheelchair and the obstacle, is the allowable minimum distance between the wheelchair and the obstacle. The computed control points $P_{c1}$, $P_{c2}$, $P_{r1}$ and $P_{r2}$ under the above constraint can provide the desired pathes for the wheelchair to pass the door with obstacle avoidance.

### C. Wheelchair Control

In order to drive the wheelchair to track the desired trajectory, a feedback control principle using PID controller is adopted. The position error between the actual position of the wheelchair and the reference trajectory is used as the input of the PID controller. As can be seen in Fig.4, and a point $F$ ahead along the heading of the wheelchair with a distance of $Z$ is used to define the position error, $F$ is projected onto the

TABLE I
THE INPUT PARAMETERS FOR EACH TRAJECTORY

| | $P_s(m)$ | $P_d(m)$ | $P_t(m)$ | $1(H_s, H_d)$ | $2(H_s, H_d)$ | $3(H_d, H_t)$ | $4(H_d, H_t)$ |
|---|---|---|---|---|---|---|---|
| A | (0.1,1.3) | (1.6,1.8) | (0.0,3.5) | (-20,90) | (-40,90) | (90,160) | (90,200) |
| B | (0.1,1.3) | (1.6,1.8) | (3.4,3.5) | (-10,90) | (10,90) | (90,20) | (90,-20) |

reference trajectory at point $S$ such that $\overline{FS}$ is perpendicular to the tangent at $S$. The position error is then represented by the distance $D_{err}$ between point $F$ and $S$. The cross track error $C_{err}$ is defined by the shortest distance between the desired trajectory and the position of the center of the gravity of the wheelchair $(x_r, y_r)$, and $C_{err}$ will be adopted as the indication of the control performance in the simulation analysis in Section IV.The position of the wheelchair $(x_r, y_r)$ is assumed to be estimated by other techniques such as SLAM algorithms.

In this paper, the longitudinal velocity $V$ is considered as constant, and the angular rate $\omega$ as the control output of the PID controller. The discretized PID controller can be expressed as

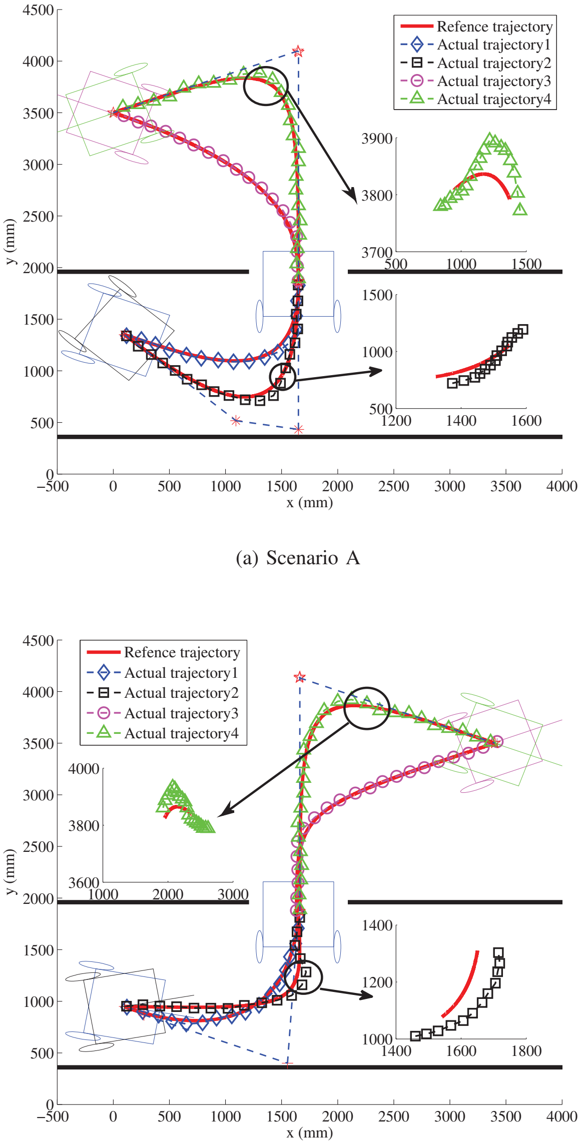
$$\omega = k_p D_{err}^k + k_i T_s \sum_{i=1}^{k} D_{err}^i + \frac{k_d}{T_s}(D_{err}^k - D_{err}^{k-1}) \qquad (13)$$

where $k_p$, $k_i$ and $k_d$ are proportional, integral and derivative gains respectively, $D_{err}^k$ is the current position error, and $D_{err}^{k-1}$ is the position error of last time instance. The reason for why $D_{err}$ rather than $C_{err}$ is used as the input of the PID controller lies in the fact that the required $k_p$ for compensating the same quantity of $C_{err}$ is much larger than that of $D_{err}$, and the larger $k_p$ is more likely to cause system oscillation.

## IV. SIMULATION RESULTS

In order to verify the proposed Bézier curve door passing strategy, simulation work has been conducted. First we carried out two simulation scenarios, namely Scenario A and B, where no obstacles are considered. For each scenario, four different reference trajectories are determined by our Bézier curve based trajectory planning algorithms using optimization technique. The parameters such as the starting position of the wheelchair and its heading, the target position and the heading should be given for determining each reference trajectory.

The required parameters for each reference trajectory are listed in Table I, where $P_s$, $P_d$ and $P_t$ represent the coordinate of starting point in corridor, the coordinate of the middle of the door and the coordinate of the target point in the room respectively. $1(H_s, H_d)$ represents the heading of the starting position and the end position for trajectory 1, and the same meaning goes with other $i(H_s, H_d)$ or $i(H_d, H_t)$ with $i$ being the $i$th trajectory. Using these parameters as the inputs, the Bézier curve based trajectory planning strategy is able to calculate the optimized smooth reference trajectory for the wheelchair to follow gracefully. The PID controller, whose gains are selected to be $k_p = 0.01$, $k_i = 0.001$ and $k_d = 0.001$, is then adopted to drive the wheelchair to track the reference trajectory.

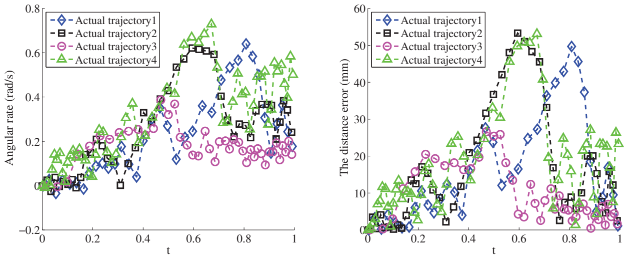


(a) Scenario A

(b) Scenario B

Fig. 5. Reference trajectory and actual trajectory of wheelchair.

Fig. 5 shows the calculated reference trajectories and their corresponding actual trajectories of the wheelchair given the parameters in Table I. As can be seen, several conclusions can be drawn: 1) All the optimized reference trajectories satisfy with the constraints mentioned in Section III-B. That is: first, the wheelchair always is within the corridor; second, the tangent of the Bézier curve at the starting position is equal to the heading of the wheelchair, and the same situation goes with the target position; the control polygon of the designed Bézier curve is convex. 2) The optimized reference trajectories are smooth. 3) The actual trajectories precisely track the reference trajectory, although there are small tracking errors (can be seen from the amplified part of trajectories). With this fact, the basic



(a) Angular rate

(b) Tracking error

Fig. 6. The angular rate and the tracking errors for each trajectory in Scenario A

goal of door passing is successfully achieved.

To analyze the smoothness and the accuracy of the actual trajectories that the wheelchair follows, the angular rate $\omega$ and the distance error $D_{err}$ with respect to the Bézier curve parameter $t$ for each trajectory in Scenario A are shown in Fig. 6a and Fig. 6b respectively. It can be seen that there is no sharp change of the angular rate, which demonstrates the wheelchair is able to move smoothly. The maximum tracking error is about 55 mm, which is both reasonable and acceptable, indicating the relatively high accuracy of the tracking. In order to validate the scheme with obstacle, a scenario where two obstacles are separately distributed in the corridor and room is simulated, see Fig. 7. The blue circles are obstacles, and the widths of wheelchair, corridor and door are assumed to be 75, 160 and 90 centimeters respectively. Compared the trajectory planning without obstacle, which is denoted by the red dash line, with the one affected by the obstacle, which is represented by the diamond and square dash lines, it can be seen that the wheelchair can dynamically adjust its planned trajectories according to the position of the obstacle, avoiding the obstacles and still reaching the destination smoothly.

## V. Conclusions

In this paper, a novel method is proposed to address the tough door crossing problem of the wheelchair. It introduces the Bézier curve based trajectory planning and optimization to produce a smooth and reasonable reference trajectory for the wheelchair to follow. Its merit is that it sufficiently considers the various constraints of the wheelchair and the environment, and can perform the door crossing with graceful and smooth trajectories even hindered by the obstacles. The good performance of the proposed approach is also verified by the means of simulation. It is not limited to the wheelchair, but also can be used for the various mobile robots which intend to pass the narrow door. Our future work will focus on implementing the proposed approach on a real wheelchair.

## Acknowledgments

The authors gratefully acknowledge the support of the UK EPSRC Global Engagements grant EP/K004638/1 and the EU Interreg IV A 2 Mers Seas Zeen Cross-border Cooperation Programme SYSIASS project: Autonomous and Intelligent Healthcare System (http://www.sysiass.eu/). Ling Chen and Sen Wang have been financially supported by scholarships from China Scholarship Council and Essex University. Our thanks also go to Robin Dowling for his technical support during the research.

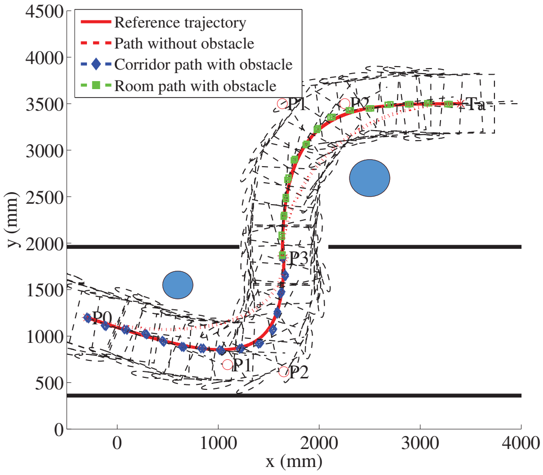


Fig. 7. Trajectories of the with and without obstacles.

## References

[1] O. Khatib, "Real-time obstacle avoidance for manipulators and mobile robots," *The international journal of robotics research*, vol. 5, no. 1, pp. 90–98, 1986.

[2] J. Barraquand, B. Langlois, and J. Latombe, "Numerical potential field techniques for robot path planning," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 22, no. 2, pp. 224–241, 1992.

[3] F. Arambula Cosío and P. Castañeda, "Autonomous robot navigation using adaptive potential fields," *Mathematical and computer modelling*, vol. 40, no. 9-10, pp. 1141–1156, 2004.

[4] R. Simpson, "Smart wheelchairs: A literature review." *Journal of rehabilitation research and development*, vol. 42, no. 4, p. 423, 2005.

[5] BSI, *Design of buildings and their approaches to meet the needs of disabled people-Code of practice*, British Standards Institution. Std. BS 3800, 2009.

[6] F. Cheein, C. De La Cruz, T. Bastos, and R. Carelli, "Slam-based cross-a-door solution approach for a robotic wheelchair," *International Journal of Advanced Robotic Systems*, vol. 7, no. 2, pp. 155–164, 2010.

[7] A. Poncela, C. Urdiales, and F. Sandoval, "A cbr approach to behaviour-based navigation for an autonomous mobile robot," in *2007 IEEE International Conference on Robotics and Automation*. IEEE, 2007, pp. 3681–3686.

[8] A. Scheuer and T. Fraichard, "Continuous-curvature path planning for car-like vehicles," in *Proceedings of the 1997 IEEE/RSJ International Conference on Intelligent Robots and Systems, 1997.*, vol. 2. IEEE, 1997, pp. 997–1003.

[9] J. Connors and G. Elkaim, "Analysis of a spline based, obstacle avoiding path planning algorithm," in *IEEE 65th Vehicular Technology Conference*. IEEE, 2007, pp. 2565–2569.

[10] K. Jolly, R. Sreerama Kumar, and R. Vijayakumar, "A bezier curve based path planning in a multi-agent robot soccer system without violating the acceleration limits," *Robotics and Autonomous Systems*, vol. 57, no. 1, pp. 23–33, 2009.

[11] J. Choi, R. Curry, and G. Elkaim, "Path planning based on bézier curve for autonomous ground vehicles," in *Advances in Electrical and Electronics Engineering-Special Edition of the World Congress on Engineering and Computer Science*. IEEE, 2008, pp. 158–166.

[12] J. Foley, A. Van Dam, S. Feiner, J. Hughes, and R. Phillips, *Introduction to computer graphics*. Addison-Wesley, 1994, vol. 55.

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# A decentralised control strategy for formation flight of unmanned aerial vehicles

Aolei Yang, Wasif Naeem, George W. Irwin, and Kang Li
School of Electronics, Electrical Engineering and Computer Science
Queen's University Belfast, BT9 5AH, UK.
Email: ayang02@qub.ac.uk, w.naeem@qub.ac.uk, g.irwin@qub.ac.uk, k.li@qub.ac.uk

*Abstract*—**This paper presents a methodology to solve formation flight problem for unmanned aerial vehicles (UAVs). It employs a unique "*extension-decomposition-aggregation*" scheme to transform the overall complex formation control problem to a group of sub-problems. The decentralised formation longitudinal and lateral autopilots are designed to support the implementation of the formation flight and manoeuvring of UAVs. Simulation studies have been carried out to verify the performance and effectiveness of the presented cooperative flight strategy.**

*Index Terms*—**Cooperative flight; Decentralised control; Formation autopilot; Unmanned aerial vehicle**

## I. Introduction

Research in cooperative flight control of multi-aircraft systems has attracted growing interest in recent years. After NASA launched the Autonomous Formation Flight (AFF) program, extensive research and development work has been carried out. One of the experiments was conducted by Dryden Flight Research Center, which demonstrated that an F/A-18 flying in the wingtip vortex behind another F/A-18 exhibited a 14% fuel savings [1]. In addition, in [2], the authors presented the benefits that birds "V" shape formation flying brings to energy saving compared to isolated flying, because of the reduction in induced flying drag. Further, its implications are very significant, not just for fuel and energy saving but for other future applications in civilian and military domains. For instance, formation flight can be used to handle increase in air traffic around airports through civil aeroplane formation taking-off and landing to increase use efficiency of airport runway, and applications in military field include aerial refuelling, aircraft logistics, air formation patrol, and carrier landing systems.

Successful formation flight requires the solution of the formation control problem. A number of approaches have been proposed in the literature. For instance, in [3] and [4], the decentralised behaviour-based approach which was inspired by the study of animal behaviour was applied to a group of mobile vehicles. In [5], the leader-following strategy was employed, which has been widely used to deal with the aircraft formation flight control. Virtual structure approach for synchronising UAV position tracking control and for maintaining formation geometry was presented in [6]. Further, the authors have introduced the extension-decomposition-aggregation (EDA) formation control strategy in [7] to translate the complex formation control problem into a group of sub-problems which are simpler to solve. There, the EDA scheme was applied to the formation control of a point mass robot in the horizontal plane.

In this paper, the EDA strategy is further extended and applied to the complex cooperative flight control of UAVs. A nonlinear *Aerosonde* UAV model is linearised about specific trim conditions, and the multivariable $H_\infty$ control methodology based on linear matrix inequalities (LMI) is then employed to design the longitudinal and lateral formation controllers to maintain the corresponding formation stability. Additionally, proportional-integral (PI) compensators are designed to eliminate steady-state formation errors. Simulation studies showing formation manoeuvres have been performed to demonstrate the effectiveness of the proposed method and the formation stability.

This paper is organised as follows. Section II presents the problem formulation and preliminaries. Section III proposes the EDA formation control methodology, whilst Section IV describes the decentralised longitudinal and lateral formation autopilots design. Section V reports the implementation and simulation results. Finally, concluding remarks and suggestions for future work are provided in Section VI.

## II. Problem formulation and preliminaries

Consider a group of three UAVs flying in a specified formation as depicted in Fig. 1. A reference vehicle (RV) in the group can be selected by a consensus between the UAVs. Note that the RV is different from the leader vehicle in the leader-following approach, since it is mainly used to define the local formation coordinate system (LFCS), which is a convenient way of describing a formation, a simple illustration being shown in Fig. 1.

In contrast to the earth axes frame ($X_E$, $Y_E$, $Z_E$) and the fixed body axes frame ($X_B$, $Y_B$, $Z_B$), the LFCS-axes, $(X_L, Y_L, Z_L)$ are always aligned with the vehicle body axes of the RV. This implies that the LFCS axes always vary with the RV's position and direction. From Fig. 1, the position vector of the $i^{th}$ vehicle in the LFCS is defined as $\mathbf{p}_{Li} = [p_{Lix}, p_{Liy}, p_{Liz}]$, and the desired formation can then be mathematically expressed as $F_d = [\mathbf{p}_{L1}; \mathbf{p}_{L2}; \cdots; \mathbf{p}_{LN}]$, where $L$ represents the LFCS and $N$ is the number of vehicles in the formation.

However, it is not enough to describe a formation by only defining the position in the LFCS, since the motion of UAVs in 3D has six degree-of-freedom (6-DOF) defining the

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

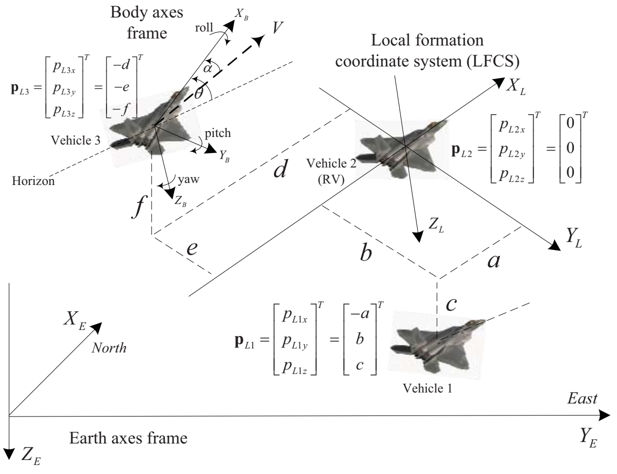

Body axes frame
roll
$X_B$
$V$
$\alpha$
$\theta$
$\mathbf{p}_{L3} = \begin{bmatrix} p_{L3x} \\ p_{L3y} \\ p_{L3z} \end{bmatrix}^T = \begin{bmatrix} -d \\ -e \\ -f \end{bmatrix}^T$
Vehicle 3
pitch
$Y_B$
Horizon
yaw
$Z_B$
$f$
$e$
$d$
Local formation coordinate system (LFCS)
$X_L$
Vehicle 2 (RV)
$\mathbf{p}_{L2} = \begin{bmatrix} p_{L2x} \\ p_{L2y} \\ p_{L2z} \end{bmatrix}^T = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}^T$
$Z_L$
$Y_L$
$b$
$a$
$c$
$X_E$
North
$\mathbf{p}_{L1} = \begin{bmatrix} p_{L1x} \\ p_{L1y} \\ p_{L1z} \end{bmatrix}^T = \begin{bmatrix} -a \\ b \\ c \end{bmatrix}^T$
Vehicle 1
East
Earth axes frame
$Y_E$
$Z_E$

Fig. 1. Local formation coordinate system

positions and attitudes. Thus, a complete formation definition in Euclidean space is given by,

$$F_d = [\mathbf{p}_{L1}; \mathbf{p}_{L2}; \cdots; \mathbf{p}_{LN}] \\ \tilde{\mathbf{x}}_i = [\phi_i, \theta_i, \psi_i, \cdots]^T : i = 1, 2, \cdots, N. \tag{1}$$

where $\phi_i$, $\theta_i$ and $\psi_i$ are the roll, pitch and yaw angles respectively constituting the attitude of the $i^{th}$ individual vehicle. Note that the above attitude angles are just partial elements of the state vector $\tilde{\mathbf{x}}_i$ of the $i^{th}$ vehicle. when a formation change is needed, $\tilde{\mathbf{x}}_i$ must generally be regulated by the formation control algorithm. Subsequently, all vehicles within the group will achieve their new desired relative distances. This implies that stabilising relative positions in a formation is a sufficient condition to achieve equal attitude angles. Thus, the primary function of formation control is to maintain the stability of the relative positions by controlling the states of each vehicle (such as the attitude angles above).

## III. Formation Control Methodology

### A. Novel strategy for decentralised formation control

It has been recognised that complex systems can be decomposed into subsystems of lower dimensionality, and those individual subsystem solutions are then combined in some way to provide an overall system response. Motivated by this idea, the EDA scheme proposed in [7] is described in detail here, and Fig. 2 shows the overall process.

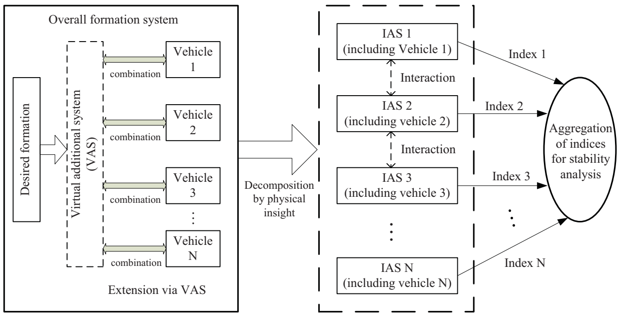

Overall formation system
Desired formation
Virtual additional system (VAS)
combination
Vehicle 1
combination
Vehicle 2
combination
Vehicle 3
combination
Vehicle N
Extension via VAS
Decomposition by physical insight
IAS 1 (including Vehicle 1)
Interaction
IAS 2 (including vehicle 2)
Interaction
IAS 3 (including vehicle 3)
IAS N (including vehicle N)
Index 1
Index 2
Index 3
Index N
Aggregation of indices for stability analysis

Fig. 2. Process of the extension-decomposition-aggregation

For the case of multi-vehicle formation control, each vehicle is a separate entity in the formation space, i.e. there is generally no explicit relationship among the vehicles to represent their formation statuses. The strategy here is to introduce a *virtual* additional system (VAS), which has three main functions: (1) to build a relationship between the isolated vehicles leading to a new overall vehicle formation control system, (2) to involve the desired formation variables or parameters in the overall formation control system, which could be combined with the related individual vehicle model, and (3) to support the subsequent decomposition, and to simplify stability analysis of the overall formation. Note that the VAS is merely an algorithm to act as an "interaction bridge" providing each vehicle with the capability of sensing its local-formation states, which can then be used in the formation control design.

Since the overall formation system involving many variables is difficult to handle as a whole, it is natural to decompose it into several local-formation subsystems. However, it is noted that there is no general systematic procedure for decomposing such a complex dynamical system. Here, using physical insight the overall formation system is decomposed into $N$ individual subsystems, each being called an individual augmented subsystem (IAS) since it combines the individual vehicle model and the local-formation variables. Thus, the initial overall complex formation control problem can now be redefined in terms of stability and set-point tracking for all the decomposed IASs.

In order to analyse the stability of the overall formation, a viable idea is to select a scalar Lyapunov function as an index to represent the stability of each IAS. These indices are then aggregated to mathematically analyse the stability of the overall formation system through the Lyapunov theory. This also brings about a considerable reduction in the dimensionality of the formation stability problem. However, the focus in this paper is on the performance of the proposed formation control system and therefore the stability analysis is not shown here.

### B. A candidate VAS: coupled multiple pendulums

To meet the requirements of the additional system, a virtual multiple-inverted-pendulum system was employed in the original algorithm [7]. However, here a virtual coupled-multiple-pendulum system (CMP) shown in Fig. 3 is used to demonstrate the flexibility of the proposed approach. Here, the CMP consists of $N$ cart-mounted pendulums coupled by $N-1$

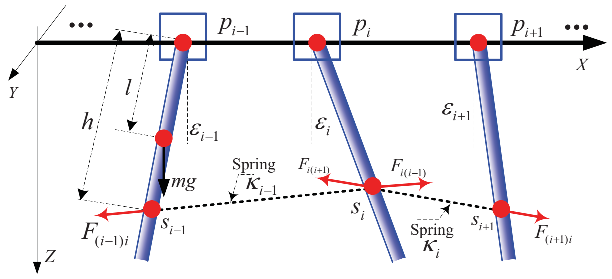


Fig. 3. Coupled multiple pendulums system

springs, and the model of the individual pendulum within the CMP is given by (2),

$$J\ddot{\varepsilon} = -B_c\dot{\varepsilon} - mla\cdot\cos\varepsilon - mgl\sin\varepsilon + W \tag{2}$$

where $J = I + ml^2 = 4ml^2/3$, $I = ml^2/3$ is the inertial moment of pendulum, $\varepsilon$ stands for the deflection angle of each pendulum, $B_c$ is the viscous damping constant at the pivot point, $a$ represents the acceleration, $l$ is the length from the pivot to the gravity centre, $m$ and $g$ represent the mass and the gravitational constant respectively, and $W$ is the resultant torque from the connected springs.

In Fig. 3, each cart of the CMP system can be associated with a vehicle in the formation. The spring connecting two pendulums corresponds to the communication channel between two vehicles, where the magnitude of the torque from the spring force depends on the spring coefficient ($k_s$), the free or natural length ($l_{\kappa_i}$), and the relative distances between the vehicles. Based on these associations, one approach for constructing the relationship is to consider $l_{\kappa_i}$ as the desired formation parameter. Thus, if a formation is disturbed, the force or torque which is caused by the springs must then impact the pendulums, resulting in a change in their deflection angles, $\varepsilon_i$. This implies that the variation of $\varepsilon_i$ is a reflection of the formation error. Thus, the aim of the individual formation control for each IAS is to regulate the deflection angle by manipulating the vehicle states. This will automatically cause the spring to return to its balanced state, i.e. the formation becomes stable. If all the springs return to their natural states or all the deflection angles are equal to zero, this in turn implies that the formation is stable with zero formation error.

Since the CMP is a virtual system, its parameters can be altered by the control algorithm according to the physical dynamics and engineering design requirements of the formation. For example, if the natural length of the springs are dynamically re-defined, the overall vehicle formation will likewise be changed. One of the most interesting aspects of this strategy is that it can be applied to handle vehicle collisions and the desertion problem, both of which are of prime importance in the motion of multiple vehicles. However, this is not discussed here due to lack of space. Recalling the EDA scheme shown in Fig. 2, the overall system is then partitioned or decomposed at the spring positions to obtain the IASs as explained in the next section.

## IV. Decentralised formation autopilot design

### A. Modelling of longitudinal and lateral IASs

Following the standard convention, the 6-DOF aircraft dynamics is approximately decomposed into its longitudinal and lateral components. The longitudinal motion occurs within the plane of symmetry of the aircraft, whereas the lateral-directional motion occurs outside of this plane of symmetry. Based on this separated characteristics and the scheme of EDA, the longitudinal IAS and the lateral IAS can in general be expressed by (3) and (4) respectively.

$$\begin{cases} \dot{\tilde{\mathbf{x}}}_{long} = f(\tilde{\mathbf{x}}_{long}, \delta_{long}) \\ \ddot{\varepsilon}_{long} = -\frac{B_c}{J}\dot{\varepsilon}_{long} - \frac{mgl}{J}\sin\varepsilon_{long} - \frac{ml}{J}a_{long}\cdot\cos\varepsilon_{long} + \frac{1}{J}\cdot W_{long} \end{cases} \tag{3}$$

$$\begin{cases} \dot{\tilde{\mathbf{x}}}_{lat} = g(\tilde{\mathbf{x}}_{lat}, \delta_{lat}) \\ \ddot{\varepsilon}_{lat} = -\frac{B_c}{J}\dot{\varepsilon}_{lat} - \frac{mgl}{J}\sin\varepsilon_{lat} - \frac{ml}{J}a_{lat}\cdot\cos\varepsilon_{lat} + \frac{1}{J}\cdot W_{lat} \end{cases} \tag{4}$$

where $\tilde{\mathbf{x}}_{long}$, $\tilde{\mathbf{x}}_{lat}$ are the state vectors of the longitudinal and lateral dynamics, $\delta_{long}$, $\delta_{lat}$ are their input vectors, $f(\cdot)$ and $g(\cdot)$ are the analytic functions modelling the dynamics of an UAV, $\varepsilon_{long}$ and $\varepsilon_{lat}$ are considered as the variables reflecting their local-formation errors, and $a_{long}$ and $a_{lat}$ are referred to as the longitudinal and lateral accelerations respectively in the inertial reference system.

The designed control framework of an individual aircraft formation system is illustrated in Fig. 4. Here, each plant or so-called longitudinal IAS is generated by combining the longitudinal model with a virtual pendulum (VP) system, and the lateral IAS is obtained likewise. The exogenous inputs $W_{long}$ and $W_{lat}$ are derived from the longitudinal and lateral formation change or error, and are assumed to be *bounded disturbances* to the IASs.

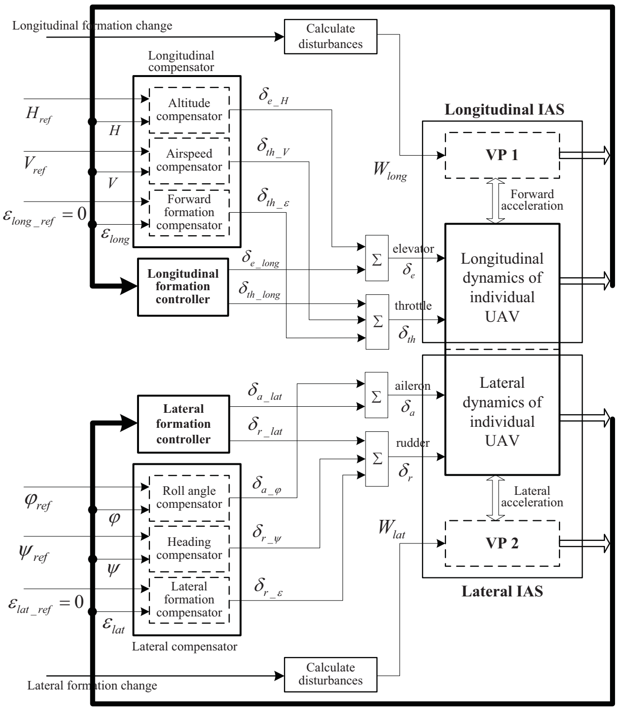


Fig. 4. Longitudinal and lateral autopilots framework

The nonlinear models of the longitudinal and lateral IASs in (3) and (4) can be generally linearised about some specific trim conditions of the aircraft, into corresponding reduced-order state-space equations as shown in (5),

$$\begin{aligned} \dot{\mathbf{x}}_{long} &= A_{long}\mathbf{x}_{long} + B_{1\_long}\mathbf{W}_{long} + B_{2\_long}\mathbf{u}_{long} \\ \dot{\mathbf{x}}_{lat} &= A_{lat}\mathbf{x}_{lat} + B_{1\_lat}\mathbf{W}_{lat} + B_{2\_lat}\mathbf{u}_{lat} \end{aligned} \tag{5}$$

where $\mathbf{x}_{long} = [\tilde{\mathbf{x}}_{long}, \varepsilon_{long}, \dot{\varepsilon}_{long}]^T$ and $\mathbf{x}_{lat} = [\tilde{\mathbf{x}}_{lat}, \varepsilon_{lat}, \dot{\varepsilon}_{lat}]^T$ are the states of the longitudinal and lateral IASs respectively, which are the augmented vectors from the vehicle states $\tilde{\mathbf{x}}_{long}$ and $\tilde{\mathbf{x}}_{lat}$. Similarly, $\mathbf{W}_{long}$ and $\mathbf{W}_{lat}$ are augmented vectors from $W_{long}$ and $W_{lat}$ respectively, whereas $\mathbf{u}_{long}$ and $\mathbf{u}_{lat}$ correspond to the control input vectors in the longitudinal and lateral directions.

### B. Lateral formation autopilot design

For the lateral formation autopilot, the lateral formation motion in the $(X_L,Y_L)$ plane is considered. As shown in Fig. 4, the proposed lateral formation autopilot consists of two components: the multivariable *lateral formation controller* and the *lateral compensators*. Here, the former is designed by using the LMI-based $H_\infty$ robust control methodology [8], [9], [10], to compute the control signals ($\delta_{a\_lat}$ and $\delta_{r\_lat}$), maintain the stability of the lateral IASs and reject the disturbances ($\mathbf{W}_{lat}$). However, it is noted that although these controllers maintain the stability of the lateral formation, zero steady-state formation error cannot be guarantee because of the system type. This is why PI-based lateral compensator given by (6) is supplemented for eliminating this error,

$$D(s) = K_p(1+K_i/s) \tag{6}$$

where $K_p$ and $K_i$ are the proportional and integral parameters respectively, which are individually tuned for each of the roll, heading and lateral formation compensators.

Since the convergence of the roll and heading angles is a necessary condition for the stability of the overall formation system, the roll and heading compensators guarantee the roll and heading angles to asymptotically converge to $\phi_{ref}$ and $\psi_{ref}$ respectively. The nominal value of $\phi_{ref}$ is usually set to zero during normal operation, whereas $\psi_{ref}$ depends on vehicle's particular role in the formation. For instance, if the current vehicle is RV, $\psi_{ref}$ can be calculated by a guidance algorithm, otherwise it should be the same as the heading angle of RV ($\psi_{RV}$). In addition, since the local-formation variable, $\varepsilon_{lat}$, is a reflection of the lateral formation change, the respective formation compensator is also required to converge $\varepsilon_{lat}$ to $\varepsilon_{lat\_ref}$ which should be regulated to zero to ensure no steady-state formation error.

As shown in Fig. 4, the resultant of the complete inputs, $\delta_a$ and $\delta_r$, for the lateral dynamics can be expressed by (7).

$$\begin{aligned}\delta_a &= \delta_{a\_lat} + \delta_{a\_\varphi} \\ \delta_r &= \delta_{r\_lat} + \delta_{r\_\psi} + \delta_{r\_\varepsilon}\end{aligned} \tag{7}$$

These control signals can then drive the individual UAV to maintain the stability of the lateral formation and compensate the steady-state lateral formation error.

### C. Longitudinal formation autopilot design

The longitudinal formation autopilot deals with the longitudinal formation motion in the $(X_L,Z_L)$ plane. It is generally more complicated than the lateral formation autopilot because the airspeed plays a significant role in different stages of the flight, such as take-off, climb/descend and altitude hold stages. The designed strategy here is that the *formation altitude steady-state error* is regulated by the pitch angle, and the *forward formation steady-state error* is eliminated by regulating the airspeed.

In Fig. 4, the longitudinal formation autopilot consist of two components: the *longitudinal formation controller* and the *longitudinal compensators*. The former can be also developed by using the LMI-based $H_\infty$ control methodology to compute the control signals ($\delta_{e\_long}$ and $\delta_{th\_long}$) to guarantee the stability of the longitudinal formation, and the latter is employed to eliminate the steady-state errors of the altitude ($H$), the airspeed ($V$) and forward formation ($\varepsilon_{long}$). The related compensators have the same structure as (6), where $K_p$ and $K_i$ are separately designed for each variable. The resultant control signals, $\delta_e$ and $\delta_{th}$, of the longitudinal formation autopilot are then calculated by (8).

$$\begin{aligned}\delta_e &= \delta_{e\_long} + \delta_{e\_H} \\ \delta_{th} &= \delta_{th\_long} + \delta_{th\_V} + \delta_{th\_\varepsilon}\end{aligned} \tag{8}$$

These control inputs can then maintain the stability of the longitudinal formation in addition to compensating for the longitudinal steady-state formation error.

## V. Implementation and simulations

### A. Implementation of linear longitudinal and lateral IASs

The performance of the proposed methodology was evaluated by using the 6-DOF *Aerosonde* UAV nonlinear simulink model [11]. The autonomous *Aerosonde* has a length of $1.74m$, a wingspan of $2.87m$, maximum payload capacity of $5kg$ and endurance of up to $30hours$. It was designed for applications including long-range weather data acquisition and reconnaissance over oceanic and remote areas.

In this paper, the IAS models were evaluated by trimming the *Aerosonde* model. The trim condition is chosen as: $V = 26m/s$, initial $H = 800m$, $\phi = 0rad$ and fuel mass is $2kg$. The parameter values of CMP system are heuristically chosen as: $m = 50kg$, $g = 9.81m/s^2$, $l = h = 0.3m$, $J = (4/3)ml^2 = 6kg\cdot m^2$, $B_c = 12N\cdot s/m$, $k_s = 0.1N/s$, but can be related to the dynamics of the physical system to be controlled. Furthermore, the relevant symbols representing the states, inputs and outputs of the longitudinal and lateral IASs are given in Table I.

TABLE I
RELEVANT SYMBOLS OF *Aerosonde* LINEAR MODEL

| Variable | Symbol | Variable | Symbol |
|---|---|---|---|
| Longitudinal velocity | $u$ | Airspeed | $V$ |
| Lateral velocity | $v$ | Sideslip angle | $\beta$ |
| Normal velocity | $w$ | Angle of attack | $\alpha$ |
| Roll angle | $\phi$ | Roll rate | $p$ |
| Pitch angle | $\theta$ | Pitch rate | $q$ |
| Yaw angle | $\psi$ | Yaw rate | $r$ |
| Altitude | $H$ | Engine rotation speed | $\omega$ |
| Longitudinal formation variable | $\varepsilon_{long}$ | Lateral formation variable | $\varepsilon_{lat}$ |
| Elevator deflection | $\delta_e$ | Throttle deflection | $\delta_{th}$ |
| Aileron deflection | $\delta_a$ | Rudder deflection | $\delta_r$ |

Specifically, the resulting dynamic equations of the longitudinal IAS are written as (9), which are obtained from the trimming of (3).

$$\begin{aligned}\dot{\mathbf{x}}_{long} &= A_{long}\mathbf{x}_{long} + B_{1\_long}\mathbf{W}_{long} + B_{2\_long}\mathbf{u}_{long} \\ \mathbf{y}_{long} &= C_{long}\mathbf{x}_{long} + D_{1\_long}\mathbf{W}_{long} + D_{2\_long}\mathbf{u}_{long}\end{aligned} \tag{9}$$

where $\mathbf{x}_{long} = [u, w, q, \theta, H, \omega, \varepsilon_{long}, \dot{\varepsilon}_{long}]^T$, $\mathbf{u}_{long} = [\delta_e, \delta_{th}]^T$, and $\mathbf{y}_{long} = [V, \alpha, q, \theta, H, \varepsilon_{long}, \dot{\varepsilon}_{long}]^T$ represent the state, input, and output vectors respectively. The resulting matrices are given by (10).

$$
\begin{aligned}
A_{long} &= \begin{bmatrix} -0.23 & 0.51 & -1.19 & -9.80 & 0 & 0.01 & 0 & 0 \\ -0.55 & -4.38 & 25.39 & -0.46 & 0 & 0 & 0 & 0 \\ 0.41 & -4.73 & -5.06 & 0 & 0 & 0.01 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0.05 & -1.00 & 0 & 26.00 & 0 & 0 & 0 & 0 \\ 33.97 & 1.58 & 0 & 0 & -0.05 & -3.13 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ -0.65 & 0.76 & -0.02 & -24.53 & 0 & 0.03 & -24.5 & -2 \end{bmatrix} \\
B_{1\_long} &= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1667 \end{bmatrix}^T \\
B_{2\_long} &= \begin{bmatrix} 0.35 & -2.60 & -35.90 & 0 & 0 & 0 & 0 & 0.59 \\ 0 & 0 & 0 & 0 & 0 & 346.5 & 0 & 0 \end{bmatrix}^T \\
C_{long} &= \begin{bmatrix} 0.9989 & 0.0466 & 0 & 0 & 0 & 0 & 0 & 0 \\ -0.0018 & 0.0384 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \\
D_{1\_long} &= [0]_{8\times1}, D_{2\_long} = [0]_{8\times2}
\end{aligned} \tag{10}
$$

Similarly, the state-space equations of the lateral IAS are expressed as (11), which were achieved by trimming of (4),

$$
\begin{aligned}
\dot{\mathbf{x}}_{lat} &= A_{long}\mathbf{x}_{lat} + B_{1\_lat}\mathbf{W}_{lat} + B_{2\_lat}\mathbf{u}_{lat} \\
\mathbf{y}_{lat} &= C_{lat}\mathbf{x}_{lat} + D_{1\_lat}\mathbf{W}_{lat} + D_{2\_lat}\mathbf{u}_{lat}
\end{aligned} \tag{11}
$$

where $\mathbf{x}_{lat} = [v, p, r, \phi, \psi, \varepsilon_{lat}, \dot{\varepsilon}_{lat}]^T$, $\mathbf{u}_{lat} = [\delta_a, \delta_r]^T$, and $\mathbf{y}_{lat} = [\beta, p, r, \phi, \psi, \varepsilon_{lat}, \dot{\varepsilon}_{lat}]^T$. Their resulting matrices are given by (12).

$$
\begin{aligned}
A_{lat} &= \begin{bmatrix} -0.68 & 1.21 & -25.97 & 9.80 & 0 & 0 & 0 \\ -4.47 & -21.99 & 10.58 & 0 & 0 & 0 & 0 \\ 0.72 & -2.85 & -1.11 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0.05 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1.00 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ -1.69 & 0 & 0 & 0 & 0 & -24.53 & -2 \end{bmatrix} \\
B_{1\_lat} &= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0.1667 \end{bmatrix}^T \\
B_{2\_lat} &= \begin{bmatrix} -1.51 & -132.34 & -5.22 & 0 & 0 & 0 & -3.77 \\ 3.85 & 2.38 & -24.31 & 0 & 0 & 0 & 9.62 \end{bmatrix}^T \\
C_{lat} &= diag(0.0385, 1, 1, 1, 1, 1, 1) \\
D_{1\_lat} &= [0]_{7\times1}, D_{2\_lat} = [0]_{7\times2}
\end{aligned} \tag{12}
$$

Using the LMI-based $H_\infty$ control methodology [10], the calculated state-feedback gain matrices of the longitudinal and lateral formation controllers are given by (13).

$$
\begin{aligned}
K_{long} &= \begin{bmatrix} 0.0455 & -0.1056 & 0.0890 & 4.7866 & 0.4322 & -0.0002 & -0.0055 & -0.1031 \\ -0.1487 & -0.0114 & 0.0094 & 0.2249 & -0.0716 & -0.0008 & 0.3373 & -0.0131 \end{bmatrix} \\
K_{lat} &= \begin{bmatrix} 0.0157 & 0.0119 & -0.0746 & 0.0970 & 0 & 0.0481 & -0.0146 \\ -0.0131 & -0.0081 & 0.0719 & -0.0166 & 0 & -0.063 & -0.0026 \end{bmatrix}
\end{aligned} \tag{13}
$$

Furthermore, the longitudinal and lateral compensators were designed heuristically to eliminate the corresponding formation error, and are listed in Table II.

TABLE II
PI PARAMETERS OF ALL THE COMPENSATORS

| Lateral compensators | | | Longitudinal compensators | | |
|---|---|---|---|---|---|
| Variables | $K_p$ | $K_i$ | Variables | $K_p$ | $K_i$ |
| $\phi$ | 1.0 | 2.0 | $H$ | 0.13 | 0.7 |
| $\psi$ | 0.5 | 2.0 | $V$ | 0.4 | 0.25 |
| $\varepsilon_{lat}$ | 0.15 | 3.80 | $\varepsilon_{long}$ | 0.5 | 4.0 |

To perform the computer simulations and evaluate the performance of the proposed formation control strategy, the nonlinear *Aerosonde* model was utilised with all the controller gains and parameters as shown previously.

## *B. Formation flight simulation*

In the simulation, the constraints on the pitch and roll angles are given as: $\pm 20 deg$ and $\pm 40 deg$, respectively. The group of UAVs is tasked to navigate the 3D waypoints provided in Table (III) which can be generated from an online or offline path planning algorithm. The *planar projection* of the 3D manoeuvring trajectories are displayed in Fig. 5 to show the formation maintenance and changing when needed, where UAV 2 is the RV and UAV 1 and 3 are its neighbours. The altitudes, airspeeds and attitudes are shown in Fig. 6 and Fig. 7 respectively. Using (1) and Fig. 1, the formation change sequences are mathematically expressed by (14), where $F_{d\_A}$ denotes the initial formation shape, $F_{d\_B}$ illustrates the first formation change about their altitudes at $t = 0s$, and $F_{d\_C}$ indicates the new longitudinal and lateral formation changes at $t = 200s$.

TABLE III
WAYPOINTS OF UAVS FORMATION MANOEUVRE

| *Number* | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| *East*$(m)$ | 0 | 2500 | 3500 | 500 | $-500$ |
| *North*$(m)$ | 0 | 600 | 3500 | 3500 | 1000 |
| *Altitude*$(m)$ | 800 | 1100 | 1400 | 1400 | 1200 |

$$
\begin{aligned}
F_{d\_A} &= \begin{bmatrix} -150 & -150 & 0 \\ 0 & 0 & 0 \\ -150 & 150 & 0 \end{bmatrix} \Rightarrow F_{d\_B} = \begin{bmatrix} -150 & -150 & \mathbf{100} \\ 0 & 0 & 0 \\ -150 & 150 & \mathbf{-100} \end{bmatrix} \\
\Rightarrow F_{d\_C} &= \begin{bmatrix} \mathbf{-300} & \mathbf{-250} & 100 \\ 0 & 0 & 0 \\ \mathbf{300} & \mathbf{450} & -100 \end{bmatrix}
\end{aligned} \tag{14}
$$

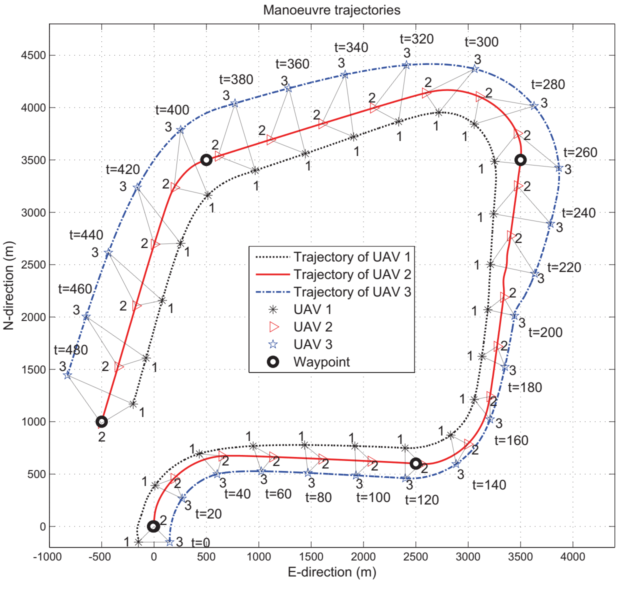


Fig. 5. Planar projection of 3D manoeuvring trajectories

As depicted, the following observations could be made:

1) The group of UAVs completed the desired manoeuvring task of tracking all the required waypoints. Furthermore, the overall formation remained stable, and small formation errors existed only during the formation change and turning manoeuvres.

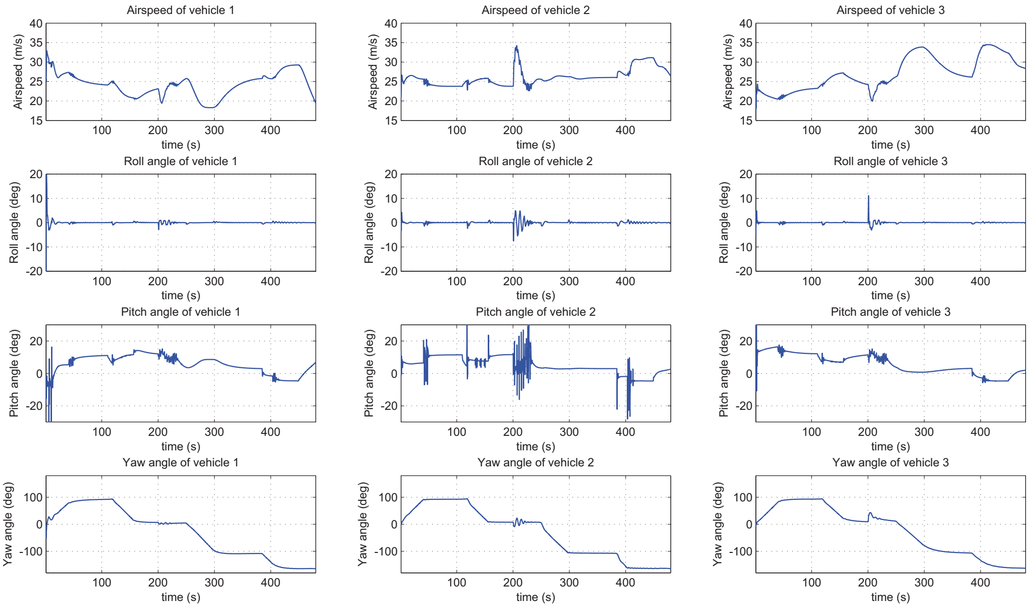


Fig. 7. Dynamics of attitudes while manoeuvring

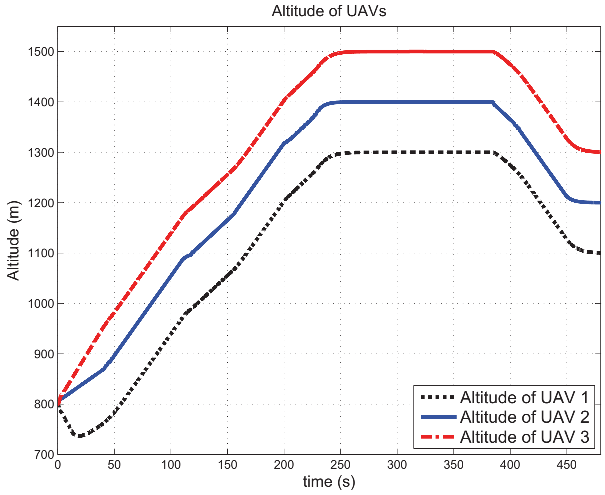


Fig. 6. Dynamics of altitudes while manoeuvring

2) During the turnings, it is clear that the attitude (pitch and roll) of each UAV remains within the specified constraints.
3) The airspeed variation with the formation changing could be observed as it was required to maintain the position of UAVs within the new desired formation as quickly as possible.

## VI. Concluding Remarks

In this paper, the extension-decomposition-aggregation (EDA) scheme was extended to support the design of the decentralised autopilots for the formation flight of UAVs in 3-D. Longitudinal and lateral formation compensators were employed to eliminate the corresponding steady-state formation error. Simulation studies have been performed to verify the feasibility and effectiveness of the EDA-based formation flight strategy. Future work includes mathematical stability analysis of the overall formation and inner/outer collision avoidance.

## Acknowledgment

This work was supported by the EPSRC under UK-China Science Bridge grant EP/G042594/1 and China Scholarship Council.

## References

[1] M. J. Vachon, R. J. Ray, K. R. Walsh, and K. Ennix, “F/A-18 performance benefit measured during the autonomous formation flight project,” in *Technical report NASA/TM-2003-210734*, NASA Dryden, Edward, CA, 2003.

[2] H. Weimerskirch, J. Martin, Y. Clerquin, P. Alexandre, and S. Jiraskova, “Energy saving in flight formation,” *Nature*, vol. 413, pp. 697–698, 2001.

[3] T. Balch and R. Arkin, “Behavior-based formation control for multi-robot teams,” *IEEE Transactions on Robotics and Automation*, vol. 14, pp. 926–939, 1998.

[4] J. R. T. Lawton, R. W. Beard, and B. J. Young, “A decentralized approach to formation maneuvers,” *IEEE Transactions on Robotics and Automation*, vol. 19, no. 6, pp. 933–941, 2003.

[5] Y. Gu, B. Seanor, G. Campa, M. R. Napolitano, L. Rowe, S. Gururajan, and S. Wan, “Design and flight testing evaluation of formation control laws,” *IEEE Transactions on Control Systems Technology*, vol. 14, no. 6, pp. 1105–1112, 2006.

[6] N. Li and H. Liu, “Formation UAV flight control using virtual structure and motion synchronization,” in *Proceeding American Control Conference, Seattle, Washington, USA*, June 11-13 2008, pp. 1782–1787.

[7] A. Yang, W. Naeem, G. W. Irwin, and K. Li, “Novel decentralised formation control for unmanned vehicles,” in *IEEE Intelligent Vehicles Symposium (IV 2012), in Alcalá de Henares, Spain*, June 2012, accepted.

[8] M. Chilali and P. Gahinet, “$H_\infty$ design with pole placement constraints: an LMI approach,” *IEEE Transactions on automatic control*, vol. 41, no. 3, pp. 358–367, 1996.

[9] C. Scherer, P. Gahinet, and M. Chilali, “Multiobjective output-feedback control via LMI optimization,” *IEEE Transactions on automatic control*, vol. 42, no. 7, pp. 896–911, 1997.

[10] D. D. Šiljak and D. M. Stipanović, “Robust stabilization of nonlinear systems: The LMI approach,” *Mathematical Problems in Engineering*, vol. 6, pp. 461–493, 2000.

[11] *AeroSim aeronautical simulation block set V1.2, users guide*, Unmanned Dynamics, 2003.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Experimental Verification of Constrained Iterative Learning Control Using Successive Projection

Bing Chu*, Zhonglun Cai†, David H Owens*‡§, Eric Rogers*, Chris T Freeman* and Paul L Lewin*
*Electronics and Computer Science,
University of Southampton, Southampton SO17 1BJ, UK
(e-mail: {b.chu, etar, cf, pll}@ecs.soton.ac.uk)
†Faculty of Engineering and the Environment,
University of Southampton, Southampton SO17 1BJ, UK
(e-mail: z.cai@soton.ac.uk)
‡Department of Automatic Control and Systems Engineering,
University of Sheffield, Sheffield, S1 3JD, UK
(e-mail: d.h.owens@shef.ac.uk)
§The Italian Institute of Technology,
Via Morego, 30, 16163-Genova, Italy

***Abstract*—In many practical applications, constraints are often present on, for example, the magnitudes of the control inputs. Recently, based on a novel successive projection framework, two constrained iterative learning control (ILC) algorithms were developed with different convergence properties and computational requirements. This paper investigates the effectiveness of these two methods experimentally on a gantry robot facility, which has been extensively used to test a wide range of linear model based ILC algorithms. The results obtained demonstrate the effectiveness of the algorithms in solving one form of the general constrained ILC problem.**

## I. Introduction

Iterative Learning Control (ILC) is a technique for controlling systems operating in a repetitive or trial-to-trial mode with the requirement that a reference trajectory $y_{ref}(p)$ defined over a finite interval $0 \leq p \leq \alpha$ is followed to a high precision, where the constant $\alpha$ denotes the finite trial duration or length. The basic idea in ILC is that information from previous trials is used to update the control input for the next trial in order to sequentially improve performance. Moreover, the next trial input is typically computed during the time taken to reset between successive trials.

Since the original work by [1], ILC has developed into an established area in control systems research and applications. Initial sources for the relevant literature are the survey papers [2] and [3]. These show that a wide range of algorithms have been developed, many of which, particularly those based on a linear plant model, have been experimentally tested.

In many practical applications, constraints are present due, for example, to physical limitations or performance requirements. Hence ILC design must take these constraints into account but most of the currently available ILC results are for unconstrained systems and there are relatively few results for the constrained case. One set of results is due to [4] where a novel nonlinear controller for process systems with input constraints is developed where the learning scheme requires relatively little knowledge of the process model. In [5] an ILC problem with soft constraints is studied where Lagrange multiplier methods are used to develop a solution. [6] uses quadratic optimal design to formulate a constrained ILC problem and suggests that a quadratic optimal design has the capability of dealing with constraints. Also [7] uses a constrained convex optimization technique to solve the constrained ILC problem for linear systems with saturation constraints.

Recently, the ILC design problem with general convex input constraints has been considered in [8]. This work shows that the constrained ILC problem can be formulated in a recently developed successive projection framework, which provides an intuitive but rigorous method for system analysis and design. Based on this, a systematic approach for constraints handling is provided and two algorithms to solve this problem developed. The convergence analysis shows that when perfect tracking is possible, both algorithms can achieve this goal whereas the computation of one algorithm is much less than the other at the cost of slightly slower convergence rate. When perfect tracking is not possible, both algorithms converge to asymptotic values representing a ”best fit” solution. Again the more computationally complex algorithm has the best convergence properties. It was also found that the input and output weighting matrices have an interesting effect on the convergence properties of the algorithms.

The main aim of this paper is to give experimental results to verify the effectiveness of the constrained ILC algorithms using a gantry robot facility previously used to test a wide range of ILC algorithms, including Norm Optimal ILC (NOILC) [9]. The paper is organized as follows. In Section 2, the required results from the derivation of the constrained ILC algorithms are given. Then in Sections 3 and 4, the gantry robot facility and the test parameters are described. The experimental results are given in Section 5 and Section 6 gives conclusions and suggestions for further research.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

## II. Iterative Learning Control for Constrained Linear Systems

Consider the following discrete linear time-invariant system

$$\begin{aligned} x_k(t+1) &= Ax_k(t) + Bu_k(t), \\ y_k(t) &= Cx_k(t), \end{aligned} \tag{1}$$

where $t$ is the time index (i.e. sample number), $k \geq 0$ is the trial index and $x_k(0) = x_0, k = 1, 2, \cdots$ is the same for all trials. The control objective is to track a given reference signal $r(t)$ and $u_k(t), x_k(t), y_k(t)$ are input, state and output vectors, respectively, of the system on trial $k$. In operation, a trial is completed, the system is reset and a new trial begins. The ILC design uses information from previous trial(s) to compute the control input for the next trial in a manner that improves tracking performance from trial-to-trial.

Before presenting the main results, the operator form of the dynamics is introduced using the well-known lifted-system representation, which provides a straightforward "$N \times N$ matrix" approach in the analysis of discrete-time ILC [10], [11].

Assume, for simplicity, the relative degree of the system is unity, i.e. the generic condition $CB \neq 0$ is satisfied (the case when the system relative degree is greater than one follows as an obvious generalization), then the system state-space model (1) on trial $k$ can be written in the form

$$y_k = Gu_k + d, \tag{2}$$

where $G$ and $d$ are the $N \times N$ and $N \times 1$ matrices

$$\begin{aligned} G &= \begin{bmatrix} CB & 0 & \cdots & 0 & 0 \\ CAB & CB & \ddots & 0 & 0 \\ CA^2B & CAB & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & CB & 0 \\ CA^{N-1}B & \cdots & \cdots & CAB & CB \end{bmatrix} \\ d &= \begin{bmatrix} CAx_0 & CA^2x_0 & CA^3x_0 & \cdots & CA^Nx_0 \end{bmatrix}^T . \end{aligned} \tag{3}$$

The $N \times 1$ vectors of input, output and reference time series $u_k, y_k$ and $r$ are defined as

$$\begin{aligned} u_k &= \begin{bmatrix} u_k(0) & u_k(1) & \cdots & u_k(N-1) \end{bmatrix}^T, \\ y_k &= \begin{bmatrix} y_k(1) & y_k(2) & \cdots & y_k(N) \end{bmatrix}^T, \\ r &= \begin{bmatrix} r(1) & r(2) & \cdots & r(N) \end{bmatrix}^T. \end{aligned} \tag{4}$$

Also no loss of generality arises from assuming that $d = 0$ (non-zero $d$ can be incorporated into the reference signal by replacing $r$ with $r - d$). Hence (2) becomes

$$y_k = Gu_k, \tag{5}$$

where $G$ is nonsingular and hence invertible.

Tracking error improvements from trial-to-trial are achieved in ILC by the design of a control law of the following general form

$$u_{k+1} = f\left(e_{k+1}, \ldots, e_{k-s}, u_k, \cdots, u_{k-r}\right). \tag{6}$$

When $s > 0$ or $r > 0$, (6) is termed a higher order updating law. This paper only considers algorithms of the form $u_{k+1} = f\left(e_{k+1}, e_k, u_k\right)$. For higher order algorithms, refer to [12], [13] and the references therein. The ILC design problem can now be stated as finding a control updating law (6) such that the system output has the asymptotic property that $e_k \to 0$ as $k \to \infty$.

There are many design methods to solve the ILC problem. The one used in this paper is based on a quadratic (norm) optimal formulation [14] where, on each trial, a performance index is minimized to obtain the system input time series vector to be used on the next trial. The basis of this paper is NOILC that designs the control input to minimize the performance index

$$J_{k+1}(u_{k+1}) = \|e_{k+1}\|_Q^2 + \|u_{k+1} - u_k\|_R^2, \tag{7}$$

subject to the constraint $e_{k+1} = r - Gu_{k+1}$, where $G$ is the operator representation of the system (1) and $Q$ and $R$ are positive definite weighting matrices. Also $\|e\|_Q^2$ denotes the quadratic form $e^TQe$ and similarly for $\|\cdot\|_R^2$. Solving this optimization problem gives the following optimal choice for the time series vector $u_{k+1}$

$$u_{k+1} = u_k + R^{-1}G^TQe_{k+1} \tag{8}$$

which, when $k \to \infty$, asymptotically achieves perfect tracking. This well-known NOILC algorithm has many appealing properties including implementation in terms of Riccati state feedback. More details concerning NOILC can be found in [14]–[17].

In practical applications, system constraints are encountered and of different forms, e.g., input constraints, input rate constraints and state or output constraints. Constraints can be divided into two classes termed hard and soft, respectively. Hard constraints are those on magnitude(s) at each point in time, for example, output limits on actuators. Soft constraints are those that are applied to the whole function rather than its point-wise values e.g. constraints on total energy usage. This paper only considers input constraints.

Suppose the input is constrained to be in a set $\Omega$, which is taken to be a closed convex set in some Hilbert space $H$. In practice, the set $\Omega$ is often of simple structure. For example, the following are often encountered:

- input saturation constraint:

$$\Omega = \{u \in H : |u(t)| \leq M(t)\}$$

- input amplitude constraint:

$$\Omega = \{u \in H : \lambda(t) \leq u(t) \leq \mu(t)\}$$

- input sign constraint:

$$\Omega = \{u \in H : 0 \leq u(t)\}$$

- input energy constraint:

$$\Omega = \{u \in H : \sum_{t=0}^{N-1} u^2(t) \leq M\}$$

If there are no constraints, the ILC design problem is relatively easy to solve and there are many design methods in the literature. However, when constraints are present, the problem becomes more complicated since it is now necessary to decide how to incorporate them into the design process and retain known performance properties. In what follows, two constrained ILC algorithms recently developed in [8] using a novel successive projection framework are summarized.

*Algorithm 1:* Given any initial input $u_0$ satisfying the constraint with associated tracking error $e_0$, the input sequence $u_{k+1}, k = 0, 1, 2, \cdots$ , defined by

$$u_{k+1} = \arg\min_{u\in\Omega} \left\{ \|r - Gu\|_Q^2 + \|u - u_k\|_R^2 \right\}, \tag{9}$$

also satisfies the constraint and iteratively solves the constrained ILC problem.

Constrained Algorithm 1 has the following properties:

*Theorem 1:* Algorithm 1 converges to point $u_s^*$ which is uniquely defined by the following optimization problem

$$u_s^* = \arg\min_{u\in\Omega} \|r - Gu\|_Q^2. \tag{10}$$

Moreover, this convergence is monotonic in the tracking error, that is,

$$\|e_{k+1}\| \leq \|e_k\|, k = 0, 1, \cdots . \tag{11}$$

In the case when perfect tracking is possible, Constrained Algorithm 1 will converge to zero tracking error and has desirable properties of monotonic convergence in tracking error norm. However, it requires the solution of a quadratic programming (QP) problem and can be computationally demanding and in [8] two efficient solution methods were developed but are omitted here for brevity.

Another algorithm that is less computationally demanding is the following.

*Algorithm 2:* Given any initial input $u_0$ satisfying the constraint with associated tracking error $e_0$, the input sequence $u_{k+1}, k = 0, 1, 2, \cdots$ , defined by the solution of the input unconstrained NOILC optimization problem

$$\tilde{u}_k = \arg\min_{u} \left\{ \|r - Gu\|_Q^2 + \|u - u_k\|_R^2 \right\}, \tag{12}$$

followed by the simple input projection

$$u_{k+1} = \arg\min_{u\in\Omega} \|u - \tilde{u}_k\| \in \Omega, \tag{13}$$

also satisfies the constraint and iteratively solves the constrained ILC problem.

*Remark 1:* The first step in Algorithm 2 requires the solution of the input unconstrained NOILC optimization problem (12). Unlike Algorithm 1, which may cause computational problems in solving the large constrained QP problem (9), (12) has a real-time Riccati solution [16]

$$\begin{array}{rcl} u_{k+1}(t) & = & u_k(t) - R^{-1}B^T M(t), \\ M(t) & := & K(t)(I + BR^{-1}B^T K(t))^{-1} \times \\ & & A(x_{k+1}(t) - x_K(t)) - \xi_{k+1}(t), \end{array} \tag{14}$$

where $K(t)$ satisfies the Riccati equation

$$\begin{array}{r} K(t) = A^T K(t+1)A + C^T QC - A^T K(t+1)B \times \\ (B^T K(t+1)B + R)^{-1} B^T K(t+1)A \end{array} \tag{15}$$

with final time condition $K(N) = 0$. Moreover, $\xi_{k+1}(t)$ satisfies the differential equation

$$\begin{array}{r} \xi_{k+1}(t) = (I + K(t)BR^{-1}B^T)^{-1}(A^T \xi_{k+1}(t+1) \\ + C^T Qe_k(t+1)), \end{array} \tag{16}$$

which is computable in reverse time as it is driven by tracking error from the previous trial $k$ [16].

*Remark 2:* The second step in Algorithm 2 requires the solution of the problem (13) and would appear to need the application of optimization methods. However, in practice the input constraint $\Omega$ is often a point-wise constraint and the solution of (13) can be computed easily. For example, when $\Omega = \{u \in H : |u(t)| \leq M(t)\}$, the solution is

$$u_{k+1}(t) = \left\{ \begin{array}{ll} M(t) & : \tilde{u}_k(t) > M(t) \\ \tilde{u}_k(t) & : |\tilde{u}_k(t)| \leq M(t) \\ -M(t) & : \tilde{u}_k(t) < -M(t) \end{array} \right. \tag{17}$$

for $t = 0, \cdots , N - 1$.

Constrained Algorithm 2 requires less computational effort but, unlike Constrained Algorithm 1, it cannot guarantee monotonic convergence of the tracking error norm. Instead, it achieves monotonic convergence of weighted error norm, as shown in the following theorem.

*Theorem 2:* When perfect tracking is not possible, Algorithm 2 converges to a point $u_s^*$ which is uniquely defined by the following optimization problem,

$$u_s^* = \arg\min_{u\in\Omega} \left\{ \|Ee\|_Q^2 + \|Fe\|_R^2 \right\}. \tag{18}$$

Moreover, this convergence is monotonic with respect to the following performance index

$$J_k = \|Ee_k\|_Q^2 + \|Fe_k\|_R^2, \tag{19}$$

where

$$\begin{array}{rcl} e & = & r - Gu \\ E & = & I - G\left(G^T QG + R\right)^{-1} G^T Q \\ F & = & \left(G^T QG + R\right)^{-1} G^T Q \end{array} . \tag{20}$$

It was also shown in [8] that when perfect tracking is not possible, the choice of $Q$ and $R$ in Algorithm 2 has an interesting effect on the convergence properties. In particular, there is a compromise between the convergence rate and the tracking performance: using a smaller $R$ will result in faster convergence, however, with a larger final tracking error.

Simulation studies have demonstrated that the two algorithms considered above can solve the constrained ILC problem efficiently, with different convergence properties and computational requirements. The remainder of this paper examines the performance of the algorithms experimentally on a gantry robot that has been used to tests an extensive range of linear model based ILC algorithms [9], [18].

## III. GANTRY ROBOT TEST FACILITY

This approach has been experimentally implemented on a 3-axis gantry robot. Figure 1 shows this experimental facility where the robot head performs a ‘pick and place’ task and is similar to systems which can be found in many industrial applications. These include food canning, bottle filling or automotive assembly, all of which require accurate tracking control, each time the operation is performed, with a minimum level of error in order to maximize production rates. This is an obvious general area for application of ILC.

Each axis of the gantry robot has been modeled based on frequency response tests where, since the axes are orthogonal, it is assumed that there is minimal interaction between them. Here we first consider the X-axis (the one parallel to the conveyor in Figure 1) and frequency response tests (via Bode approximate gain plots in Figure 2) result in a 7th order continuous-time transfer-function as an adequate model of the dynamics on which to base control systems design.

$$G_X(s) = \frac{13077183.4436(s+113.4)}{s(s^2+61.57s+1.125\times10^4)} \times \cdots$$
$$\frac{(s^2+30.28s+2.13\times10^4)}{(s^2+227.9s+5.647\times10^4)(s^2+466.1s+6.142\times10^5)}. \tag{21}$$

The dynamics have been sampled at $T_s = 0.01$ seconds to yield the discrete state space model which can be used to compute the model matrix $G$ according to (3). The same procedure has been applied to the Y-axis and Z-axis. For the details see [9].

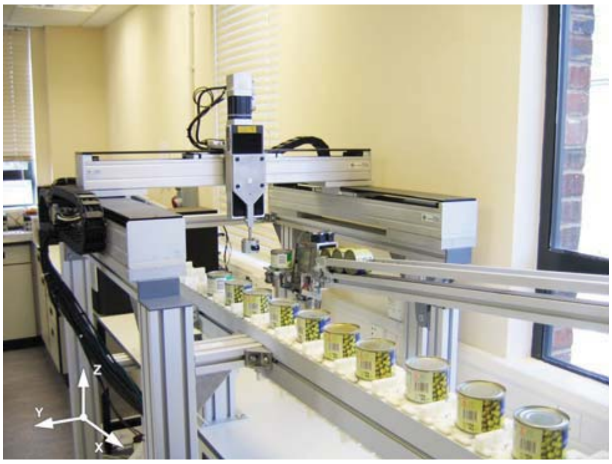


Fig. 1. The multi-axis gantry robot.

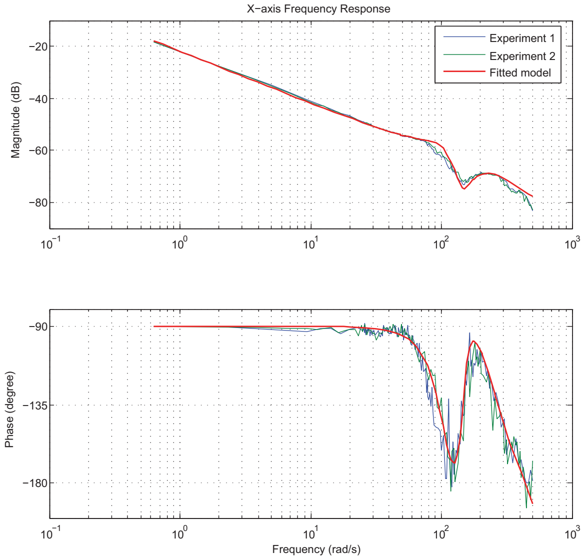


Fig. 2. Frequency response test results and fitted model

$$A_x =$$
$$\begin{bmatrix} 0.035 & 1.000 & 0 & 0 & 0 & 0 & 0 \\ -0.008 & 0.035 & -0.071 & 0.156 & 0 & 0 & 0 \\ 0 & 0 & -0.158 & 1.000 & 0 & 0 & 0 \\ 0 & 0 & -0.078 & -0.158 & 0.569 & 0.698 & 1.319 \\ 0 & 0 & 0 & 0 & 0.388 & 1.000 & 0 \\ 0 & 0 & 0 & 0 & -0.390 & 0.388 & 1.080 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1.000 \end{bmatrix},$$
$$B_x = \begin{bmatrix} 0 & 0 & 0 & 0.0164 & 0 & 0.0134 & 0.0197 \end{bmatrix}^T,$$
$$C_x = \begin{bmatrix} -0.0003 & -0.0054 & -0.0145 & 0.0316 & 0 & 0 & 0 \end{bmatrix}.$$

## IV. TEST PARAMETERS

The gantry robot is designed to repeatedly complete a pick-and-place motion in synchronization with a moving conveyor. A reference trajectory for the gantry movement has been predefined with the purpose of synchronizing its motion with that of the conveyor, which is running at a constant speed. Each axis is controlled individually and has its own reference trajectory and they are combined to form the 3D reference trajectory given in Figure 3 which clearly shows the ‘pick and place’ action. The signal duration is 2 seconds.

## V. EXPERIMENTAL RESULTS

This section gives the results of experimental performance of the two constrained ILC algorithms given in Section 2. Input constraints were imposed such that for each axis, the amplitude of the input voltage was limited to be 90% that of the optimal input $u^*$, which produces performance tracking. Note that under these constraints perfect tracking is not possible, which is more practically relevant than the trivial case where perfect tracking is achievable).

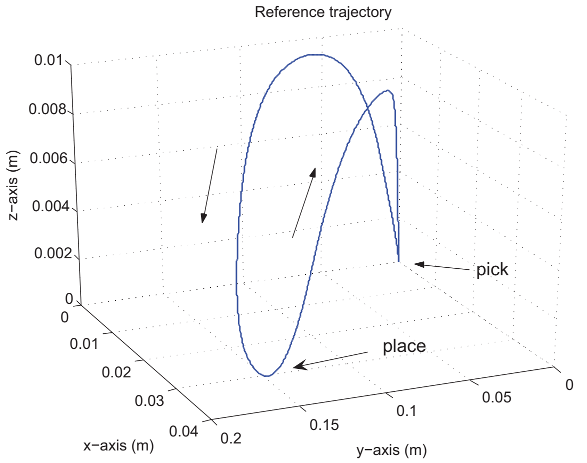


Fig. 3. 3-D combined reference trajectory.

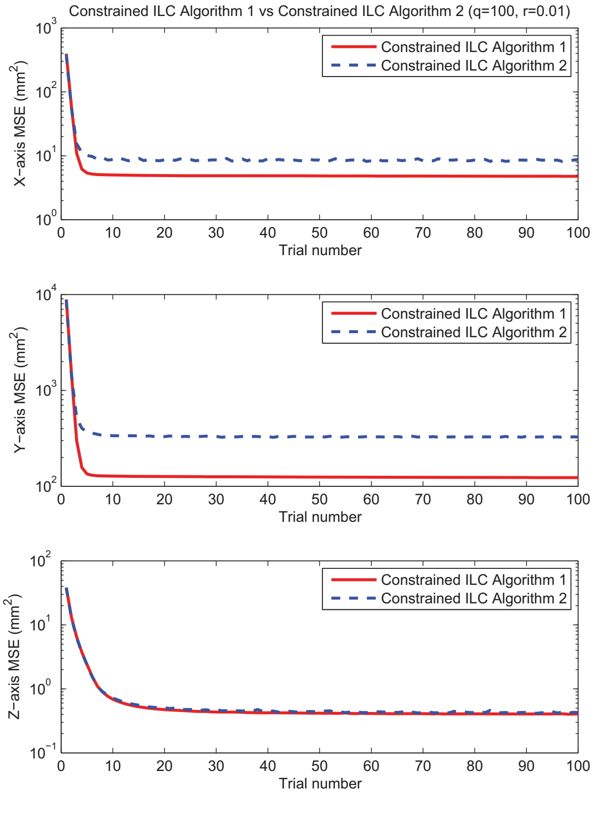


Fig. 4. Comparison of Convergence with $Q = 100I, R = 0.01I$

The cost function weighting matrices are chosen as diagonal matrices with common diagonal entries of 100 and 0.01 for all three axes and the experimentally results are given in Figure 4. These results confirm that both algorithm solve the constrained ILC problem and converge to some final values. The convergence of Constrained ILC Algorithm 1 is monotonic in the tracking error norm, whereas Constrained ILC Algorithm 2 is not. Moreover, the final tracking error norm of Constrained ILC Algorithm 1 is smaller than that of Algorithm 2, which is consistent with the theoretical predictions.

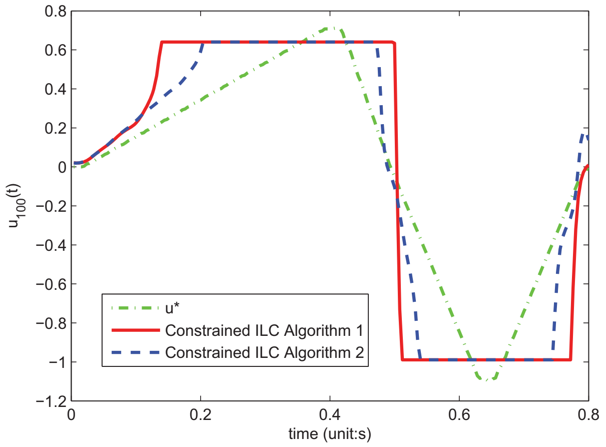


Fig. 5. Input on the $100^{th}$ trial for the X-axis

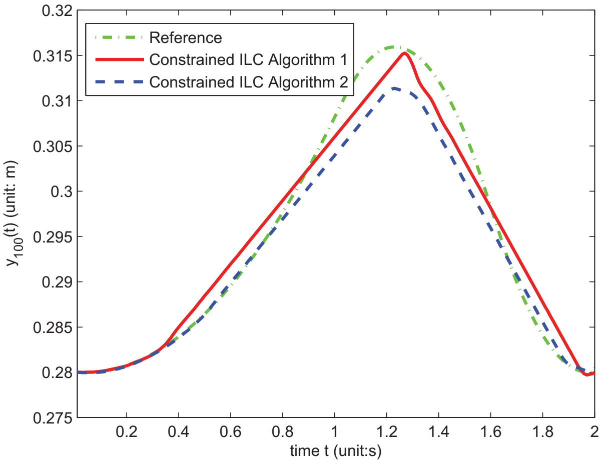


Fig. 6. Output on the $100^{th}$ trial for the X-axis

To expand the discussion of these results , the input, output and tracking error on $100^{th}$ trial for X-axis are given in Figures 5-7. Using Constrained ILC Algorithm 1, a smaller tracking error (on average) is obtained, compared to Constrained ILC Algorithm 2. Note also that the input computed by Algorithm 2 does not just enforce saturation on the original input but adds some compensation (Figure 5).

Figures 8 shows the effects of varying the selection of the weighting matrices keeping the diagonal structure but changing the control input weighting to 0.1. Compared to the results in Figure 4, this new choice puts larger weighting on the input change, leading to slower convergence. Varying this weighting value has no effect on the final tracking error for Algorithm 1

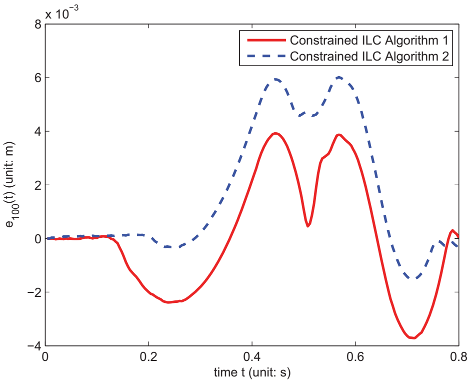


Fig. 7. Tracking error on the $100^{th}$ trial for the X-axis

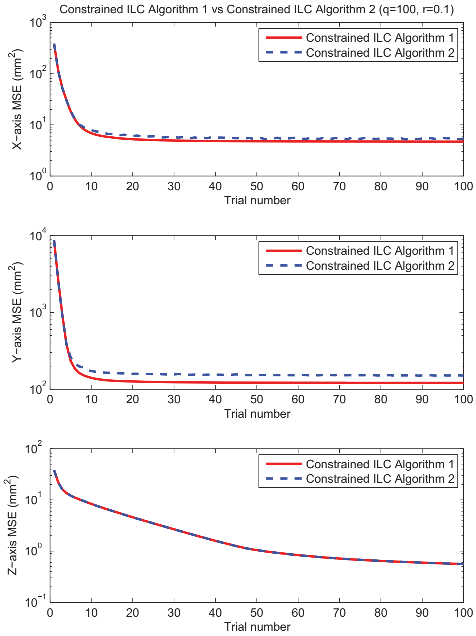


Fig. 8. Comparison of Convergence with control weighting increased to 0.1.

but in Algorithm 2 it results in a better (almost optimal) tracking accuracy, which verifies the theoretical results of Section 2.

Using these results, it can be concluded that Constrained ILC Algorithm 1 performs better than Algorithm 2. However, this is achieved at the expense of higher computational load, which may be not acceptable in some applications but Algorithm 2 achieves nearly optimal performance using a quite simple computation, which is equally (if not more) important in many cases.

## VI. Conclusions

In this paper, two constrained ILC algorithms developed based a successive project framework in [19] have been tested on a gantry robot facility. The results confirm that both algorithms can solve the constrained ILC problem efficiently, while the computation of one algorithm is much less than the other at the cost of slightly sacrificed convergence performance. This requires a compromise between the performance/accuracy and the computational cost.

The experimental results in this paper are based on a linear model of the gantry robot, where nonlinearities are neglected and can be treated as model uncertainty. The results clearly demonstrate certain robustness of the algorithms used. However, further theoretical robustness analysis still needs to be done and constitutes part of planned future research.

## References

[1] S. Arimoto, S. Kawamura, and F. Miyazaki, “Bettering operations of robots by learning,” *Journal of Robotic Systems*, vol. 1, no. 1, pp. 123–140, 1984.

[2] H. S. Ahn, Y. Chen, and K. L. Moore, “Iterative learning control: brief survey and categorization,” *IEEE Transactions on Systems, Man and Cybernetics, Part C*, vol. 37, no. 6, pp. 1109–1121, 2007.

[3] D. Bristow, M. Tharayil, and A. Alleyne, “A survey of iterative learning control: A learning-based method for high-performance tracking control,” *IEEE Control Systems Magazine*, vol. 26, no. 3, pp. 96–114, 2006.

[4] C. Chen and S. Peng, “Learning control of process systems with hard input constraints,” *Journal of Process Control*, vol. 9, no. 2, pp. 151–160, 1999.

[5] S. Gunnarsson and M. Norrlof, “On the design of ILC algorithms using optimization,” *Automatica*, vol. 37, no. 12, pp. 2011–2016, 2001.

[6] J. Lee, K. Lee, and W. Kim, “Model-based iterative learning control with a quadratic criterion for time-varying linear systems,” *Automatica*, vol. 36, no. 5, pp. 641–657, 2000.

[7] S. Mishra, U.Topcu, and M. Tomizuka, “Iterative learning control with saturation constraints,” in *Proceedings of American Control Conference*, 2009, pp. 943–948.

[8] B. Chu and D. H. Owens, “Iterative learning control for constrained linear systems,” *International Journal of Control*, vol. 83, no. 7, pp. 1397–1413, 2010.

[9] J. D. Ratcliffe, P. L. Lewin, E. Rogers, J. J. Hatonen, and D. H. Owens, “Norm-optimal iterative learning control applied to gantry robots for automation applications,” *IEEE Transactions on Robotics*, vol. 22, no. 6, pp. 1303–1307, 2006.

[10] J. Hatonen, “Issues of algebra and optimality in iterative learning control,” Ph.D. dissertation, University of Oulu, Finland, 2004.

[11] J. Hatonen, D. Owens, and K. Moore, “An algebraic approach to iterative learning control,” *International Journal of Control*, vol. 77, no. 1, pp. 45–54, 2004.

[12] Z. Bien and K. Huh, “Higher-order iterative learning control algorithm,” *IEE Proceedings, Part D: Control Theory and Applications*, vol. 136, no. 3, pp. 105–112, 1989.

[13] J. Hatonen, D. Owens, and K. Feng, “Basis functions and parameter optimisation in high-order iterative learning control,” *Automatica*, vol. 42, no. 2, pp. 287–294, 2006.

[14] N. Amann, D. Owens, and E. Rogers, “Iterative learning control using optimal feedback and feedforward actions,” *International Journal of Control*, vol. 65, no. 2, pp. 277–293, 1996.

[15] N. Amann, “Optimal algorithms for iterative learning control,” Ph.D. dissertation, University of Exeter, UK, 1996.

[16] N. Amann, D. Owens, and E. Rogers, “Iterative learning control for discrete-time systems with exponential rate of convergence,” *IEE Proceedings: Control Theory and Applications*, vol. 143, no. 2, pp. 217–224, 1996.

[17] ——, “Predictive optimal iterative learning control,” *International Journal of Control*, vol. 69, no. 2, pp. 203–226, 1998.

[18] B. Chu, Z. Cai, D. H. Owens, E. Rogers, C. T. Freeman, and P. L. Lewin, “Experimental verification of accelerated norm-optimal iterative learning control,” in *UKACC International Conference on Control*. Coventry, UK, 2010, pp. 211–216.

[19] B. Chu and D. H. Owens, “Accelerated norm-optimal iterative learning control algorithms using successive projection,” *International Journal of Control*, vol. 82, no. 8, pp. 1469–1484, 2009.

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# Control Allocation for Fault Tolerant Control of a VTOL Octorotor

Aryeh Marks
Department of
Aerospace Engineering
Cranfield University
Bedfordshire, U.K.
MK43 0AL
Email:a.marks@cranfield.ac.uk

James F Whidborne
Department of
Aerospace Engineering
Cranfield University
Bedfordshire, U.K.
MK43 0AL
Email:j.f.whidborne@cranfield.ac.uk

Ikuo Yamamoto
Graduate School of
Environmental Engineering
The University of Kitakyushu
Fukuoka
Japan
Email:yamamoto@env.kitakyu-u.ac.jp

*Abstract*—**For the fault tolerant control of an eight-rotor VTOL Unmanned Air Vehicle (UAV), a control allocation scheme is proposed. The eight-rotor configuration provides actuator redundancy to ensure safe operation under rotor/motor failures. A PD controller is used to generate total thrust and moment demands. A cascade inverse method of control allocation is proposed to allocate the controller commands to the actuators whilst ensuring that actuator saturation does not occur. If the vehicle is subjected to rotor failures, the scheme re-allocates the commands to maintain the vehicle stability and performance. Until actuator saturation occurs the response of the vehicle is the same when operating with all motors or fewer. The response of the vehicle to several combinations of complete actuator failures is investigated by simulation and it is shown that the proposed method is able to maintain control after failure of up to four actuators. The controller is invarient and the vehicle response to commands is identical until motor saturation occurs.**

*Index Terms*—**Control Allocation, VTOL UAV, Octorotor, Fault Tolerance**

## I. Introduction

Unmanned Air Vehicles (UAVs) capable of Vertical Take-off and Landing (VTOL) operations can provide many advantages over conventional manned aircraft and UAVs which are not capable of such flight. They give mission flexibility in that a runway is not needed for launch and recovery, give a stable platform for capturing images due to their hover capability, are capable of high agility maneuvers such as vertical drops [1] and allow for operations in harsh and hostile environments since they do not put a human operator at risk. One major drawback with using conventional quadrotor vehicles such as OS4 [2], STARMAC [3], Qball-X4 [4] and tri-rotor vehicles [5] is that there is no effector redundancy. If a rotor fails completely then control is lost and the vehicle may crash. This is an unacceptable scenario when operating over populated urban areas.

It is fairly straightforward to demonstrate that complete loss of a rotor for a quadrotor results in a vehicle that is not fully controllable. However, with a partial failure in one rotor, a quadrotor is still controllable. Fault detection and recovery schemes have been investigated for a 50% thrust reduction in one rotor by [6] and, using a sliding mode method, by [7]. Other methods for partial failures can be found in [4], [8]. Fault tolerant control for a quadrotor subject to sensor failures has been investigated by [9] and by [10]. Even though a total loss of a rotor results in an uncontrollable vehicle, it has been shown that partial control can be maintained [11] with a loss of yaw control.

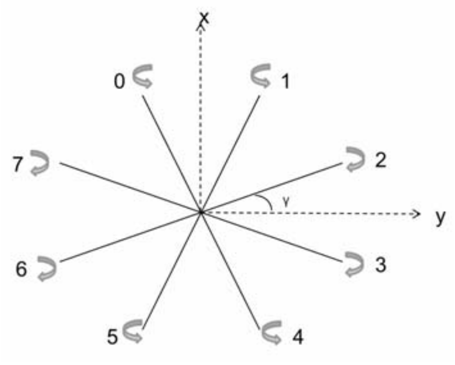


Fig. 1. Octorotor Schematic Layout

TABLE I
Vehicle Parameters

| | |
|---|---|
| Thrust factor, $b$ | $3.13 \times 10^{-5}$ |
| Drag factor, $d$ | $7.5 \times 10^{-7}$ |
| Inertia $I_x, I_y$ | $7.5 \times 10^{-3}$kg m$^2$ |
| Inertia $I_z$ | $1.3 \times 10^{-2}$kg m$^2$ |
| Rotor Inertia, $J$ | $6 \times 10^{-5}$kg m$^2$ |
| Length of arm, $l$ | 0.4m |
| Vehicle mass, $m$ | 1.2kg |
| $\gamma$ | $22.5°$ |

The octorotor has been proposed as a solution to the problem of safe operation of quadrotor-like UAVs [12], [13]. A schematic of the vehicle is shown in Figure 1 with parameters shown in table I. With eight independently controllable rotors the vehicle has in-built hardware redundancy. Combinations of various rotor thrusts will provide moments causing the vehicle to roll around the x axis and pitch around the y axis. Yaw control is achieved by varying the thrust from the clockwise and counter-clockwise rotors whilst keeping a constant overall thrust value. This generates an imbalance in the gyroscopic drag causing the vehicle to yaw around the z axis. Furthermore, if the thrust of any single rotor is changed then it will generate a rolling, pitching and yawing moment. This is a property that can be exploited when fewer than eight rotors are utilized.

It is possible to use various combinations of rotors to generate moments across the body. The allocation and mixing of the thrust demands to achieve a desired objective is the control allocation problem. Various methods for linear control allocation have been proposed including explicit ganging [14], rule based systems [15, pp. 89-106] which have switches in the control laws depending on the failure scenario, daisy chaining [16] which divides the actuators into sets that can perform the same task and are then ranked according to preference for usage and effectiveness until the requirements are met or the maximum performance from the set is reached. Online optimization methods have been proposed, see [17] for a review. The most common approach uses constrained quadratic programming (e.g. [18], [19], [20], [21], [22]) and this results in efficient solutions. A Redistributed Pseudo Inverse (RPI) method [21] is used in this paper. This method can increase the possibility of reaching an optimal solution to an inverse problem and allows for actuator saturation to be considered. A comparison of control allocation methods with control effectiveness uncertainties [23] has shown that the RPI method leads to low errors and high performance.

This paper proposes the use of controller reallocation as a means of obtaining fault tolerant control of the octorotor vehicle subject to failures in one or more rotors. Stability and performance of the vehicle is maintained by means of an RPI control reallocation scheme. The performance of the vehicle subject to multiple rotor failures is shown by simulation. The authors believe this is the first work to show full controllability of all states for a VTOL UAV after rotor failures.

In Section II the dynamics of the octorotor are presented. The essential dynamics are well known [24], [25], [26] but this work focusses on the application to a vehicle with eight rotors rather than the four found in other work [2], [3]. The controller tasked with stabilizing the body angles and global altitude is developed in Section III. A description of rotor saturation and control allocation via the RPI method is given in Section IV. Section V shows results from numerical simulations showing how the control re-allocation can control the vehicle with hardware failures.

## II. Octorotor Dynamics

### A. Dynamics Model

The dynamics of small VTOL UAVs are well developed. Here, the Newton-Euler approach is used [24], [25], [26] with the following assumptions:

- the structure is rigid and symmetric,
- the propellers are rigid,
- the thrust and the drag are proportional to the square of the speed of the rotor,
- ground effect is neglected,
- the inertia matrix is diagonal,
- the rotor Coriolis force and wind forces are not included,
- and the motor dynamics are ignored.

The state variables used in this analysis are:

$$X = [U \; V \; W \; P \; Q \; R \; x \; y \; z \; \phi \; \theta \; \psi]^T \tag{1}$$

where $U, V, W$ are the body-centric velocities of the vehicle, $P, Q, R$ are the rotation rates, $x, y, z$ describe the global position of the vehicle in the inertial frame and $\phi, \theta, \psi$ are the Euler angles. Consider a body-fixed frame with the x,y, and z axes originating at the center of mass of the vehicle. An inertial frame is fixed to the Earth and has axes in the conventional North-East-Down arrangement. The orthogonal rotation matrix $S_b$ converts from a body-fixed coordinate system to the Earth-fixed coordinate system with the assumptions that

- the Earth is flat and stationary and
- the center of gravity lies at the origin of the body axis reference frame

and is given by

$$S_b = \begin{bmatrix} c\theta c\psi & c\theta s\psi & -s\theta \\ s\phi s\theta c\psi - c\phi s\psi & c\phi c\psi + s\phi s\theta s\psi & c\theta s\phi \\ c\phi s\theta c\psi + s\phi s\psi & c\phi s\theta s\psi - s\phi c\psi & c\theta c\phi \end{bmatrix} \tag{2}$$

where $s\phi = \sin\phi$, $c\phi = \cos\phi$ etc. This notation is used throughout the paper.

The total forces and moments in the body axis are given by

$$F_{net} = \frac{d}{dt}[m\mathbf{V}] + \omega' \times [m\mathbf{V}] \tag{3}$$

$$M_{net} = \frac{d}{dt}[I\omega'] + \omega' \times [I\omega'] \tag{4}$$

where $\mathbf{V}$ is the vector of linear velocities, $\omega'$ is the vector of angular velocities, $I$ is the inertia matrix and $m$ is the mass of the vehicle. The gravitational force $F_g$ is

$$F_g = m \; S_b \begin{bmatrix} 0 \\ 0 \\ g \end{bmatrix} = mg \begin{bmatrix} -s\theta \\ c\theta s\phi \\ c\theta c\phi \end{bmatrix} \tag{5}$$

where $g$ is the acceleration due to gravity. The total force $F_{net}$ is the force of gravity and the forces generated through the rotors, $F_p$,

$$F_{net} = F_g + F_p, \tag{6}$$

which from (3) gives

$$\begin{bmatrix} \dot{U} \\ \dot{V} \\ \dot{W} \end{bmatrix} = \frac{1}{m}\begin{bmatrix} F_{px} \\ F_{py} \\ F_{pz} \end{bmatrix} + g \begin{bmatrix} -s\theta \\ c\theta s\phi \\ c\theta c\phi \end{bmatrix} - \begin{bmatrix} QW - RV \\ RU - PW \\ PV - QU \end{bmatrix} \tag{7}$$

From (4), the total moments $M_{net}$ acting on the vehicle are

$$M_{net} = \begin{bmatrix} M_x \\ M_y \\ M_z \end{bmatrix} = \begin{bmatrix} I_x & 0 & 0 \\ 0 & I_y & 0 \\ 0 & 0 & I_z \end{bmatrix} \begin{bmatrix} \dot{P} \\ \dot{Q} \\ \dot{R} \end{bmatrix} + \begin{bmatrix} P \\ Q \\ R \end{bmatrix} \times \begin{bmatrix} I_x & 0 & 0 \\ 0 & I_y & 0 \\ 0 & 0 & I_z \end{bmatrix} \begin{bmatrix} P \\ Q \\ R \end{bmatrix} \tag{8}$$

Rearranging in terms of the state variable derivatives gives

$$\begin{bmatrix}\dot{P}\\ \dot{Q}\\ \dot{R}\end{bmatrix} = \begin{bmatrix}M_x/I_x\\ M_y/I_y\\ M_z/I_z\end{bmatrix} - \begin{bmatrix}((I_z-I_y)/(I_x))\,QR\\ ((I_x-I_z)/(I_y))\,RP\\ ((I_y-I_x)/(I_z))\,PQ\end{bmatrix} \tag{9}$$

The rotation matrix, $S_b$, from (2) is used to express the movement of the vehicle in the global axes once the body-centric velocities are known:

$$\begin{aligned}\begin{bmatrix}\dot{x}\\ \dot{y}\\ \dot{z}\end{bmatrix} &= S_b^T \begin{bmatrix}U\\ V\\ W\end{bmatrix}\\ &= \begin{bmatrix}c\psi c\theta & c\psi s\theta s\phi - s\psi c\phi & c\psi s\theta c\phi + s\psi s\phi\\ s\psi c\theta & s\psi s\theta s\phi + c\psi c\phi & s\psi s\theta c\phi - c\psi s\phi\\ -s\theta & c\theta s\phi & c\theta c\phi\end{bmatrix} \begin{bmatrix}U\\ V\\ W\end{bmatrix}\end{aligned} \tag{10}$$

The flight path is found by integrating (10). It contains the body-centric Euler angles and these are related to the global body angles through

$$\begin{bmatrix}P\\ Q\\ R\end{bmatrix} = \begin{bmatrix}1 & 0 & -s\theta\\ 0 & c\phi & s\phi c\theta\\ 0 & -s\phi & c\theta c\phi\end{bmatrix} \begin{bmatrix}\dot{\phi}\\ \dot{\theta}\\ \dot{\psi}\end{bmatrix} = S_e^{-1} \begin{bmatrix}\dot{\phi}\\ \dot{\theta}\\ \dot{\psi}\end{bmatrix}$$

giving

$$\begin{bmatrix}\dot{\phi}\\ \dot{\theta}\\ \dot{\psi}\end{bmatrix} = S_e \begin{bmatrix}P\\ Q\\ R\end{bmatrix} \tag{11}$$

where

$$S_e = \begin{bmatrix}1 & t\theta s\phi & t\theta c\phi\\ 0 & c\phi & -s\phi\\ 0 & s\phi/c\theta & c\phi/c\theta\end{bmatrix} \tag{12}$$

### B. State Space Model

The general state space model $\dot{X} = f(X, \Gamma)$ is obtained from (7), (9), (10), (11) and (17) with state variables given by (1) and control given by

$$\Gamma = \begin{bmatrix}F_{pz}\\ M_p\end{bmatrix} \tag{13}$$

where $M_p = M_{net}$. This gives the state equation

$$\frac{d}{dt}\begin{bmatrix}U\\ V\\ W\\ P\\ Q\\ R\\ x\\ y\\ z\\ \phi\\ \theta\\ \psi\end{bmatrix} = \begin{bmatrix}-g\ s\theta - (QW - RV)\\ g\ c\theta s\phi - (RU - PW)\\ (-c\theta c\phi F_{pz}/m) + g\ c\theta c\phi - (PV - QU)\\ (M_x/I_x) - ((I_z - I_y)/I_x)\,QR\\ (M_y/I_y) - ((I_x - I_z)/I_y)\,RP\\ (M_z/I_z) - ((I_y - I_x)/I_z)\,PQ\\ (c\psi c\theta)U + (c\psi s\theta s\phi - s\psi c\phi)V + (c\psi s\theta c\phi + s\psi s\phi)W\\ (s\psi c\theta)U + (s\psi s\theta s\phi + c\psi c\phi)V + (s\psi s\theta c\phi - c\psi s\phi)W\\ -s\theta U + (c\theta s\phi)V + (c\theta c\phi)W\\ P + (t\theta s\phi)Q + (t\theta c\phi)R\\ c\phi Q - s\phi R\\ (s\phi/c\psi)Q + (c\phi/c\theta)\,R\end{bmatrix} \tag{14}$$

### C. Actuator Model

The total force provided by the rotors in the body frame is

$$F_{pz} = \sum_{i=0}^{7} b\ \Omega_i^2 \tag{15}$$

where $b$ is the term relating the rotor thrust to the squared rotor speed $\Omega_i^2$. The rotor axes are mounted vertically on the vehicle so

$$F_{px} = F_{py} = 0 \tag{16}$$

The moments, $M_x, M_y, M_z$, are generated by the differences in the thrusts of the eight rotors. The relationship between the control $\Gamma$ and the rotor speeds is given by

$$\Gamma = \Lambda\Sigma \tag{17}$$

where $\Sigma$ is the vector containing the squared rotor speeds

$$\Sigma = \begin{bmatrix}\Omega_0^2 & \Omega_1^2 & \Omega_2^2 & \Omega_3^2 & \Omega_4^2 & \Omega_5^2 & \Omega_6^2 & \Omega_7^2\end{bmatrix}^T, \tag{18}$$

and $\Lambda$ is the control allocation matrix

$$\Lambda = \begin{bmatrix}b & b & b & b & b & b & b & b\\ b_s & -b_s & -b_c & -b_c & -b_s & b_s & b_c & b_c\\ -b_c & -b_c & -b_s & b_s & b_c & b_c & b_s & -b_s\\ d & d & -d & -d & d & d & -d & -d\end{bmatrix}, \tag{19}$$

where $b_s = bl\sin\gamma$, $b_c = bl\cos\gamma$ and where $\gamma$ denotes the angle between the arms of the vehicle and the major axis lines as shown in Figure 1, $l$ denotes the arm length from the center of the vehicle to the rotor axis, the factor $b$ relates the squared rotor speed $\Omega_i^2$ to the thrust and the drag factor $d$ relates the gyroscopic drag to the squared rotor speed.

## III. CONTROLLER

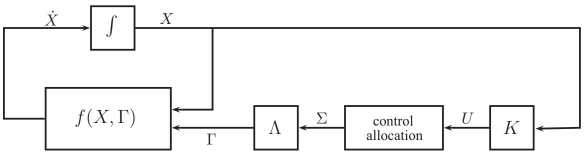


Fig. 2. System Block Diagram

Figure 2 shows the block diagram of the system which is assumed to provide perfect control allocation, that is $\Gamma = U$. The number of rotors and the allocation of the control demands, $\Sigma$, are not considered in the control block where the output, $U$, is simply a moment demand. This means that any maneuvers are completed with the same performance regardless to the number of functional rotors. The operator should not notice any performance change with rotor/motor failures unless all actuators are saturated.

Even though the vehicle has eight physical actuators the system is under actuated in that the actuators can only directly affect four of the six degrees of freedom. For this reason only four control demands can be performed. The control $\Gamma$ (17) contains the total thrust and three moment demands and corresponds to commands $U_i$, $i = 1, 2, 3, 4$.

To design the controller, several further assumptions are made on the model presented in Section II. The vehicle is assumed to operate near hover (non-acrobatic flight), hence the cross coupling terms are ignored in (9) and state derivatives $\dot{P}, \dot{Q}, \dot{R}$ are assumed to be proportional to the controller demands:

$$\begin{bmatrix} \dot{P} \\ \dot{Q} \\ \dot{R} \end{bmatrix} = \text{diag}\left(\frac{1}{I_x}, \frac{1}{I_y}, \frac{1}{I_z}\right) \begin{bmatrix} U_2 \\ U_3 \\ U_4 \end{bmatrix} \tag{20}$$

Similar approximations are made to (10) and (11) to obtain

$$\dot{z} = W \tag{21}$$
$$\dot{\phi} = P \tag{22}$$
$$\dot{\theta} = Q \tag{23}$$
$$\dot{\psi} = R \tag{24}$$

The total thrust is calculated in the Earth-fixed frame meaning the thrust from the rotors $F_{pz}$ must be multiplied by the appropriate factors in the $S_b$ matrix (2). Using (3) and (7) the total thrust demand is known. Based on these approximations, a PD scheme is used for the controller block, $K$, with the individual commands:

$$U_1 = -\frac{m}{c\phi c\theta}(K_{pz}(z_d - z) - K_{dz}(\dot{z}) + g) \tag{25}$$
$$U_2 = (K_{p\phi}(\phi_d - \phi) - K_{d\phi}(\dot{\phi})) \tag{26}$$
$$U_3 = (K_{p\theta}(\theta_d - \theta) - K_{d\theta}(\dot{\theta})) \tag{27}$$
$$U_4 = (K_{p\psi}(\psi_d - \psi) - K_{d\psi}(\dot{\psi})) \tag{28}$$

The gain values were tuned by trial and error and are identical for the roll and pitch controllers (26) and (27) since the vehicle has planes of symmetry along the x and y body axes and so it is possible to equate the pitch and roll responses to a rotor thrust. The gain values for the yaw controller, (28), are set lower than for the roll and pitch since the gyroscopic drag generated by the rotor turning is lower than the moment created when a rotor speed is increased. This means that the yaw response is more sluggish than the roll and pitch response.

## IV. Rotor Saturation and Linear Control Allocation

### A. Rotor Saturation

The allowable values of the rotor speeds $\Omega_i(t)$ are limited between absolute lower and upper bounds $\underline{\Omega}_i$, $\overline{\Omega}_i$ such that

$$\underline{\Omega}_i \leq \Omega_i(t) \leq \overline{\Omega}_i \tag{29}$$

holds for all $t$. There is a maximum thrust that can be generated by each motor-propeller combination due to the constraint on the maximum rotational speed, as well as a minimum thrust due to the lowest rotation speed of the rotor. Furthermore, it is assumed that the motors cannot turn backwards so a negative thrust cannot be generated. Hence, for the octorotor, $\underline{\Omega}_i \geq 0$. Test bed modeling can reveal the limit to the rate at which the motors can respond but in this paper no motor dynamics are modeled and it is assumed that the rotors respond instantaneously. Hence no constraint on the control rates is imposed. From (18), (29) can be rewritten as

$$\underline{\Sigma}_i \leq \Sigma_i(t) \leq \overline{\Sigma}_i \tag{30}$$

Repeating (17), the mapping of the generation of moments to the thrusts from the rotors is

$$\Gamma = \Lambda\Sigma \tag{31}$$

where for the octorotor, with all rotors operational, the dimension of $\Lambda$ is $4 \times 8$ and the full control allocation matrix is given by (19). It should be noted that this differs from the control mapping of a quadrotor where the dimension of $\Lambda$ is $4 \times 4$ [27] and a simple matrix inversion is used to determine the individual rotor thrust values based on the control demands.

For the octorotor a pseudo-inverse method is used. Ideally the thrust demands are shared between all eight rotors. This ensures that no single rotor is close to its maximum threshold which for these simulations was set at 700 rad/s. If an unattainable command is demanded by the controller then the allocation block will not pass it forward to the dynamics block. This is achieved using the redistributed pseudo inverse (RPI) method which is outlined in Section IV-B. This method allows for a control reallocation following a rotor failure. The dimension of the matrix $\Lambda$ is reduced to $4 \times (8 - q)$ where $q$ is the number of failed rotors. The outputs of the individual rotors are then capable of providing the required thrust until their maximum is reached. At this point they are saturated and no more thrust can be generated. In such a state the vehicle may not be controllable.

### B. Redistributed Pseudo Inverse (RPI) Method

The redistributed pseudo inverse method extends the pseudo inverse method by explicitly accounting for actuator saturations. The method is originally attributable to [21], the method we use here is based on the description in [14]. A similar approach, called the Cascaded Generalized Inverse method is proposed by [28]. The process is iterative in that a succession of pseudo inverse solutions are calculated with position saturated control effectors removed from subsequent pseudo inverse solutions. The algorithm is

Step 1. Set $\tilde{\Lambda} = \Lambda$ and $c_i = 0$ for all $i$.

Step 2. Solve the modified pseudo inverse control allocation problem:

$$\Sigma = -c + \tilde{\Lambda}^{+}[\Gamma + \Lambda c] \tag{32}$$

where $\cdot^{+}$ denotes the Moore-Penrose pseudo inverse.

Step 3. If $\underline{\Sigma}_i < \Sigma_i < \overline{\Sigma}_i$ for all $i$, then end. Otherwise,

- for all $i$ such that $\Sigma_i \leq \underline{\Sigma}_i$, set $c_i = -\underline{\Sigma}_i$ and remove the $i$th column of $\tilde{\Lambda}$ and
- for all $i$ such that $\Sigma_i \geq \overline{\Sigma}_i$, set $c_i = -\overline{\Sigma}_i$ and remove the $i$th column of $\tilde{\Lambda}$

and return to Step 2.

The scheme is simply adapted for a control reallocation scheme in the event of an actuator failure by augmenting Step 1. Following a total failure in the $j$th actuator, remove the $j$th column of $\tilde{\Lambda}$ and remove $c_j$. Then proceed as above.

## V. Simulations

In order to test the performance of the redistributed pseudo inverse reallocation method combined with the PD controller, simulations of a number of scenarios were carried out using MATLAB/Simulink. All scenarios began with the vehicle in a hover with all of the state derivatives equal to zero, and the aim is to maintain hover despite multiple rotor failures. The failure condition was initiated at a time of $t = 3$s via a switch in the simulation. It is assumed that the time taken for rotor fault detection and controller reallocation is 0.5s.

Four scenarios were investigated:

1) Failure in rotor 7.
2) Failure in rotor 0 and 7.
3) Failure in rotor 1, 5 and 7.
4) Failure in rotor 1, 3, 5 and 7.

The flight path of the vehicle is described by its altitude $(z)$ and three Euler angles $(\phi, \theta, \psi)$. After the failure the vehicle should recover level flight and regain any lost altitude.

### A. Altitude Recovery

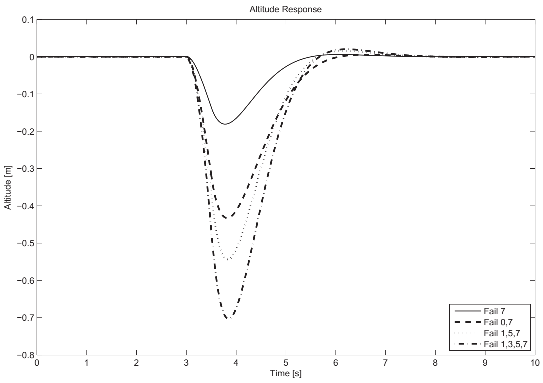


Fig. 3. Altitude Response

Figure 3 shows that as the number of failed rotors increases the altitude drop increases. This is as expected since the total thrust available drops. The vehicle recovers after the reconfiguration to regain the initial altitude with a similar performance after all rotor failures.

### B. Roll Recovery

Figure 4 shows that failure in the rotor caused the vehicle to roll towards the failed rotors. The largest variation from level attitude was with two failed rotors that were next to each other. This generated a large adverse rolling moment before the vehicle recovered to level hover. When four rotors failed, because of the symmetry, no adverse rolling moment was generated and the failed-rotor vehicle acted like a quadrotor with only altitude disturbance.

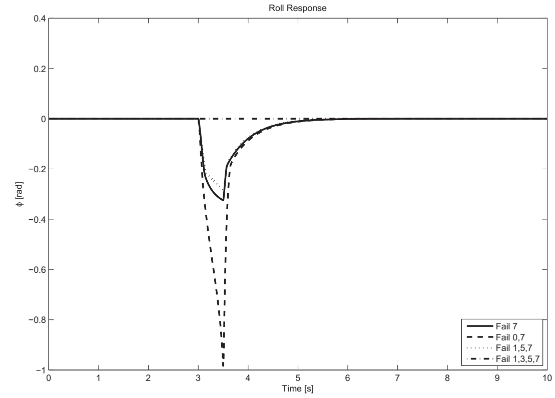


Fig. 4. Roll Response

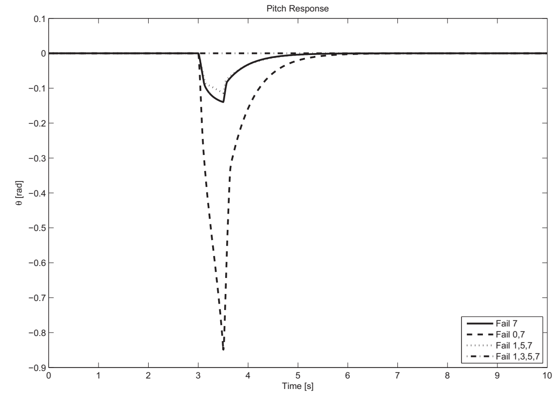


Fig. 5. Pitch Response

### C. Pitch Recovery

Figure 5 shows that the pitch responses are similar to the roll response due to the symmetry over the x and y body axes. The differences in response after the failures were due to the physical location of the failed rotors. The response after reconfiguration is similar to the roll recovery and the vehicle regains hover. Again, with four rotor failures the vehicle only deviates in altitude.

### D. Yaw Recovery

Figure 6 shows that the yaw moment generated from a failed rotor caused the vehicle to yaw slightly. But due to the highly coupled nature of the octorotor dynamics (9) the roll and pitch lead to a large adverse yaw disturbance.

The large roll and pitch generated with a failure in rotors 0 and 7 leads to a large divergence in yaw. The response of the vehicle once reconfigured is not the same as the roll and pitch responses due to the different gains chosen for this controller as described in Section III

## VI. Conclusion

The paper proposes the use of a redistributed pseudo inverse method of control reallocation for the fault tolerant control of

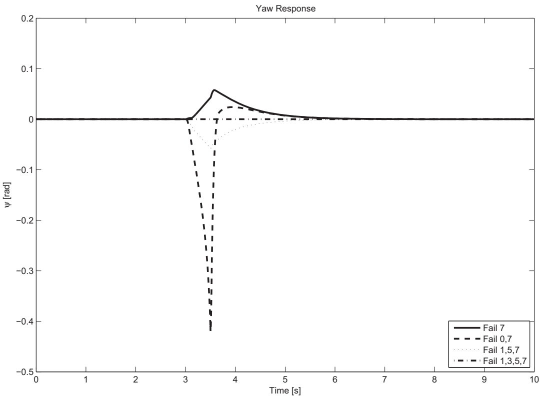


Fig. 6. Yaw Response

an octorotor VTOL aircraft. Unlike the quadrotor, the vehicle stability and performance is resilient to single rotor failure. Furthermore multiple rotor failures can be tolerated; control can be maintained for up to four rotor failures. Four scenarios were investigated by simulation and in all of them the vehicle retained enough control authority to recover from upset angles up to 60°. Note that the proposed method depends on the availability of a fault detection system, this aspect has not been addressed in this paper.

Ongoing theoretical work is investigating the effect of redistributing the rotors so they turn clockwise, counter-clockwise alternating around the vehicle and finding an optimum configuration which will maximize the resilience of the system to multiple rotor failure. In future work, the method will be flight tested.

## References

[1] B. Michini, J. Redding, N. Ure, M. Cutler, and J. How, "Design and flight testing of an autonomous variable-pitch quadrotor," in *2011 IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China, May 2011, pp. 2978 –2979.

[2] S. Bouabdallah, A. Noth, and R. Siegwart, "PID vs LQ control techniques applied to an indoor micro quadrotor," in *Proceedings EEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004).*, vol. 3, Sendai, Japan, Sep. 2004, pp. 2451 – 2456.

[3] S. Waslander, G. Hoffmann, J. S. Jang, and C. Tomlin, "Multi-agent quadrotor testbed control design: integral sliding mode vs. reinforcement learning," in *Proceedings 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2005)*, Edmonton, Canada, Aug. 2005, pp. 3712 – 3717.

[4] Q.-L. Zhou, Y. Zhang, C.-A. Rabbath, and D. Theilliol, "Design of feedback linearization control and reconfigurable control allocation with application to a quadrotor UAV," in *Conference on Control and Fault-Tolerant Systems (SysTol)*, Oct. 2010, pp. 371 –376.

[5] S. Salazar-Cruz and R. Lozano, "Stabilization and nonlinear control for a novel trirotor mini-aircraft," in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA 2005)*, Barcelona, Spain, Apr. 2005, pp. 2612 – 2617.

[6] M. Ranjbaran and K. Khorasani, "Fault recovery of an under-actuated quadrotor aerial vehicle," in *49th IEEE Conference on Decision and Control (CDC)*, Dec. 2010, pp. 4385 –4392.

[7] F. Sharifi, M. Mirzaei, B. Gordon, and Y. Zhang, "Fault tolerant control of a quadrotor uav using sliding mode control," in *2010 Conference on Control and Fault-Tolerant Systems (SysTol)*, Oct. 2010, pp. 239 –244.

[8] Y. Zhang and A. Chamseddine, *Automatic Flight Control Systems - Latest Developments*. InTech, 2012, ch. 5. Fault Tolerant Flight Control Techniques with Application to a Quadrotor UAV Testbed.

[9] C. Berbra, S. Lesecq, and J. Martinez, "A multi-observer switching strategy for fault-tolerant control of a quadrotor helicopter," in *16th Mediterranean Conference on Control and Automation*, Ajaccio-Corsica, France, Jun. 2008, pp. 1094 –1099.

[10] A. Freddi, S. Longhi, and A. Monteriù, "A model-based fault diagnosis system for a mini-quadrotor," in *7th workshop on Advanced Control and Diagnosis*, Zielona Góra, Poland, Nov. 2009, pp. 2055 –2060.

[11] A. Freddi, A. Lanzon, and S. Longhi, "A feedback linearization approach to fault tolerance in quadrotor vehicles," in *Proceedings of the 18th IFAC World Congress*, Milano, Italy, Sep. 2011, pp. 5413 – 5418.

[12] T. Matsuzaki, I. Yamamoto, N. Inagawa, T. Nakamura, W. Batty, and J. F. Whidborne, "Development of unmanned flying observation robot with real time video transmission system," in *Proc. World Automation Congress (WAC 2010)*, Kobe, Japan, Sep. 2010, p. 5 pages, article number 5665710.

[13] V. G. Adîr, A. M. Stoica, A. Marks, and J. F. Whidborne, "Modelling, stabilization and single motor failure recovery of a 4Y octorotor," in *Proc. 13th IASTED International Conference on Intelligent Systems and Control (ISC 2011)*, Cambridge, U.K., Jul. 2011, pp. 82–87.

[14] M. W. Oppenheimer, D. B. Doman, and M. A. Bolender, "Control allocation for over-actuated systems," in *Proc. 14th Mediterranean Conference on Control and Automation (MED 06. )*, Ancona, Italy, Jun. 2006, paper no. FEA4-3.

[15] G. J. Ducard, *Fault-tolerant Flight Control and Guidance Systems – Practical Methods for Small Unmanned Aerial Vehicles*, ser. Advances in Industrial Control. Springer London, 2009. [Online]. Available: http://dx.doi.org/10.1007/978-1-84882-561-1_5

[16] J. Berg, K. Hammett, C. Schwartz, and S. Banda, "An analysis of the destabilizing effect of daisy chained rate-limited actuators," *IEEE Transactions on Control Systems Technology*, vol. 4, no. 2, pp. 171 – 176, Mar. 1996.

[17] M. Bodson, "Evaluation of optimization methods for control allocation," *J. Guid. Control Dyn.*, vol. 25, pp. 703–711, 2002.

[18] M. Bodson and W. Pohlchuk, "Command limiting in reconfigurable flight control," *J. Guid. Control Dyn.*, vol. 21, pp. 639–646, 1998.

[19] D. Enns, "Control allocation approaches," in *Proc AIAA Guid., Nav. and Control Conf.*, Boston, MA, 1998, pp. 98–108.

[20] J. J. Burken, P. Lu, Z. L. Wu, and C. Bahm, "Two reconfigurable flight-control design methods: Robust servomechanism and control allocation," *J. Guid. Control Dyn.*, vol. 24, pp. 482–493, 2001.

[21] J. Virnig and D. Bodden, "Multivariable control allocation and control law conditioning when control effectors limit," in *Proc. AIAA Guid., Nav. and Control Conf.*, Scottsdale, AZ, Aug. 1994, pp. 572–582.

[22] O. Harkegard, "Dynamic control allocation using constrained quadratic programming," *J. Guid. Control Dyn.*, vol. 27, pp. 1028–1034, 2004.

[23] J. Ma, P. Li, W. Li, and Z. Zheng, "Performance comparison of control allocation for aircraft with control effectiveness uncertainties," in *Proc. Asia Simulation Conference - 7th International Conference on System Simulation and Scientific Computing (ICSC 2008)*, Beijing, P.R. China, Oct. 2008, pp. 164–169.

[24] S. Bouabdallah, P. Murrieri, and R. Siegwart, "Design and control of an indoor micro quadrotor," in *Robotics and Automation, 2004. Proceedings. ICRA '04. 2004 IEEE International Conference on*, vol. 5, May 2004, pp. 4393 – 4398.

[25] A. Mian and W. Daobo, "Nonlinear flight control strategy for an underactuated quadrotor aerial robot," in *IEEE International Conference on Networking, Sensing and Control (ICNSC 2008)*, Apr. 2008, pp. 938 –942.

[26] Y. Yu, C. Jiang, and H. Wu, "Backstepping control of each channel for a quadrotor aerial robot," in *International Conference on Computer, Mechatronics, Control and Electronic Engineering (CMCE)*, vol. 3, Aug. 2010, pp. 403 –407.

[27] S. Bouabdallah and R. Siegwart, "Backstepping and sliding-mode techniques applied to an indoor micro quadrotor," in *Proceedings IEEE International Conference on Robotics and Automation (ICRA 2005)*, Apr. 2005, pp. 2247 – 2252.

[28] K. A. Bordignon, "Constrained control allocation for systems with redundant control effectors," Ph.D. dissertation, Virginia Polytechnic Institute, VA, 1996.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Attitude Control of VTOL-UAVs

Maryam Heidarian[†] and Attaullah Y. Memon[‡]

*Abstract*— **This paper presents a novel control approach to obtain asymptotic attitude stability of a quadrotor as a representative of Planar Vertical Take Off and Landing (PVTOL) Unmanned Aerial Vehicles (UAVs). The considered quadrotor is a symmetric VTOL-UAV with four rigid mono-directional propellers, which has been modeled based on quaternion representation with taking Coriolis and gyroscopic torques into account. In the proposed approach, two nearly equivalent control laws (model independent as well as model dependent) have been used to obtain exponential stability of attitude angles and asymptotic stability of attitude angular velocity of the quadrotor UAV. The proposed approach also presents how the attitude parameters i.e. attitude angles and attitude angular velocity can be quickly regulated to their desired values as required.**

*Index Terms*— **Attitude Control, Velocity Control, VTOL, UAV, Quadrotor.**

## I. INTRODUCTION

It is more than two decades since the first VTOL-UAVs were fully demonstrated in practical research works. VTOL-UAVs constitute the class of Unmanned Aerial Vehicles which have the ability of Vertical Take Off and Landing and thus have the capability of fast target acquisition. Due to this, their applications have found a growing interest in performing certain tasks which require high maneuverability and robustness with respect to unknown external disturbances. These UAVs can be used both as individual vehicles and in a team formation of multiple vehicles [1]. Furthermore, use of VTOL-UAVs has been envisaged in a variety of applications, e.g. in environmental protection, intervention in hostile sites, natural risk management, remote inspections, rescue missions, agriculture and, commercial video production. More recently, small quadrotor UAVs have attracted considerable attention by researchers, due to its less complicated mechanical design and maintenance aspects in comparison to helicopters. A quadrotor is essentially a helicopter which has four propellers in cross configuration.

In order to effectively control a VTOL-UAV, a composite control scheme comprising of two different controllers is required, namely: the attitude controller and the position controller. Design of these two controllers constitutes challenging tasks and the same have been addressed separately in the literature. In [2], a dynamical model of quadrotor based on quaternion representation has been derived from Newton-Euler equations. Hamel et al., in [3], identify dynamics of the vehicle beyond the basic nonlinear equations of motion, with gyroscopic torque and Coriolis terms. Based on this model, Tayebi and McGilvray [4], have represented a model independent $PD$ controller with asymptotic stability and a model dependent $PD^2$ controller with exponential stability. A method to obtain attitude control stabilization of a quadrotor through using backstepping technique and adding saturation functions has been analyzed in [5]. Precise measurement of the angular velocity and the initial orientation are required for attitude stabilization of these vehicles. Due to various uncertainties (related to gyroscope and other effects), there may be some errors in these measurements. Using inertial measurements units (IMU's) information to estimate these required values is one of the possible ways to reduce the errors [6]. In their more recent work [7], Tayebi et al., have considered control designs that do not necessarily require exact knowledge of the angular velocity of the aerial vehicle.

In this paper, we consider two nearly equivalent control laws (model independent as well as model dependent) to obtain exponential stability of attitude angles and asymptotic stability of attitude angular velocity of the quadrotor UAV. The rest of the paper is organized as follows. Section II states the problem formulation and presents the necessary mathematical foundation. Section III presents the control design and main results. The simulation results are presented in Section IV, and finally Section V draws the conclusions.

## II. PROBLEM FORMULATION

A quadrotor simply consists of four lift generating propellers mounted on motors. These motors are located at the lateral sides of a cross shaped frame with an angle of 90 degrees between the arms. Center of mass is placed at the intersection of the line joining rotors 1 and 3 and the line joining rotors 2 and 4, which is middle of the connecting links. Furthermore, the quadrotor is assumed to be symmetric and a basic schematic is depicted in Fig. 1.

Flight control of this vehicle is achieved by varying the angular speed of rotors, $w_i, \imath \in 1, 2, 3, 4$. For example, pitching is possible with increasing (reducing) the speed of the rear motor while reducing (increasing) the speed of the front motor. This means, the pitch torque is a function of the difference $w_1 - w_3$. Similarly, roll movement is obtained by using the lateral motors. Thus, the roll torque will be a function of $w_2 - w_4$. Similarly, the yaw motion is obtained by increasing (reducing) the speed of the front and rear motors together while reducing (increasing) the speed of the lateral motors together. These movements can be accomplished while keeping the total thrust, $T$, constant. Also the vertical movement is generated by increasing the total thrust. The

[†‡] The authors are with the Department of Electronics and Power Engineering, PN Engineering College, National University of Sciences & Technology (NUST), Karachi, Pakistan.
(E-mail: maryamh@pnec.edu.pk, attaullah@pnec.edu.pk)

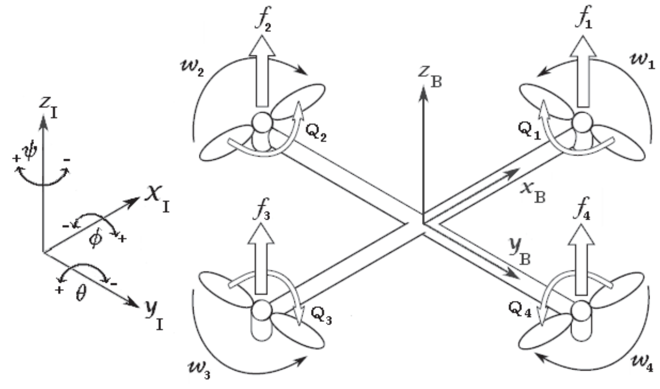


Fig. 1. **Illustration of frames**

main thrust is expressed as:

$$T = \sum_{i=1}^{4} |f_i| \tag{1}$$

$$f_i = b\ w_i^2 z_B \tag{2}$$

Where $f_i$ is the vertically upward lifting force produced by $i$th motor. As we can see in Fig. 1, the rotation direction of two of the rotors is clockwise while the same for the other two is counterclockwise. This is so to balance the movement of the quad rotor and to prevent any yaw drift caused by the unbalanced reactive torques. The reactive torque of $i$th rotor is given by

$$Q_i = l\ w_i^2 \tag{3}$$

Since, each motor turns in a fixed direction, the produced force $f_i$ is always positive. Thus,

$$T = b \sum_{i=1}^{4}\ w_i^2 \tag{4}$$

The constants $l > 0$ and $b > 0$ in the above two equations are dependent on different aerodynamical parameters [8]. The generalized torques (e.g. roll torque $\tau_\phi$, pitch torque $\tau_\theta$, and yaw torque $\tau_\psi$) according to [3] can be represented by

$$\begin{aligned}
\tau_\phi &= b\ d\ (\ w_2^2 -\ w_4^2) &(5)\\
\tau_\theta &= b\ d\ (\ w_1^2 -\ w_3^2) &(6)\\
\tau_\psi &= l(\ w_1^2 +\ w_3^2 -\ w_2^2 -\ w_4^2) &(7)
\end{aligned}$$

in which $d$ is the length of arms between the motors and the center of gravity.

It is noteworthy to mention here that a good working knowledge of the quadrotor dynamical model is essential to improve the performance of the aircraft. Simple vector algebraic laws (e.g. commutativity) cannot be applied to finite rotation vectors of rigid bodies. Due to this, we cannot find the attitude of the aircraft from integrating the angular velocities. Dynamical modeling of the quadrotor requires us to define two reference frames: an inertial frame $I$ defined by set of unit vectors $\{x_I, y_I, z_I\}$ and a body fixed-frame $B$ with orthogonal axes defined by set of unit vectors $\{x_B, y_B, z_B\}$. In order to define the orientation of the aircraft between these two reference frames, one can use Euler angle description, in which a 3 × 3 direction matrix will represent rotation of the aircraft with respect to the body fixed-frame. Euler angle description has an inherent geometric singularity problem. In order to overcome this problem, one can use quaternion representation which defines the rotation of the aircraft with four parameters. The quaternion description is essentially based on Euler's theorem which states that any rotation of an aircraft can be described by a single rotation about a fixed axis.

In what follows, the dynamical model of the quadrotor is obtained via Newton-Euler approach as given in [4]. The basic assumption is that the quadrotor and its propellers are rigid and the external aerodynamic effects (air friction, wind pressure etc.) can be neglected. A simplified model with consideration of Coriolis and gyroscopic torques is given by

$$\begin{aligned}
\dot{q} &= \frac{1}{2}\begin{pmatrix} -(\bar{q})^T \\ s(\bar{q}) + q_0 I \end{pmatrix}\Omega &(8)\\
I_f\dot{\Omega} &= -s(\Omega)I_f\Omega - G_a + \tau_a &(9)\\
I_r\dot{w}_i &= \tau_i - Q_i \qquad i \in 1,2,3,4 &(10)
\end{aligned}$$

where $q$ represents the quaternion equations as given by

$$\begin{aligned}
q &= \begin{pmatrix} q_0 \\ \bar{q} \end{pmatrix}\\
&= \begin{pmatrix}
\cos\frac{\phi}{2}\cos\frac{\theta}{2}\cos\frac{\psi}{2} + \sin\frac{\phi}{2}\sin\frac{\theta}{2}\sin\frac{\psi}{2}\\
\sin\frac{\phi}{2}\cos\frac{\theta}{2}\cos\frac{\psi}{2} - \cos\frac{\phi}{2}\sin\frac{\theta}{2}\sin\frac{\psi}{2}\\
\cos\frac{\phi}{2}\sin\frac{\theta}{2}\cos\frac{\psi}{2} + \sin\frac{\phi}{2}\cos\frac{\theta}{2}\sin\frac{\psi}{2}\\
\cos\frac{\phi}{2}\cos\frac{\theta}{2}\sin\frac{\psi}{2} - \sin\frac{\phi}{2}\sin\frac{\theta}{2}\cos\frac{\psi}{2}
\end{pmatrix} &(11)
\end{aligned}$$

The parameters $\phi$, $\theta$ and $\psi$ respectively represent the roll, pitch and yaw angular displacements about their related axes, and are defined by the following relations [9];

$$\begin{aligned}
\dot{\phi} &= \Omega_\phi + (\Omega_\theta \sin\phi + \Omega_\psi \cos\phi)\tan\theta &(12)\\
\dot{\theta} &= \Omega_\theta \cos\phi - \Omega_\psi \sin\phi &(13)\\
\dot{\psi} &= (\Omega_\theta \sin\phi + \Omega_\psi \cos\phi)\sec\theta &(14)
\end{aligned}$$

in which $\Omega = (\Omega_\phi, \Omega_\theta, \Omega_\psi)^T$ describes the angular velocity of the quadrotor. Furthermore, $I_f \in \mathbb{R}^{3\times3}$ is a symmetric positive-definite constant inertia matrix of the airframe with respect to this frame, and $\tau_a = (\tau_\phi, \tau_\theta, \tau_\varphi)^T$ denotes the control torque. The notation $I_r$ represents the constant moment of inertia of the rotor, and $\tau_i$ , $i \in 1,2,3,4$ represents the rotor torques. The function $s(x)$ in (9) represents a skew-symmetric matrix defined as

$$s(x) = \begin{pmatrix} 0 & -x_3 & x_2 \\ x_3 & 0 & -x_1 \\ -x_2 & x_1 & 0 \end{pmatrix}$$

and is used to define the term $-s(x)I_f\Omega$ in (9), which is due to Coriolis torque. Finally, the term $G_a$ denotes the gyroscopic torque and is defined as

$$G_a = I_r s(\Omega) z_I(-w_1 + w_2 - w_3 + w_4)$$

in which $z_I = (0, 0, 1)^T$.

## III. Control Design and Analysis

The proposed control algorithm to stabilize the attitude of quadrotor follows closely the approach presented in [4], with some technical differences incorporated to seek for an improvement in overall transient performance. We first design the control torque $(\tau_a)$ that stabilizes the attitude of quadrotor dynamically and then, we synthesize a rotor torque $(\tau_i)$ to obtain the control torque $(\tau_a)$ designed earlier, while considering that the real dynamical input to the quadrotor is angular speed of rotors.

### A. Designing the Control Torque $\tau_a$

The dynamical model of the quadrotor, as described in the last section, possesses a cascade structure, in which $\tau_a$ controls $\Omega$ and $\Omega$ controls $q$, i.e. $(\tau_a \to \Omega \to q)$. This means:

$$\dot{\Omega} = g(\Omega, \tau_a)$$

$$\dot{q} = f(q, \Omega)$$

From (8), we have

$$\dot{q} = f(q, \Omega) = F(\Omega)q \tag{15}$$

$$F(\Omega) = \frac{1}{2}\begin{pmatrix} 0 & -\Omega_\phi & -\Omega_\theta & -\Omega_\psi \\ \Omega_\phi & 0 & -\Omega_\psi & \Omega_\theta \\ \Omega_\theta & \Omega_\psi & 0 & -\Omega_\phi \\ \Omega_\psi & -\Omega_\theta & \Omega_\phi & 0 \end{pmatrix} \tag{16}$$

Similarly from (9), we have

$$\dot{\Omega} = g(\Omega, \tau_a) = I_f^{-1}(-s(\Omega)I_f\Omega - G_a + \tau_a) \tag{17}$$

With this in mind, our goal is to find a suitable control law $\tau_a = H(q, \Omega)$. We achieve this objective through the following two steps:

- By finding desired angular velocity $\Omega_d = h(q)$ such that when $\Omega_d$ is given as input to (9), the solution to the nonlinear equation $\dot{q} = f(q, h(q))$ is asymptotically stable.
- By ensuring that the angular velocity $\Omega$ asymptotically tracks the desired angular velocity $\Omega_d$, i.e. $\lim_{t\to\infty}(sup|\Omega - \Omega_d|) = 0$.

*1) Desired Angular Velocity:* The desired angular velocity $\Omega_d$, has to be chosen in such a way that the solution of the nonlinear differential equation $\dot{q} = f(q, \Omega_d)$ converges to its equilibrium point. The equilibrium point, with assuming $0 \leq q_0 \leq 1$ is $q_e = (1, 0, 0, 0)^T$. The quaternion regulation error can be described by

$$\widetilde{q} = q - q_e = (q_0 - 1, q_1, q_2, q_3)^T$$

Comparing $\tilde{q}$ with $\dot{q}$ will give us:

$$\begin{aligned} \dot{\tilde{q}} &= \frac{1}{2}\begin{pmatrix} -(\bar{q})^T \\ s(\bar{q}) + (q_0 - 1)I \end{pmatrix}\Omega \\ &= \frac{1}{2}\begin{pmatrix} -q_1 & -q_2 & -q_3 \\ q_0 - 1 & -q_3 & q_2 \\ q_3 & q_0 - 1 & -q_1 \\ -q_2 & q_1 & q_0 - 1 \end{pmatrix}\Omega \\ &= B(q)\Omega \end{aligned} \tag{18}$$

**Theorem 1.** *Let $\alpha$ be a positive constant and $Q$ is any positive definite symmetric matrix. If the desired angular velocity is given by*

$$\Omega_d = \alpha I B(q)^T Q q \tag{19}$$

*Then, under the stated assumptions and conditions, the overall quaternion system will be exponentially stable. Furthermore, the desired quaternion regulation settling time can be obtained by choosing a suitable value of $\alpha$.*

*Proof:* For simplicity, we will consider $Q = 2I$ in Equation (19) so that

$$\Omega_d = -\alpha\bar{q} \tag{20}$$

Consider the stable unforced system

$$\begin{aligned} \dot{\tilde{q}} &= N(\tilde{q}, \Omega) \\ &= N(\tilde{q}, 0) = 0 \end{aligned}$$

Substituting the value of $\Omega_d$ from (20) we have

$$\dot{\tilde{q}} = -B(q)\alpha\bar{q}$$

Defining

$$V = \tfrac{1}{2}(\tilde{q})^T\tilde{q} = 1 - q_0$$

With substitutions from (8) and (20), we have

$$\dot{V} = -\dot{q_0} = -\tfrac{1}{2}\bar{q}^T\alpha_i(\bar{q})$$

Using the property of quaternion representation [12] that $\bar{q}^T\bar{q} + {q_0}^2 = 1$, and since $0 \leq q_0 \leq 1$, we get

$$\begin{aligned} \dot{V} &= -\frac{1}{2}\alpha(1 + q_0)V \\ &< 0 \end{aligned}$$

which shows that for the desired input the system is input-to-state stable. This means $\lim_{t\to\infty}\tilde{q} = 0$ and from definition of $\tilde{q}$ we can conclude $\lim_{t\to\infty} q = q_e$.

According to the treatment given in [10], it can be shown that the system is asymptotically stable. For exponential stability; substituting (20) into (8) we have

$$\dot{q_0} = \frac{1}{2}\alpha(1 - q_0^2)$$

$$\dot{\bar{q}} = -\frac{1}{2}\alpha q_0\bar{q}$$

The time response of $q(t)$, by solving these differential equations can be found as

$$q_0(t) = 1 - 2c_1\frac{e^{-\alpha t}}{1 + c_1 e^{-\alpha t}} \tag{21}$$

$$\bar{q}(t) = \frac{1 + c_1}{1 + c_1 e^{-\alpha t}}e^{-0.5\alpha t}q(0) \tag{22}$$

where $c_1$ can be defined as $c_1 = \frac{1 - q_0(0)}{1 + q_0(0)}$. From (21) and (22) we can conclude that the quaternion system (8) is exponentially stable [10]. Also from (21) and (22), it can be seen that the parameter $\alpha$ is related to the settling time of the quaternion regulation and according to definition of the regulation settling time $t_q$ in [11], this relationship is given as

$$t_q = \frac{4.6}{0.5\alpha} = \frac{9.2}{\alpha} \tag{23}$$

*2) Desired Angular Velocity Tracking:* In the next step, we design $\tau_a$ such that it makes the angular velocity $\Omega$, asymptotically follows the desired angular velocity (20). The angular velocity tracking error can be described as $\tilde{\Omega} = \Omega - \Omega_d$. Assume that

$$\dot{\tilde{\Omega}} = -\lambda f(\tilde{\Omega}) \tag{24}$$

in which $\lambda$ is a positive constant and $f(\tilde{\Omega})$ is any function of $\tilde{\Omega}$ which satisfies

$$\tilde{\Omega} f(\tilde{\Omega}) > 0 \qquad \tilde{\Omega} \neq 0 \tag{25}$$

$$f(\tilde{\Omega}) = 0 \qquad \tilde{\Omega} = 0 \tag{26}$$

Defining the Lyapunov function candidate as

$$V = \tfrac{1}{2}(\tilde{\Omega})^T \tilde{\Omega}$$

The time derivative of $V$ while considering (24) is

$$\begin{aligned} \dot{V} &= -\lambda \tilde{\Omega} f(\tilde{\Omega}) \\ &< 0 \end{aligned}$$

which shows that $\lim_{t\to\infty} \tilde{\Omega} = 0$ and subsequently, we have $\lim_{t\to\infty} \Omega = \Omega_d$.

A model-dependent control law, $\tau_a$ can now be designed as

$$\tau_a = s(\Omega) I_f \Omega + G_a + I_f \dot{\Omega} \tag{27}$$

Using the definition of $\tilde{\Omega}$, we have that $\dot{\Omega} = \dot{\tilde{\Omega}} + \dot{\Omega}_d$, which with respect to (24) gives

$$\tau_a = s(\Omega) I_f \Omega + G_a - \lambda I_f f(\tilde{\Omega}) + I_f \dot{\Omega}_d \tag{28}$$

From (20), we have $\dot{\Omega}_d = J(q)\dot{q}$, where $J(q)$ is Jacobian matrix of $\Omega_d$ as given by

$$J(q) = \begin{pmatrix} 0 & -\alpha_1 & 0 & 0 \\ 0 & 0 & -\alpha_2 & 0 \\ 0 & 0 & 0 & -\alpha_3 \end{pmatrix}$$

Finally from (15) we get

$$\dot{\Omega}_d = J(q) F(\Omega) q$$

which yields

$$\tau_a = s(\Omega) I_f \Omega + G_a - \lambda I_f f(\tilde{\Omega}) + I_f J(q) F(\Omega) q \tag{29}$$

**Remark 1.** *As an example, one of the functions that can satisfy Equations (25) and (26) is*

$$f(\tilde{\Omega}) = sat(\tilde{\Omega}) = \begin{cases} \tilde{\Omega} & |\tilde{\Omega}| < a \\ sgn(\tilde{\Omega}) & |\tilde{\Omega}| \geq a \end{cases}$$

*Where the positive constant $a$ is the width of the boundary layer of the saturation function.*

**Remark 2.** *The control law (29) will ensure the asymptotic stability of the quadrotor if and only if the regulation of $\Omega$ to its equilibrium point (zero) is faster than regulation of $q$ to $q_e$ which means that*

$$t_\Omega < t_q \tag{30}$$

*Here, the angular velocity tracking error settling time $t_\Omega$ approximately is:*

$$t_\Omega = \frac{a + \alpha}{\lambda \alpha} \tag{31}$$

*Notice that the boundary layer width $a$, has to be sufficiently small such that it is in the angular velocity settling range. Also the control law (29) with respect to the airframe inertia uncertainties $\Delta I_f$ is robustly stable if the angular velocity tracking parameter $\lambda$ is*

$$\lambda > \lambda_0 = a(a + \frac{1}{2}\|J\|_\infty)\frac{\delta}{\sigma_{min}(I_{f0})}$$

*where $I_{f0}$ is nominal value of airframe's inertia matrix and $\Delta I_f \in \{y|\ \|y\|_\infty \leq \delta\}$.*

It is important to point out here that one can simply design a model-independent control law as

$$\begin{aligned} \tau_a &= \dot{\Omega} \\ &= -\lambda f(\tilde{\Omega}) + J(q) F(\Omega) q \end{aligned} \tag{32}$$

Further, with both the control laws of (29) and (32), the regulation problem of attitude angles to their desired values (i.e. at zero as in hovering case) results in an exponentially stable system, and the regulation settling time is a function of $\alpha$.

### *B. Designing the Rotor Torque $\tau_i$*

Having achieved the task of designing control torque (model dependent or model independent), we now proceed to designing $\tau_i$ such that the angular speed of the rotors ($w_i$'s), follow the desired angular speeds generated by our designed control torque $\tau_a = (\tau_\phi, \tau_\theta, \tau_\psi)^T$. From Equations (1)-(3) and (7), we find the desired angular speeds as

$$\begin{aligned} \begin{pmatrix} w_{d_1}^2 \\ w_{d_2}^2 \\ w_{d_3}^2 \\ w_{d_4}^2 \end{pmatrix} &= \begin{pmatrix} 0 & bd & 0 & -bd \\ bd & 0 & -bd & 0 \\ l & -l & l & -l \\ b & b & b & b \end{pmatrix}^{-1} \begin{pmatrix} \tau_\phi \\ \tau_\theta \\ \tau_\psi \\ T \end{pmatrix} \\ &= A^{-1} \begin{pmatrix} \tau_\phi \\ \tau_\theta \\ \tau_\psi \\ T \end{pmatrix} \end{aligned} \tag{33}$$

in which the parameters $b$, $d$ and $l > 0$ are assumed to be positive, in order to ascertain that the matrix $A$ remains nonsingular.

Tracking error of angular speed of rotor's can be described as $\tilde{w_i} = w_i - w_{d_i}$. For making the angular speeds of rotors asymptotically approach their respective $w_{d_i}$'s, assume that

$$\dot{\tilde{w_i}} = -h f(\tilde{w_i}) \tag{34}$$

in which $h$ is a positive constant and $f(\tilde{w_i})$ is any function of $\tilde{w_i}$ that satisfies (25) and (26) by replacing $\tilde{\Omega}$ with $\tilde{w_i}$. Defining the Lyapunov function candidate as

$$V = \frac{1}{2}(\tilde{w_i})^T\tilde{w_i}$$

The time derivative of $V$ while considering (34) is

$$\begin{aligned}\dot{V} &= -h\tilde{w_i}f(\tilde{w_i}) \\ &< 0\end{aligned}$$

which shows that $\lim_{t\to\infty}\tilde{w_i} = 0$ and subsequently, we have $\lim_{t\to\infty} w_i = w_{d_i}$.

We can simply define $\tau_i$ with respect to (10), as

$$\tau_i = Q_i + I_r\dot{w_i}$$

which can be written as

$$\tau_i = Q_i + I_r\dot{w_{d_i}} - hI_rf(\tilde{w_i}) \qquad (35)$$

One of the functions that can satisfy the conditions is

$$f(\tilde{w_i}) = sat(\tilde{w_i}) = \begin{cases} \tilde{w_i} & |\tilde{w_i}| < a \\ sgn(\tilde{w_i}) & |\tilde{w_i}| \geq a \end{cases} \qquad (36)$$

Notice that in Equation (35), one of the possible means of finding $\dot{w_{d_i}}$ can be by using the dirty derivative filter [4];

$$\dot{w_{d_i}} = \frac{s}{1+T_fs}w_{d_i}$$

## IV. Simulation Results

In order to ascertaining the performance of the quadrotor using the proposed control algorithm, we consider a quadrotor with dynamical parameters as listed in Table 1. Using Equation (23), the desired regulation settling time (attitude angles settling time) can give us the required parameters for our control law. Notice that, there is a compromise between choosing a small value of $t_q$ and having large peak value for the angular velocity $\Omega$, since with replacing the control law (29) or (32) in (9), we will find that peak value of the angular velocity $\Omega$ would be a function of $\alpha$, $\lambda$ and $a$. In other words, if we decrease $t_q$, we have increased $\alpha$ and decreased $\lambda$, which will increase the peak value. Furthermore, choosing $a$ as small as possible will cause to smaller peak values. In this paper we have considered the value of $a$ as:

$$a = 0.02\alpha \qquad (37)$$

Simulation 1 (Fig. 2 and Fig. 3) shows the performance of the control law (29) for desired regulation settling time ($t_q$) $0.2s$. From (23) and (37) respectively we will get $\alpha = 46$ and $a = 0.92$. One can choose $t_\Omega = 0.1$ and then from (31), $\lambda = 510$. Simulation 2 (Fig. 4 and Fig. 5) shows the performance of the control law (32) for the previous values.

Consider the initial angles as $\phi = -25°, \theta = 30°, \psi = -10°$, gain $h_i$ , $i\epsilon\{1,2,3,4\}$ as 0.002, $T = 1.5N$ and $T_f = 0.008$ (cutoff frequency of 20 Hz). In Simulations 1 and 2, the desired situation of quad rotor is in hovering. Notice that, performance of both the control laws is nearly equivalent. Trajectories of angular speed of rotors for the system described in Simulation 1 have been depicted in Fig. 6. In Simulation 3 (Fig. 7 and Fig. 8) the system is to be regulated to $\phi_r = 10°, \theta_r = -15°, \psi_r = 5°$. As before, performance of model-independent control law (32) is same as model-dependent control law (29). Trajectories of angular speed of rotors are shown in Fig. 9.

TABLE I

Considered dynamical values

| Dynamical Parameter | Considered Value |
|---|---|
| d | 0.225 m |
| $I_r$ | $3.4 \times 10^{-5}\ kg.m^2$ |
| $I_{f_\phi}$ | $4.9 \times 10^{-3}\ kg.m^2$ |
| $I_{f_\theta}$ | $4.9 \times 10^{-3}\ kg.m^2$ |
| $I_{f_\psi}$ | $8.8 \times 10^{-3}\ kg.m^2$ |
| $b$ | $2.9 \times 10^{-5}$ |
| $l$ | $1.1 \times 10^{-6}$ |

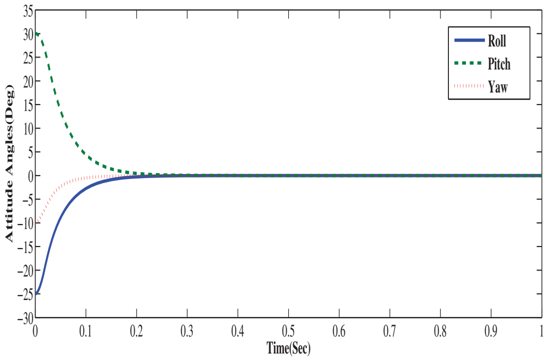


Fig. 2. **Simulation 1: Attitude Angles with Controller (29)**

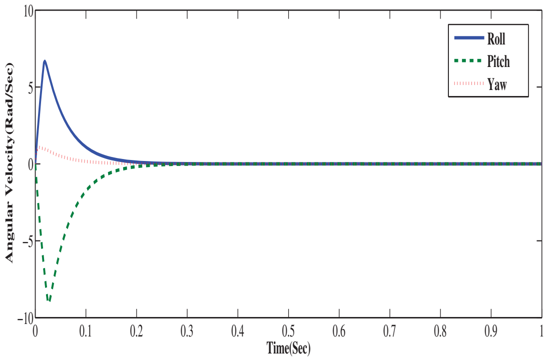


Fig. 3. **Simulation 1: Angular Velocities with Controller (29)**

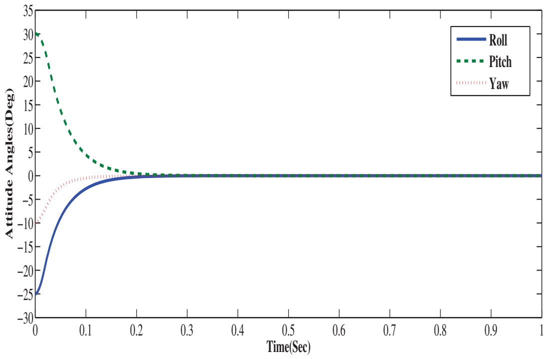


Fig. 4. **Simulation 2: Attitude Angles with Controller (32)**

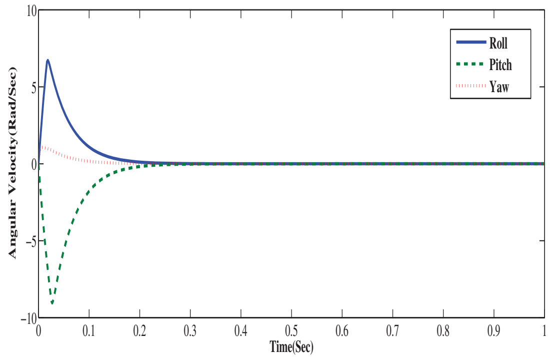


Fig. 5. **Simulation 2: Angular Velocities with Controller (32)**

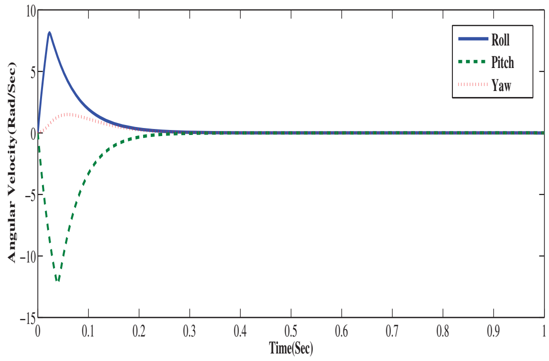


Fig. 8. **Simulation 3: Angular Velocities with Controller (29)**

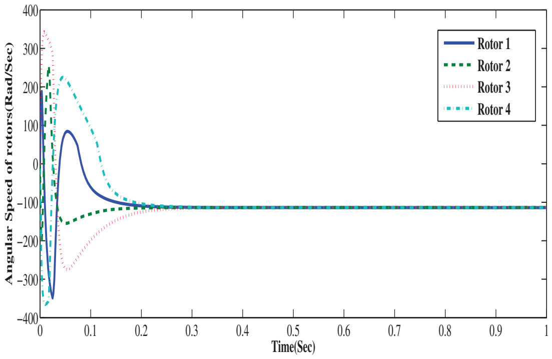


Fig. 6. **Simulation 1: Angular Speed of Rotors**

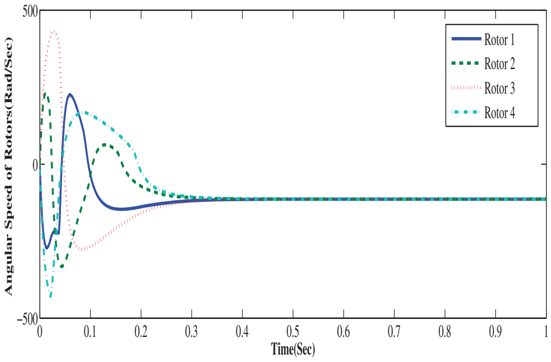


Fig. 9. **Simulation 3: Angular Speed of Rotors**

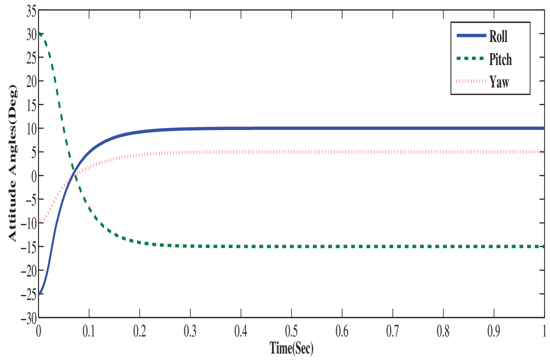


Fig. 7. **Simulation 3: Attitude Angles with Controller (29)**

## V. CONCLUSIONS

We presented a novel control approach to obtain asymptotic attitude stability of a quadrotor as a representative of VTOL-UAVs. Based upon the quaternion representation, we defined a robust control law for attitude control of the quadrotor in which the regulation of attitude angles to their desired values is shown to be exponentially stable, and that the desired settling values can be adjusted by the operator. In the proposed methodology, the tracking of desired angular velocity is shown to be asymptotically stable. The performance of the proposed control design has been ascertained using Simulation.

## REFERENCES

[1] A. Abdessameud and A. Tayebi, *Formation Stabilization of VTOL UAVs Subject to Communication Delays*, In proc. of the 49th IEEE CDC, Atlanta, GA, USA, 2010.

[2] J. T-Y. Wen and K. Kreutz-Delgado, *The Attitude Control Problem*, IEEE Transaction on Automatic Control, Vol. 36, No. 10, 1991.

[3] T. Hamel, R. Mahony, R. Lozano and J. Ostrowski, *Dynamic Modelling and Configuration Stabilization for an X4-Flyer*, In Proc. of IFAC World Congress, Barcelona, Spain, 2002.

[4] A. Tayebi and S. McGilvray, *Attitude Stabilization of a VTOL Quadrotor Aircraft*, IEEE Transactions on Control Systems Technology, Vol. 14, No. 3, 2006.

[5] P. Castillo, P. Albertos, P. Garcia, and R. Lozano, *Simple Real-time Attitude Stabilization of a Quad-rotor Aircraft with Bounded Signals*, In proc. of the 45th IEEE CDC, San Diego, CA, USA, 2006.

[6] A. Tayebi, S. McGilvray, A. Roberts, and M. Moallem, *Attitude estimation and stabilization of a rigid body using low-cost sensors*, In proc. of the 46th IEEE CDC, New Orleans, LA, USA, 2007.

[7] A. Tayebi, A. Roberts and A. Benallegue, *Inertial Measurements Based Dynamic Attitude Estimation and Velocity-Free Attitude Stabilization*, In Proc. of ACC, San Francisco, CA, USA, 2011.

[8] R. Mahony and T. Hamel, *Adaptive compensation of aerodynamic effects during takeoff and landing maneuvers for a scale model autonomous helicopter*, European Journal of Control, Vol.7, 2001.

[9] R. M. Murray, Z. Li and S. S. Sastry , *A Mathematical introduction to Robotic Manipulation*, CRC Press, 1994.

[10] H. K. Khalil, *Nonlinear Systems*, 3rd ed. Prentice Hall, 2002.

[11] J. Lu and B. Wie, *Nonlinear Quaternion Feedback Control for Spacecraft via Angular Velocity Shaping*, In proc. of ACC, Baltimore, MD, USA, 1994.

[12] S. M. Joshi, A. G. Kelkar and J. T-Y. Wen, *Robust Attitude Stabilization of Spacecraft Using Nonlinear Quaternion Feedback*, IEEE Transactions on Automatic Control, Vol. 40, No. 10, 1995.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Experimental validation of a geometric method for the design of stable and broadband vibration controllers using a propeller blade test rig

Ubaid Ubaid*, Steve Daley†, Simon Pope*, Ilias Zazas†
*Automatic Control and Systems Engineering, University of Sheffield, S1 3JD, UK.
†Institute of Sound and Vibration Research, University of Southampton, SO17 1BJ, UK.

***Abstract*—A systematic geometric design methodology to generate a stable controller for simultaneous local and remote attenuation that was previously proposed is experimentally validated on a structure. The local control path transfer function for this experimental system is non-minimum phase due to which the original broadband controller design would yield an unstable controller. Here a modified procedure for systems with local non-minimum phase dynamics is used to generate a stable controller. According to this method, reduction in vibration at local and remote points on a structure can be parameterised in terms of the available design freedom and a controller is realised in terms of the optimal selection of this using the minimum phase counterpart of the local control path transfer function. The modified method results in a controller that is both stable and stabilizing and which achieves the desired vibration attenuation at the local and remote points on the structure. An experimental facility that replicates the vibration transmission through the shaft of a propeller blade rig system is used to demonstrate the method. Vibration for excitation near the first bending mode frequency of the resonating part of this structure is attenuated at the non-resonating part of the system without deteriorating vibration at the resonating end.**

## I. Introduction

Active control for reduction of vibration at a specific point on a complex interconnected structure can potentially enhance vibration at other points on the structure [1]. A controller design technique to address the vibration attenuation problem at local and remote points on a structure simultaneously using only a single locally placed sensor actuator pair was presented for a discrete frequency excitation case in [2] and later extended for the broadband case [3]. For vibration attenuation over an arbitrary frequency band, controller implementation involves inversion of the local control path transfer function. When the local control path transfer function is non minimum phase, then the controller itself would be unstable. This problem can be solved using a new design freedom [4] to parameterise reduction in vibration at local and remote points whereby the controller is implemented in terms of the minimum phase counterpart of the local control path transfer function. Additionally, to improve robustness to unmodelled high frequency dynamics, a filter is incorporated into the design freedom selection [5] such that the gain of the closed loop system rolls off at high frequency without deteriorating controller performance in the excitation frequency bandwidth. The aim of this paper is to present experimental verification of this design technique for attenuation of vibration at both local and remote points on a blade rig simultaneously. Trade-offs between stability robustness and disturbance attenuation are also highlighted in terms of the values of the design freedom parameter.

## II. Experimental set-up

A schematic diagram of the blade rig is shown in figure 1. The primary excitation signal $f_p(t)$ is a common signal fed to the two smaller shakers attached at both ends of the bar which acts as the transient loading force due to rotation of the propeller blades. The vibration at the blade end of the shaft $q_p(t)$ is the summation of outputs measured by two accelerometers connected near each of the disturbance shakers. The control input $f_c(t)$ is applied to the control shaker attached at the other end of propeller shaft on the thrust block and a local accelerometer on the thrust block measures local vibration levels $q_c(t)$. Vibration is transmitted from the blade end along the shaft to the thrust block end and is particularly detrimental at the blade resonant frequency. Due to difficulties in measuring and actuating at the blade end for most applications, it is desired to control both blade vibration and its transmission using sensors and actuators placed at the thrust block only[1]. This blade system can be considered as a two input two output system with the transfer function matrix relating the disturbance and control inputs to the remote and local vibration outputs as

$$\begin{bmatrix} q_c(j\omega) \\ q_p(j\omega) \end{bmatrix} = \begin{bmatrix} g_{cc}(j\omega) & g_{cp}(j\omega) \\ g_{pc}(j\omega) & g_{pp}(j\omega) \end{bmatrix} \begin{bmatrix} f_c(j\omega) \\ f_p(j\omega) \end{bmatrix} \quad (1)$$

In the next section, a design freedom parameter is introduced which parameterizes reduction in vibration at the thrust block and blade end. A controller implemented in terms of the optimally selected values of this design freedom would achieve the targeted vibration reduction.

## III. General algorithm

The detailed synthesis of a stable controller using the geometric approach is given in [4], [5]. For the system described in (1), the aim is to design a feedback controller $k(j\omega)$ using as feedback signal only from the thrust block to achieve vibration

[1] Note that this general concept is the subject of several BAE Systems patents

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

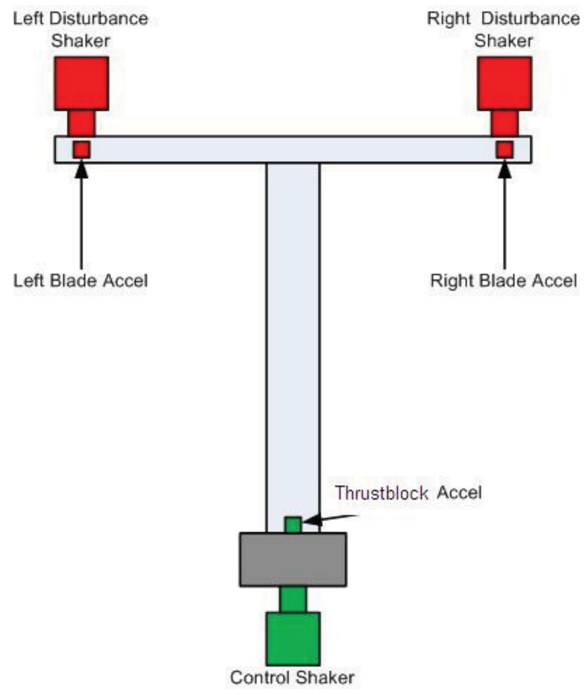


Figure 1. Schematic of blade rig

reduction at both the thrust block and blade end simultaneously. Using control action, $f_c(j\omega) = -k(j\omega)q_c(j\omega)$, the closed loop output respectively at the local and remote points is then given as

$$q_c(j\omega) = \left[1 + \frac{-g_{cc}(j\omega)k(j\omega)}{1 + g_{cc}(j\omega)k(j\omega)}\right] g_{cp}(j\omega)f_p(j\omega) \tag{2}$$

and

$$q_p(j\omega) = \left[1 + \frac{-g_{cc}(j\omega)k(j\omega)}{1 + g_{cc}(j\omega)k(j\omega)} \frac{g_{cp}(j\omega)g_{pc}(j\omega)}{g_{cc}(j\omega)g_{pp}(j\omega)}\right] g_{pp}(j\omega)f_p(j\omega) \tag{3}$$

Denote a design freedom parameter $\gamma$ which is related to the sensitivity function $S(j\omega)$ as

$$\gamma(j\omega) = \frac{1}{g_{AP}(j\omega)f_{LP}(j\omega)}\left[S(j\omega) - 1\right] \tag{4}$$

where $g_{AP}(j\omega)$ is the all pass transfer function formed from the right half plane zeros of $g_{cc}(j\omega)$ and $f_{LP}(j\omega)$ is a low pass or a bandpass filter which improves robustness at out-of-band frequencies. The closed loop local and remote outputs in (2) and (3) can be represented in terms of this parameter variable as

$$q_c(j\omega) = \left[1 + \gamma(j\omega)g_{AP}(j\omega)f_{LP}(j\omega)\right] g_{cp}(j\omega)f_p(j\omega) \tag{5}$$

and

$$q_p(j\omega) = \left[1 + \gamma(j\omega)g_{AP}(j\omega)f_{LP}(j\omega)\frac{g_{cp}(j\omega)g_{pc}(j\omega)}{g_{cc}(j\omega)g_{pp}(j\omega)}\right] g_{pp}(j\omega)f_p(j\omega) \tag{6}$$

The magnitude of the expression inside brackets in RHS of (5) and (6) determines reduction in closed loop local and remote output. The magnitude of these terms can be represented as a circle at each discrete frequency in a $\gamma-$plane. A value for $\gamma$ from inside the circle corresponds to reduction in vibration. If values for $\gamma$ at discrete frequencies in the disturbance frequency bandwidth $[\omega_L, \omega_H]$ is selected from inside the circles given by inequalities (7) and (8), then reduction in vibration at the local and remote points is possible.

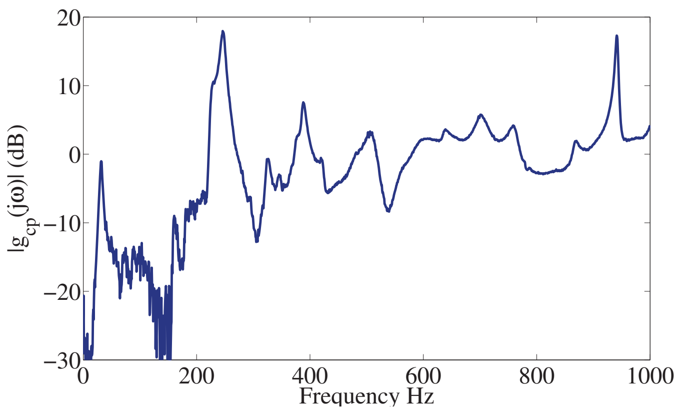


Figure 2. Measured frequency response of primary excitation input to thrust block vibration output

$$\left|\gamma(j\omega_i) + \frac{1}{g_{AP}(j\omega_i)f_{LP}(j\omega_i)}\right| < \left|\frac{1}{g_{AP}(j\omega_i)f_{LP}(j\omega_i)}\right| \tag{7}$$

$$\left|\gamma(j\omega_i) + \frac{1}{g_{AP}(j\omega_i)f_{LP}(j\omega_i)}\frac{g_{cc}(j\omega)g_{pp}(j\omega)}{g_{cp}(j\omega)g_{pc}(j\omega)}\right| < \left|\frac{1}{g_{AP}(j\omega_i)f_{LP}(j\omega_i)}\frac{g_{cc}(j\omega)g_{pp}(j\omega)}{g_{cp}(j\omega)g_{pc}(j\omega)}\right| \tag{8}$$

These optimal $\gamma$ points are interpolated by a stable transfer function and the controller is implemented as

$$k(j\omega) = -\frac{\gamma(j\omega)f_{LP}(j\omega)}{\left[1 + \gamma(j\omega)g_{AP}(j\omega)f_{LP}(j\omega)\right]g_{MP}(j\omega)} \tag{9}$$

where $g_{MP}(j\omega)$ is the minimum phase counterpart of the local control path transfer function.

The magnitude frequency response of the open loop path from primary excitation input on the blade to the local and remote outputs, denoted as $g_{cp}(j\omega)$ and $g_{pp}(j\omega)$ respectively, is plotted in figures 2 and 3. It shows that near the frequency of the first bending mode of the blade (i.e. the iron bar connected to one end of the shaft), vibration transmission to the thrust block is amplified. A feedback controller to achieve simultaneous reduction in the thrust block and blade vibration outputs will be designed for primary excitation around this frequency range.

The first step in controller design is to determine an LTI model for the open loop control path $g_{cc}(j\omega)$ from the measured FRF. As the controller will target vibration reduction in the low frequency region, the measured FRF of the path from control shakers to acceleration on thrust block at frequencies below 800 Hz is fitted with a $15^{th}$ order transfer function model using least squares. The dynamics neglected at frequencies higher than 800 Hz will not be a problem as they will not be excited by the control action due to low pass

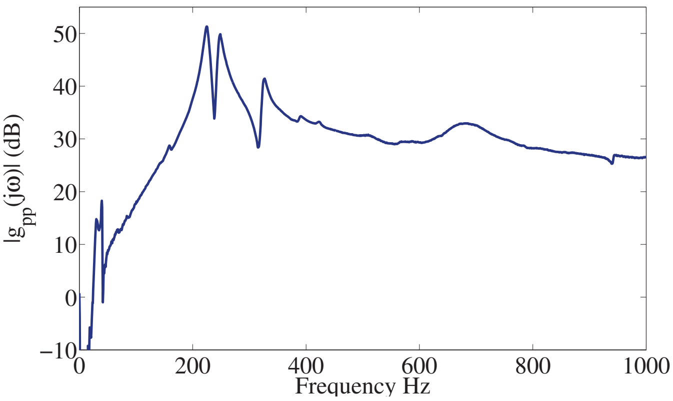


Figure 3. Measured frequency response of primary excitation input to blade vibration output

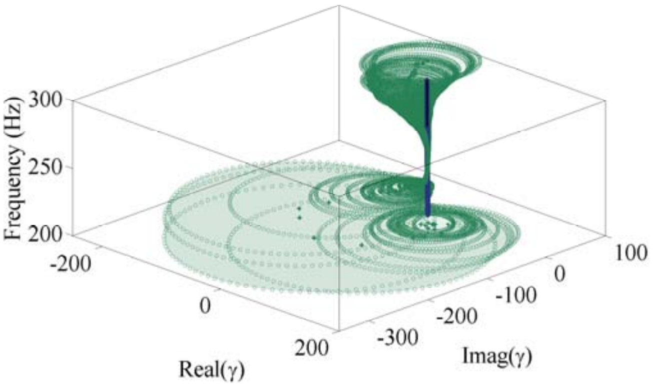


Figure 4. Circles that represent remote vibration attenuation in $\gamma-$ plane at discrete frequencies between 200 Hz and 300 Hz. Unit radius circles corresponding to local vibration attenuation appear as a cylinder that passes through the origin

filter $f_{LP}(j\omega)$. This identified transfer function has 1 right half plane zero so that $g_{AP}(j\omega)$ is of order 1 and $g_{MP}(j\omega)$ has this RHP zero reflected into the LHP.

The circles in the $\gamma-$plane corresponding to reduction in vibration at the local and remote outputs given by (7) and (8) for frequencies around the first bending mode is shown in figure 4. The circles corresponding to reduction in vibration at the blade end are very large in the frequency region from 200 Hz to 220 Hz and above 270 Hz. The distance between the centre of both circles will be large and so it will not be possible to achieve considerable vibration reduction at the blade end without amplifying vibration output at the thrust block end. Optimal $\gamma$ points at these frequencies are selected such that the vibration level at the thrust block is reduced without enhancing the vibration level at the blade end. $f_{LP}(j\omega)$ is chosen as a bandpass filter with lower and higher cut-off frequency as 100 Hz and 600 Hz, respectively.

## IV. Nevanlinna-Pick interpolation

The set of selected optimal $\gamma$ points at discrete frequencies in the disturbance frequency bandwidth is interpolated by a stable transfer function using the Nevanlinna Pick interpolation algorithm [6]. This interpolation problem can be stated as follows: given $n$ distinct points $s_1, \ldots, s_n$ in the right half plane $\Pi^+$ and a collection of complex numbers $H_1, \ldots, H_n$, determine a transfer function $f(s)$ that is analytic in $\Pi^+$ with

$$\sup |f(s)| \leq 1$$

such that $f(s_i) = H_i$, for all $i = 1, \ldots, n$. The solution of this interpolation exists if and only if the associated Pick matrix $P$

$$P = \left[ \frac{1 - H_k \overline{H}_l}{s_k + \overline{s}_l} \right]_{k,l=1}^{n}$$

is positive definite, where $\bar{\bullet}$ denotes complex conjugate. The points $s_i$ according to the above theorem should strictly belong to the Right Half Plane, whereas the set of optimal selected values for the design freedom $\gamma$ have to be interpolated on the imaginary $j\omega$ axis. The frequency points are shifted into the RHP using transformation of Lemma 2 in [8] and is stated as follows: for the optimal $\gamma$ data values at $n$ discrete frequencies $\omega_i$, for $i = 1, \ldots, n$, a stable transfer function $\gamma(j\omega)$ exists if and only if the associated Pick matrix

$$P = \left[ \frac{1 - W_k \overline{W}_l}{s_k + \overline{s}_l} \right]_{k,l=1}^{n} \tag{10}$$

is positive definite, where $W_i = {}^{\gamma_i}/_{M}$ and $s_i = \sigma + j\omega_i$, for $i = 1, \ldots, n$. where $M$ is the maximum modulus of interpolated transfer function and $\sigma$ a positive real number. Increasing $M$ or decreasing $\sigma$ increases the positive definitness of the pick matrix but for a stable controller and good performance at intermediate frequencies, $M$ and $\sigma$ values have to be finely tuned. It should be noted that small values of $\sigma$ will give interpolated transfer function $\gamma(j\omega)$ with poles that are close to the imaginary axis. This would cause oscillations in the frequency response of the identified transfer function $\gamma(j\omega)$ leading to gain and phase crossover at intermediate frequencies. As the non-interpolated points in the disturbance frequency band for small values of $\sigma$ may lie outside (7) and (8) circles in $\gamma-$plane, this would deteriorate controller performance. Although large value of $\sigma$ gives better approximation at intermediate frequencies, $M$ values will have to be increased to get a positive definite pick matrix. A large value of $M$ can result in the nyquist contour of $\gamma(j\omega)g_{AP}(j\omega)f_{LP}(j\omega)$ encircling the critical point which will cause the controller to become unstable as can be seen from (9). If the Pick matrix (10) of optimally selected $\gamma$ points is not positive definite, then an approximate set of sub-optimal points adjusted to lie inside the circles is obtained using Linear Matrix Inequalities that also satisfy the pick condition. This new set of $\gamma$ points is used to obtain the interpolating function using the classical N-P interpolation algorithm.

The first step in iterative classical N-P interpolation algorithm is to compute the elements of Fenyves array $T$.

$$T_{k,l} = \frac{s_l + \bar{s}_{k-1}}{s_l - s_{k-1}} \frac{T_{k-1,l} - T_{k-1,k-1}}{1 - T_{k-1,l}\overline{T}_{k-1,k-1}}$$
$$2 < k < n, \; k < l < n \tag{11}$$

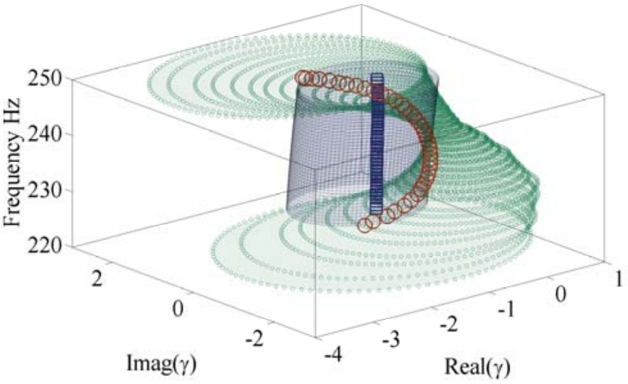


Figure 5. Final operating $\gamma$ points as the frequency response of a stable bounded real interpolated transfer function $\gamma(j\omega)$

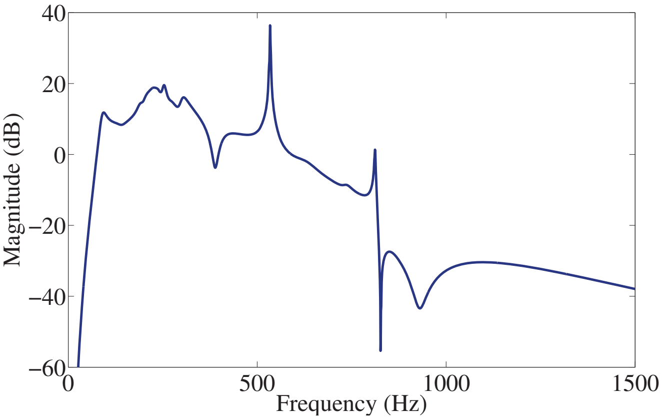


Figure 6. Magnitude plot of the frequency response of controller

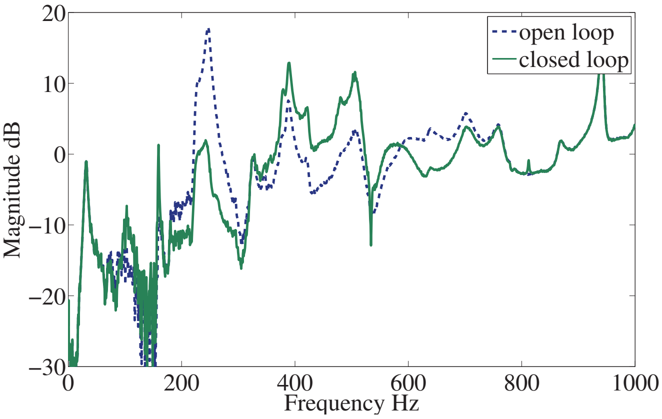


Figure 7. Magnitude plot of the frequency reponse from disturbance to local output with (dashed) and without (solid) feedback controller

where $T_{1,l} = W_l$, for $1 < l < n$. The next step is to recursively estimate $W_1(s)$ from

$$W_k(s) = \frac{T_{k,k} + W_{k+1}(s)\frac{s-s_k}{s+\bar{s}_k}}{1 + \overline{T}_{k,k}W_{k+1}(s)\frac{s-s_k}{s+\bar{s}_k}}, \qquad k = n, n-1, \ldots, 2, 1 \tag{12}$$

If the set of data points $(s_i, W_i)$ for interpolation is augmented with its complex conjugate $(\bar{s}_i, \overline{W}_i)$, then a stable bounded real analytic interpolating function is given by $\gamma(s) = M \times \frac{1}{2}\left[W_1(s+\sigma) + \overline{W}_1(s+\sigma)\right]$ for any arbitrarily selected initial stable bounded analytic function $W_{k+1}(s)$ in (12). There will be at least 4 poles and zeros in $\gamma(j\omega)$ transfer function for every interpolated $\gamma$ point which will affect the order of the final compensator transfer function $k(j\omega)$. In the frequency interval from 200 Hz to 300 Hz there are 164 discrete frequencies at which an optimal $\gamma$ point is selected. Only 6 of the optimal $\gamma$ points are used as interpolation data points in order to get a lower order controller.

The final operating $\gamma$ points obtained from the frequency response of the interpolating function $\gamma(j\omega)$ for frequency 225 Hz to 250 Hz is shown in figure 5. It is seen that in the frequency range 230 Hz to 245 Hz, circles representing remote vibration reduction converge towards the origin and become very small. Due to this several more optimal $\gamma$ points have to be selected in this frequency band alone to get a good transfer function approximation, but this will increase the order of interpolated transfer function considerably. At all other frequencies in the disturbance frequency band, circles representing reduction in vibration at the remote point are considerably larger than the unit circle that corresponds to local vibration attenuation. Hence, final operating $\gamma$ points from inside the unit circle will lie on the boundary of the remote vibration reduction circle. This is predicted to achieve only slight reduction in the blade vibration output using a $58^{th}$ order controller transfer function. This vibration attenuation problem is a case of very extreme magnitude for the dimensionless parameter discussed in [7], which is equivalent to the function formed by the centre of remote vibration reduction circle given as $-g_{cc}(j\omega)g_{pp}(j\omega)\left[g_{cp}(j\omega)g_{pc}(j\omega)g_{AP}(j\omega)f_{LP}(j\omega)\right]^{-1}$. The magnitude of this function is a measure of the severity of the trade-off between disturbance attenuation and stability robustness.

A controller realized in terms of this $\gamma(j\omega)$ transfer function by substituting in (9) has a magnitude frequency spectrum as shown in figure 6. The gain of controller starts to roll-off at 600 Hz due to the filter action thereby improving robustness to unmodelled high frequency dynamics. The theoretical closed loop frequency response of local and remote points with the designed controller is compared with the open loop frequency response as shown in figures 7 and 8.

## V. Experimental Results

The controller obtained in the previous section is a compensator transfer function in continuous time domain. It is integrated with the experimental set-up through Simulink using a dSPACE real time interface prior to which it has to be converted to a discrete time model. A discrete model of compensator using first order hold method with a sampling frequency of 5 kHz is obtained which matches exactly the frequency characteristic of the continuous time domain compensator in the disturbance frequency bandwidth. For different primary excitation inputs $f_p(t)$, the acceleration output at the thrust block $q_c(t)$ and blade end $q_p(t)$ are measured to compare closed loop output against open loop output.

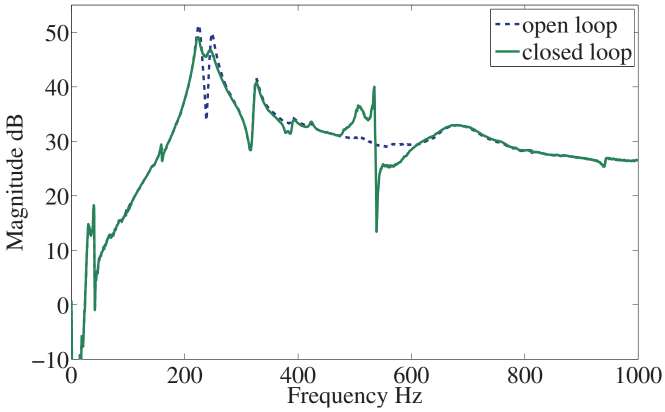


Figure 8. Magnitude spectrum of the frequency reponse from disturbance to remote output with (dashed) and without (solid) feedback controller

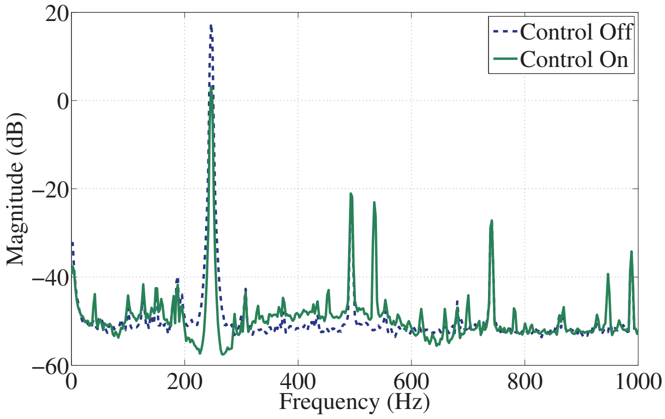


Figure 9. Power spectral density of output from thrust block when primary excitation is discrete frequency excitation at 247 Hz

### A. *Sinusoidal excitation at discrete frequency 247 Hz*

The power spectral density of measured acceleration at the thrust block with and without feedback control for a primary excitation signal at 247 Hz is plotted in figure 9. It shows more than 14 dB reduction in magnitude at 247 Hz using feedback controller and the power spectral density of measured acceleration at the blade end in figure 10 shows around 3 dB reduction in magnitude at 247 Hz. A peak at 534 Hz in the PSD of both closed loop outputs can be noticed which is not present for the open loop case. This is caused by the peak at this frequency in the magnitude of controller frequency response spectrum. In the next section, this peak is reduced without affecting controller performance in the design frequency bandwidth using a notch filter.

## VI. Results after augmenting a notch filter

The controller implementation involves inversion of the minimum phase counterpart of the local control path transfer function so an antiresonance at frequency 534 Hz in the local control path transfer function appears as a peak in the controller FRF. Filter $f_{LP}(j\omega)$ has a high cut-off frequency 600 Hz which is higher than the frequency (534 Hz) at which the peak in controller FRF appears. If the high cut-off frequency of $f_{LP}(j\omega)$ is reduced below 534 Hz and the order of filter is increased in order to take account of the sharp increase in this peak then due to limitations as quantified by Bode's sensitivity integral, amplification at out-of bound frequencies will not be spread over a large frequency range and there will be peaks appearing in the closed loop frequency response. Therefore, the controller is implemented in series with a notch filter which has a notch at 534 Hz in order to reduce the peak at this frequency. The magnitude and phase of the controller is unaffected in the disturbance frequency bandwidth. The acceleration measurements at the thrust block and blade end are taken for different disturbance excitation signals to compare reduction in closed loop output.

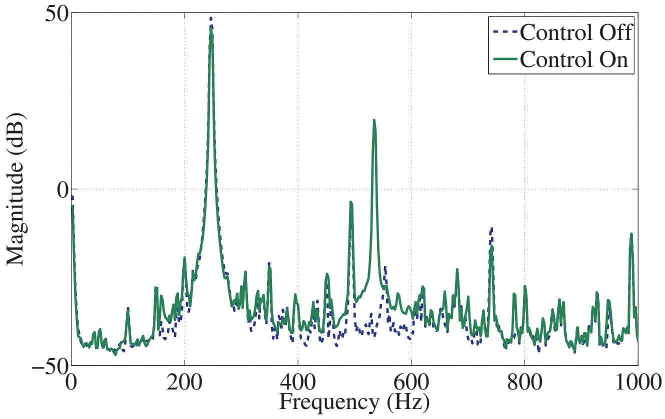


Figure 10. Power spectral density of output from blade end when primary excitation is discrete frequency excitation at 247 Hz

### A. *Sinusoidal excitation at discrete frequency 247 Hz*

The power spectral density of acceleration measured at the thrust block with and without feedback control for a primary excitation signal at 247 Hz shows 16 dB reduction in magnitude as shown figure 11. The power spectral density of acceleration measured at the blade end in figure 12 shows around 3 dB reduction in magnitude at 247 Hz. The peak at 534 Hz is reduced considerably because of the notch filter at this frequency.

### B. *Broad band frequency white noise excitation*

The primary excitation signal input to the disturbance shaker is a random white noise signal and the power spectral density of acceleration outputs from the thrust block $q_c(t)$ and blade end $q_p(t)$ with and without feedback control action is plotted in figures 13 and 14. The PSD of acceleration measured at the thrust block shows around 12 dB attenuation in the disturbance frequency bandwidth while the PSD of measured blade acceleration shows no amplification and a small attenuation as designed.

## VII. Conclusion

A geometric design methodology for vibration control using remotely located stable control systems has been demonstrated

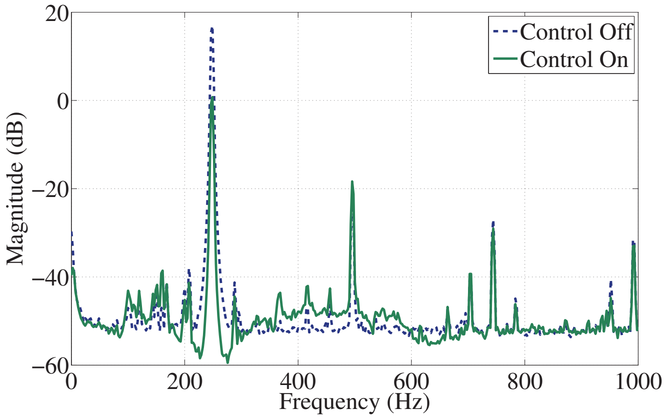


Figure 11. Power spectral density of output from thrust block when primary excitation is discrete frequency excitation at 247 Hz

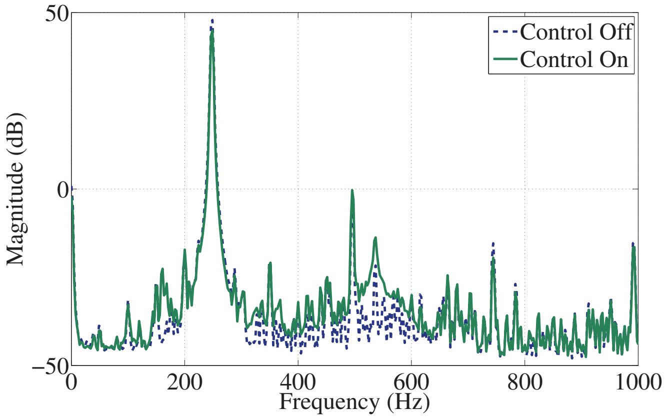


Figure 12. Power spectral density of output from blade end when primary excitation is discrete frequency excitation at 247 Hz

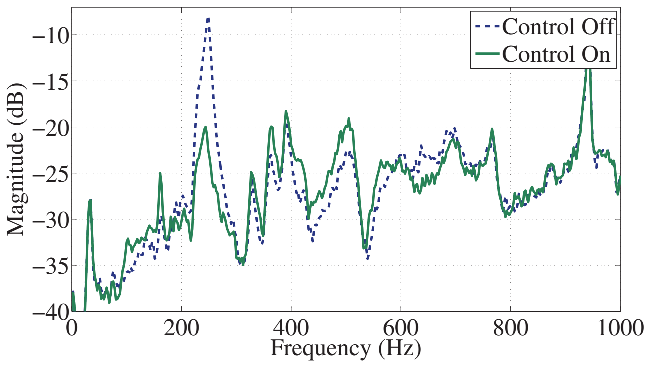


Figure 13. Power spectral density of output at thrust block with primary excitation as random white noise excitation

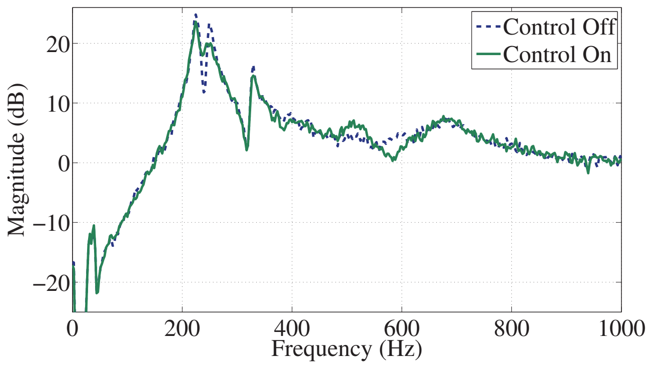


Figure 14. Power spectral density of output at blade end with primary excittion as random white noise excitation

experimentally on a blade rig experimental set-up, which mimicks the vibration transmission problems encountered due to propeller blade excitations encountered in many aerospace and maritime applications. The limitations on actuator and sensor placement can be overcome using this control design approach and shows considerable reduction in closed loop output at the thrust block side near the problematic blade resonant frequency.

## REFERENCES

[1] J. Post and R. Silcox, *Active control of the forced response of a finite beam*, Noise Control Engineering Journal, 1 (1990), pp. 197-202.

[2] S. Daley and J. Wang, *A geometric approach to the design of remotely located vibration control systems*, Journal of Sound and Vibration, 318 (2008), pp. 702-714.

[3] J. Wang and S. Daley, *Broad band controller design for remote vibration using a geometric approach*, Journal of Sound and Vibration, 329 (2010), pp. 3888-3897.

[4] U. Ubaid, S. Daley and S. Pope, *Broad band design of remotely located vibration control systems: a stable solution for non-minimum phase dynamics*, Internoise 2011, Osaka, Japan, September 4-7 2011.

[5] U. Ubaid, S. Daley and S. Pope, *Design of remotely located stable vibration controllers for non-minimum phase systems*, in 14th Asia Pacific Vibration Conference, Hong Kong, December 5-8 2011.

[6] Ph. Delsarte, Y. Genin and Y. Kamp, *The Nevanlinna-Pick problem for matrix-valued functions*, SIAM Journal on Applied Mathematics, Vol. 36, 1 (1979), pp. 47-61.

[7] J. Freudenberg, C. Hollot and R. Middleton, *A tradeoff between disturbance attenuation and stability robustness*, in American Control Conference, 2003. Proceedings of, vol. 6, IEEE, 2003, pp. 4816-4821.

[8] G. Ferreres and G. Puyou, *Feasibility of $H\infty$ design specifications: an interpolation method*, International Journal of Control, 78 (2005), pp. 927-936.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# RESEARCH ON SPEED-SENSORLESS INDUCTION MOTOR CONTROL SYSTEM BASED ON AMESIM-SIMULINK SIMULATION

JIA Xiaoyan
State Key Laboratory of Automobile Simulation and Control
Department of Control Science and Engineering,Jilin University
Changchun, PR China
E-mail: jiaxy09@mails.jlu.edu.cn

XIE Xiaohua
State Key Laboratory of Automobile Simulation and Control
Department of Control Science and Engineering,Jilin University
Changchun, PR China
E-mail: xiexh@jlu.edu.cn

WANG Xue
State Key Laboratory of Automobile Simulation and Control
Department of Control Science and Engineering,Jilin University
Changchun, PR China
E-mail: xuew11@mails.jlu.edu.cn

CHEN Hong
China First Automobile Works Group
Research and Development Center
Changchun, PR China
E-mail: chenhong1@rdc.faw.com.cn

***Abstract*—This paper will establish the simulation model and the inverter model of induction motor based on the multi function simulation software AMESim. Considering the influence on the motor from the temperature and other complex factors, and Using Matlab / Simulink software to establish a simulation model of direct torque control system, then we go on combined simulation through the combination of the two aspects. The simulation result shows that the controller make the speed faster and the robustness stronger of the motor control system, but at the same time it exposes the shortcoming that torque pulsing will enlarge at the low speed. Pointing at the shortcoming, this article starts with the principle of extended kalman filtering algorithm, uses of measuring current, voltage on the pattern of stator, deduces from a new motor speed, flux state observer model without speed sensor, and estimates the rotor speed and stator flux. The simulation result shows that this method can solve the problem of flux and torque pulsing effectively.**

***Keywords- Extended Kalman filter; direct torque control; the simulation software AMESim; torque ripple***

## I. INTRODUCTION (HEADING 1)

The full name of AMESim is LMS Imagine.Lab AMESim, that is complex system modeling and simulation platform of the interdisciplinary field. It provides a complete platform of system project design, this makes it possible that the user can build complex model of multidisciplinary field system on a platform. By using of various components provided by system, the user can using the physical model to build the simulation platform, and this improved the difficulties that general modeling software to do it by using of complicated mathematical model. AMESim model is more close to the reality.

Direct torque control is a new type of high performanced AC adjustable-speed transmission control technique. It abandonees the decoupling control theory in vector control. By using of the stator flux orientation and instantaneous space vector theory, and detecting the stator voltages and currents, this control method observes flux and torque of motor in the stator coordinate, then compares the observation value with the given one. By hysteresis controller regulating the error value gets corresponding control signal. To control motor, through synthesizing flux and torque signal to select the corresponding voltage space vector. The advantage of direct torque control is that the torque has faster dynamic response, and the rotor parameters change has certain robustness. The defect of this control method is that torque and flux pulsing is strong especially in the low speed.

The motor speed and flux observer given in this paper check the feasibility of the algorithm, by computer simulation and the comparison between the actual rotor speed and stator flux. The simulation results show that the design of new state observer can reduce the motor flux and torque pulsing effectively, and improve the motor torque performance. And it make high performance motor control possible.

## II. THE INDUCTION MOTOR STATE ESTIMATION BASED ON THE EXTENDED KALMAN FILTER

The object of this paper is the AC induction motor, which is a typical nonlinear system. Under normal circumstances, the Kalman filter is used for state estimation of linear systems, if it is used in the state estimation of nonlinear systems, you must consider to use the extended Kalman filter (Extended Kalman

Filter, EKF). The largest difference between the general Kalman filter and EKF is attempting to make the nonlinear systems linearization [3-7]. After the Jacobian matrix linearization, there is little difference with the general Kalman filter in form.

### *A. The Extended Kalman Filter Design Based On AC Induction Motor*

In two-phase stationary coordinate, in order to use EKF to predict the rotor speed, we need to add to a new state variable [8] [9]. The block diagram of the direct torque control system based on the extended Kalman filter is shown in figure 1. That is where we use voltage and current value which is easy to measure to estimate the value of the stator flux and achieve the speed identification based on the extended Kalman filter algorithm. So as to improve the performance of induction motor direct torque control system [10] [11].

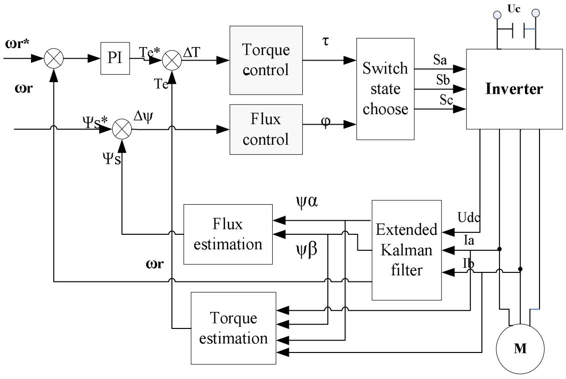


Figure 1 the system block diagram of EKF-DTC

Under the conditions that the sampling period is very short, we can think that the change is zero. Rewriting of the induction motor state-space model:

$$\begin{aligned}\dot{x} &= A\cdot x(t)+B\cdot u(t)\\ y(t) &= C\cdot x(t)\end{aligned} \tag{1}$$

Among the formula:

State variable $x=[i_{\alpha s}\ i_{\beta s}\ \psi_{\alpha r}\ \psi_{\beta r}\ \omega_r]^{\mathrm{T}}$ , measurement variable $y=[i_{\alpha s}\ i_{\beta s}]^{\mathrm{T}}$ , input variable $u=[u_{\alpha s}\ u_{\beta s}]^{\mathrm{T}}$ ,

$$A=\begin{bmatrix} -\dfrac{L_m^2R_r+L_r^2R_s}{\sigma L_sL_r^2} & 0 & \dfrac{L_mR_r}{\sigma L_sL_r^2} & \dfrac{n_pL_m\omega_r}{\sigma L_sL_r^2} & 0\\ 0 & \dfrac{L_m^2R_r+L_r^2R_s}{\sigma L_sL_r^2} & -\dfrac{n_pL_m\omega_r}{\sigma L_sL_r^2} & \dfrac{L_mR_r}{\sigma L_sL_r^2} & 0\\ \dfrac{R_rL_m}{L_r} & 0 & -\dfrac{R_r}{L_r} & -n_p\omega_r & 0\\ 0 & \dfrac{R_rL_m}{L_r} & n_p\omega_r & -\dfrac{R_r}{L_r} & 0\\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$B=\begin{bmatrix} \dfrac{1}{\sigma L_s} & 0\\ 0 & \dfrac{1}{\sigma L_s}\\ 0 & 0\\ 0 & 0\\ 0 & 0 \end{bmatrix} \qquad C=\begin{bmatrix} 1 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Where $R_s$ is the stator winding resistance, $R_r$ is the rotor winding resistance, $L_s$ is the stator cyclic inductance, $L_r$ is the rotor cyclic inductances, $L_m$ is mutual cyclic inductance, $\sigma=1-\dfrac{L_m^2}{L_s\tau_r\sigma}$ , $n_p$ is the number of pairs. $\omega_r$ is the rotor speed. It can be seen from the above state equation expression, in addition to the speed, the rest of the amount are all the stator side variables or motor constants, but also it shows the system is a nonlinear system. In other words we can use the extended Kalman filter algorithm. The actual system contains the system and measurement noise inevitably, so we add the noise matrix to the fifth-order model of the AC induction motor. It constitutes the following stochastic nonlinear control system:

$$\begin{aligned}\dot{x}(t) &= f(x(t))+W(t)\\ y(t) &= h(x(t))+V(t)\end{aligned} \tag{2}$$

Among the formula:

$f(x(t))=A\cdot x(t)+B\cdot u(t)$ , $h(x(t))=C\cdot x(t)$ , $W(t)$ is the system noise matrix, $V(t)$ is the process noise matrix. So the space state equation of an AC induction motor conforms the form of general nonlinear systems state equation in 2.1 section. So this paper may be using extended Kalman filter algorithm to research the AC induction motor. Firstly we should linear and discrete the state space model (1) and (2) of the AC induction motor. Specific process is as follows: we can expand the model using of the Taylor series mentioned in the previous section, and then ignore the higher order terms and retain only first order, get the approximation linear model. Assuming that the state equation of the discrete system is as follows:

$$\begin{aligned}\dot{x} &= A_kx(k)+B_ku(k)+W(k)\\ y &= C_kx(k)+V(k)\end{aligned} \tag{3}$$

The dispersion coefficient is :

$$A_k=e^{AT}\approx I+AT\ ,B_k=\int_0^T e^{A\xi}Bd\xi\approx BT\ ,C_k=C\ .$$

We substitute A, B, C into the above equation to get:

$$A_k=\begin{bmatrix} -\frac{L_m^2R_r+L_r^2R_s}{\sigma L_sL_r^2}T_s & 0 & \frac{L_mR_r}{\sigma L_sL_r^2}T_s & \frac{n_pL_m\omega_r}{\sigma L_sL_r^2}T_s & 0 \\ 0 & \frac{L_m^2R_r+L_r^2R_s}{\sigma L_sL_r^2}T_s & -\frac{n_pL_m\omega_r}{\sigma L_sL_r^2}T_s & \frac{L_mR_r}{\sigma L_sL_r^2}T_s & 0 \\ \frac{R_rL_m}{L_r}T_s & 0 & -\frac{R_r}{L_r}T_s & -n_p\omega_rT_s & 0 \\ 0 & \frac{R_rL_m}{L_r}T_s & n_p\omega_rT_s & -\frac{R_r}{L_r}T_s & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$B_k=\begin{bmatrix} \frac{1}{\sigma L_s} & 0 \\ 0 & \frac{1}{\sigma L_s} \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \qquad C_k=\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Known by the five formulas of the extended Kalman filter, we must also know the nonlinear state matrix $f$ , output matrix $h$ and vector partial differential matrix $F$ 、 $H$ . The $F$ is called the Jacobian matrix, and the $H$ is called the transfer matrix.

We can get $f=A_k\cdot x+B_k\cdot u$ from the formula $f(x(t))=A\cdot x(t)+B\cdot u(t)$ , and get $h(x(t))=C\cdot x(t)$ from $h=C_k\cdot x$ . So

$$F(\hat{x}(t))=\frac{\partial f(x(t))}{\partial x(t)}\Big|_{x(t)=\hat{x}(t)}$$

$$=\begin{bmatrix} -R_s/(\sigma L_s)-1/(\sigma T_r) & -\omega_r & R_r/L_r & \omega_r/(\sigma L_s) & -i_{\beta s}+\psi_{\beta r}/(\sigma L_s) \\ \omega_r & -R_s/(\sigma L_s)-1/(\sigma T_r) & -\omega_r/(\sigma L_s) & R_r/L_r & i_{\alpha s}-\psi_{\alpha r}/(\sigma L_s) \\ -R_s & 0 & 0 & 0 & 0 \\ 0 & -R_s & 0 & 0 & 0 \\ (-n_p^2/J)\psi_{\beta r} & (n_p^2/J)\psi_{\alpha r} & (n_p^2/J)i_{\beta s} & (-n_p^2/J)i_{\alpha s} & 0 \end{bmatrix}$$

$$H=\frac{\partial C}{\partial x(t)}\Big|_{x(t)=\hat{x}(t)}=\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

After calculating the above functions and the corresponding matrix, we can apply the five steps of the Extended Kalman Filter. Substituting into the following recursive formula:

$$x(k+1/k)=x(k)+T*f(x(k),u(k)) \tag{4}$$

$$P_{k+1/k}=P_k+(F(x(k))*P_k+P_k*F^T(x(k)))*T+Q \tag{5}$$

$$K(k+1)=P_{k+1/k}*H^T*[HP_{k+1/k}H^T+R]^{-1} \tag{6}$$

$$P_{k+1}=P_{k+1/k}-K(k+1)*H*P_{k+1/k} \tag{7}$$

$$x(k+1)=x(k+1/k)+K(k+1)(y(k+1)-H*x(k+1/k)) \tag{8}$$

Among the formulas, $K$ is the Kalman gain, $P$ is the state error covariance matrix, $Q$ and $R$ are noise covariance matrix whose initial value is artificially given. This article gets each parameter by trial and error method under the premise of ensuring the steady-state tracking and filtering divergence.

### B. System Simulation Model Based On The EKF-DTC System

When the simulation model is set up, the extended Kalman filter is using of the S function. Figure 2 is the Simulink. Simulation diagram with the EKF-DTC.

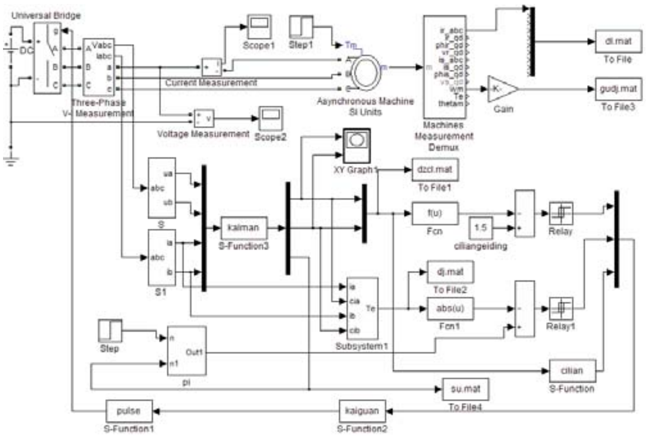


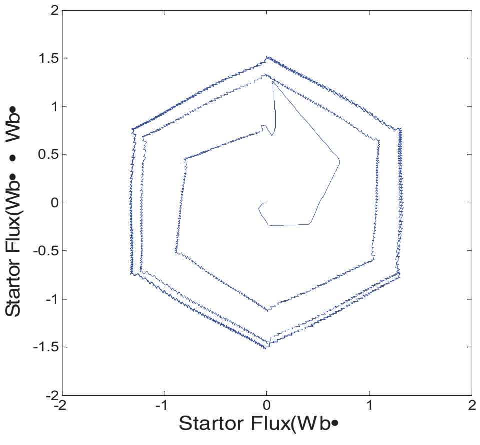
Figure 2. Simulink Simulation model based on EKF-DTC.

Induction motor simulation parameters:

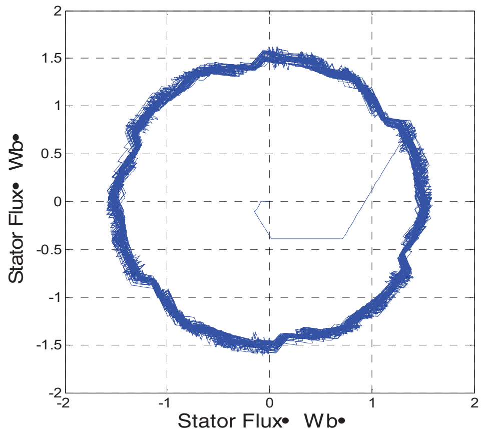
rated power $P_N$ =1.1kw, stator winding resistances $R_s$ =5.793Ω, rotor winding resistances $R_r$ =3.421Ω, stator cyclic inductances $L_s$ =0.368H, rotor cyclic inductances $L_r$ =0.368H,mutual cyclic inductance $L_m$ =0.363H,moment of inertia $J$ =0.0267Kg.m2,the number of pairs $P$ =2.

Based on the EKF-DTC method and the traditional DTC stator flux vector, electromagnetic torque and current simulation waveform, we compare the simulation results.

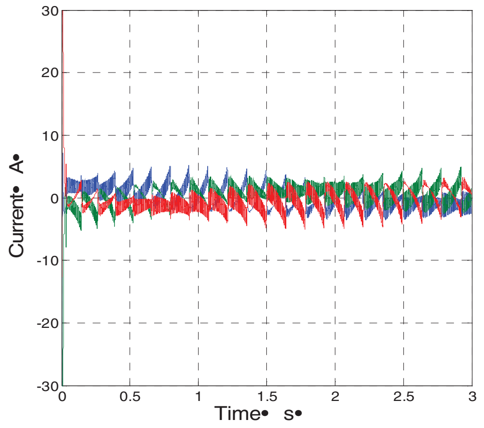
the given speed is 30 r / s (the rated speed is 150 r/s), and the load torque is 3 N. m.



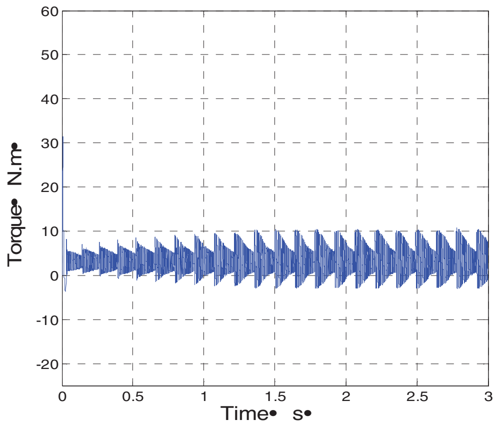
(a1) stator flux waveform of traditional DTC



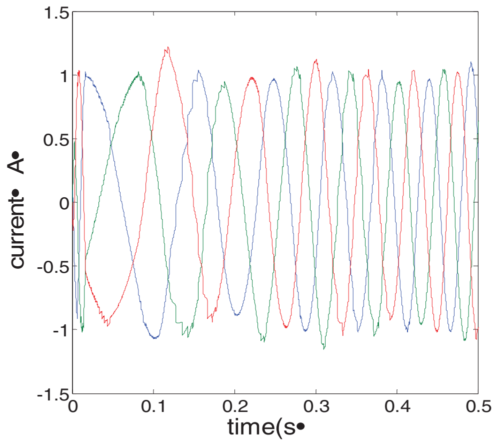
(a2) stator flux waveform of EKF-DTC



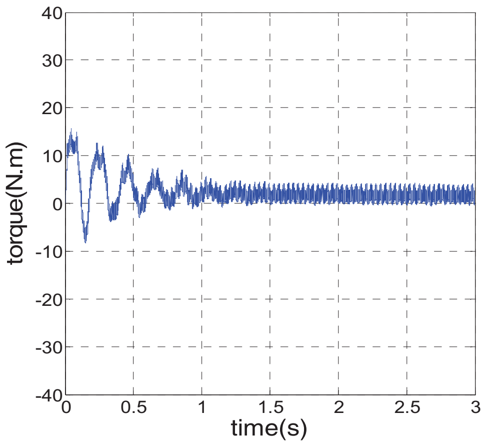
(c1) Current waveform of traditional DTC



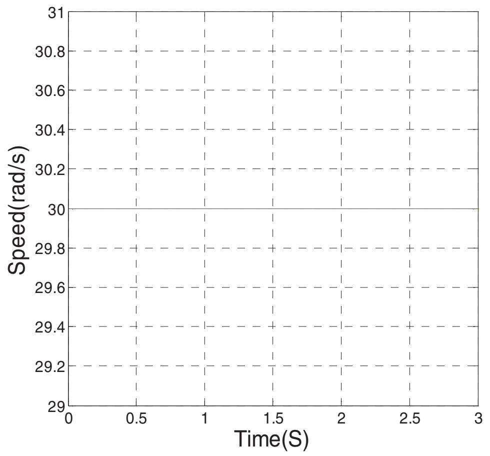
(b1) Electromagnetic torque waveform of traditional DTC



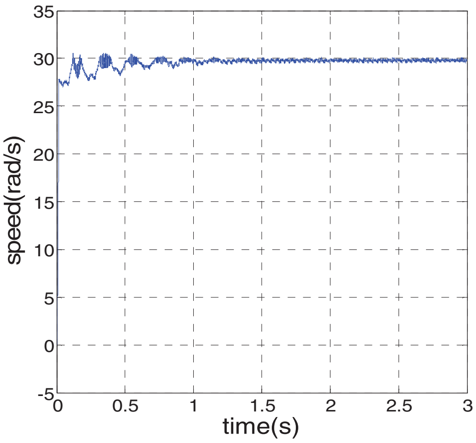
(c2) Current waveform of EKF-DTC



(b2) Electromagnetic torque waveform of EKF-DTC



(d1) Given speed waveform



(d2) speed waveform of EKF-DTC

Figure 3 Comparative experiment waveforms

## III. SIMULINK AND AMESIM SIMULATION REALIZATION AND VALIDATION OF THE ALGORITH

### A. The Structures Of The Joint Simulation Model

We use the motor model in AMESim to instead of the induction motor model of control system in front. At the same time its inverter drive section is also created in the software, to verify the effect of this control algorithm designed. In AMESim, the mathematical model of the induction motor takes the relationship between the temperature and the parameters into account, the system makes it closer to the actual motor system environment, and makes the powerful validation of the effectiveness of the proposed algorithm. In AMESim model, the relationship between the temperature and the motor parameters is as follows:

$$R_s = R_{s_0}(1 + alphaR_s(Temp - T_0))$$

$$tauR = T_{r_0}(1 + alphaT_r(Temp - T_0))$$

Co-simulation electronic control system and Simulink single software simulation model are different. In the co-simulation electronic control system, we introduce the voltage, current sensor signal and speed signal in AMESim into the Simulink model. We can calculate the inverse change switch signal by the computing in the Simulink model and then act on the inverter device of AMESim model. As shown in Figure 4, the S-function in the simulation model is corresponding to the AMESim co-simulation interface in figure 5, to complete the data reception and transmission.

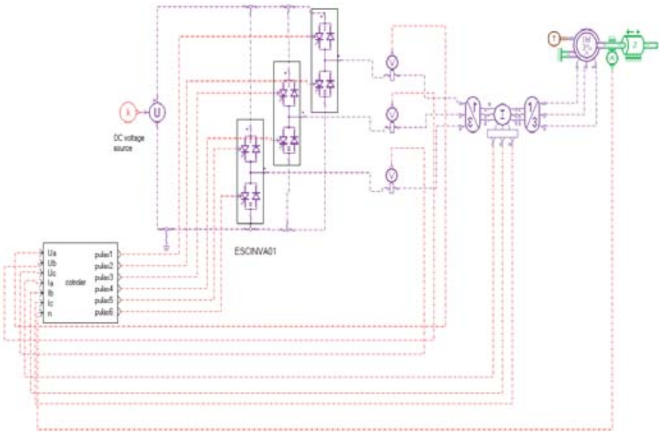


Figure 4 AMESim simulation model of Co-simulation

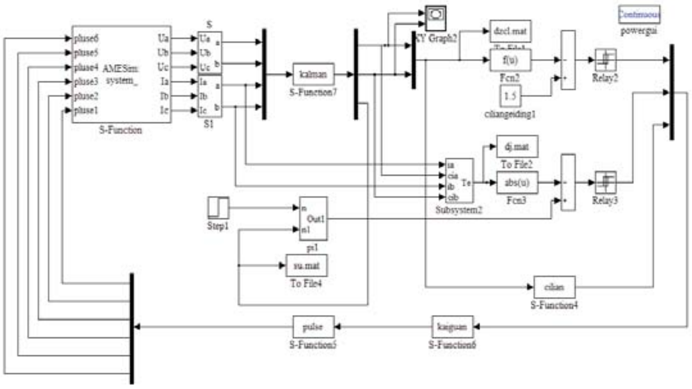


Figure 5 Simulink simulation model of Co-simulation

### B. The Simulation Experiment Of Algorithm Verification

Simulation: the motor parameters still choose the same parameters as the previous ones.

The experimental setup: the reference temperature is 25 degrees Celsius. The given speed is 30rad / s. Observing the stator flux and speed simulation waveform.

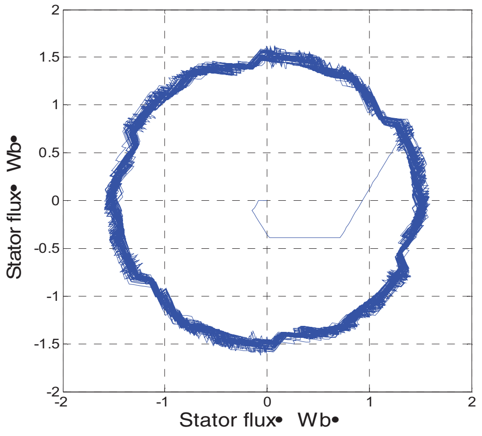


Figure 6. Stator flux waveform of Co-simulation

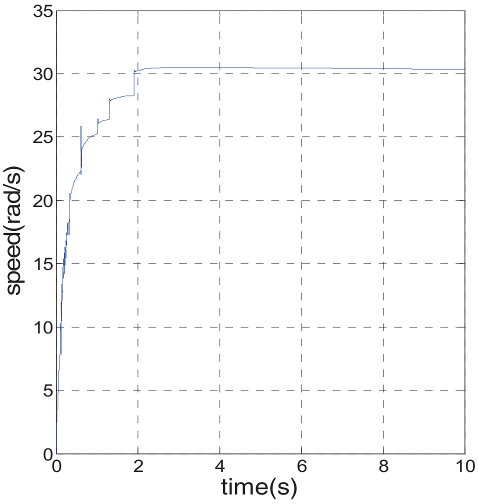


Figure 7. The waveform of the observables under Low-speed

## IV. Conclusion

Analyzing figure 3:

1. It is visible from the comparison between Figure (a1) and Figure (a2). When the motor is running at low speed 30rad / s, the stator flux trajectory is a hexagon in the Traditional DTC method. The control effect is far from ideal. But the stator flux trajectory is approximately circular in EKF-DTC method, and the pulse also significantly reduces.

2. It is visible from the comparison between Figure (b1) and Figure (b2). The electromagnetic torque ripple reduces to a large extent based on the EKF-DTC method.

3. It is visible from the comparison between Figure (c1) and Figure (c2). The current waveform pulse is obvious in traditional DTC method. The current sine wave has been significantly improved based on EKF-DTC method.

4. It is visible from the comparison between Figure (d1) and Figure (d2). The value of a given speed is 30rad / s. The EKF identification value is also about 30rad / s. That is to say that we can achieve the speed recognition based on the EKF-DTC and the no speed sensor technology.

It can clearly be seen that the method reduce the stator flux, electromagnetic torque, the current waveform pulsation to a large extent and improve the steady-state performance of a direct torque control system based on EKF-DTC. Analyzing the reason is that the stator flux observer model is the UI model in the traditional DTC, the accuracy declines seriously at a low speed. However the EKF can remove the error caused by the cumulative integral. The introduction of EFK is equivalent to change the original open-loop stator flux model into a closed-loop control, and eliminate the error in the ring. At the same time EKF achieve speed identification and the speed sensor control, and also maintain the advantages of the DTC fast dynamic response.

Analyzing Figure 6 and Figure 7

The stator flux trajectory is almost circular, but the trajectory is coarse, and it shows that the pulse is large. However, the shape and thickness of the track can still determine that it is limited within an acceptable range. The given speed is 30rad / s. It can be seen from the experimental results that the actual speed can track a given speed. Above all we can conclude that the design algorithm of this paper is effective.

## References

[1] M. Depenbrock, “Direct self-control of inverter-fed machine”. IEEE Trans. on PE, vol. 3, no. 4, pp. 420-429, 1988J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] I. Takahashi, T. Noguchi, “A new quick-response and high-efficiency control strategy of an induction motor”. IEEE Trans. on IA, vol. 22, no. 5, pp. 820-827, 1986.

[3] ANDERSON, B.O., and MOORE, J.B.: ‘Optimal filtering’ (Pren-tice-Hall, 1979).

[4] JAZINSKI, A.H.: ‘Stochastic processes and filtering theory’ (Academic Press, 1970).

[5] KIM, Y., SUL, S., and PARK, M.: ‘Speed sensor less vector con- troll of induction motor using extended Kalman filter’, IEEE Trans., 1994, IA-30, (5), pp. 1225-1233.

[6] KUBOTA, H, and MATSUSE, K.: ‘Speed sensorless field-ori-ented control of induction motor with rotor resistanc adapta- tion’, IEEE Trans., 1994, IA-30, (5), pp. 1219-1224

[7] HENNEBERGER, G., BRUNSBACH, B.J., and KLEPSCH, T.: ‘Field-oriented control of synchronous and asynchronous drives without mechanical sensors using a Kalman filter’. Proceedings of EPE, Firenze, Italy, 1991, Vol. 3, pp. 664671

[8] Zaky, M.S, Khater, M.M, Shokralla, S.S, Yasin, H.A Wide-Speed-Range Estimation With Online Parameter Identification Schemes of Sensorless Induction Motor Drives[J]. IEEE Transactions Industrial Electronics, 2009, 56(5): 1699-1707.

[9] Brandstetter, P. Kuchar, M. Vinklarek, D.Estimation Techniques for Sensorless Speed Control of Induction Motor Drive[J]. IEEE International Symposium Industrial Electronics, 2006, 1(7) :154-159.

[10] Ziane, H., Retif, J.M., Rekioua, T., “Fixed-switching-frequency DTC control for PM synchronous machine with minimum torque ripples”[J]. Canadian Journal of Electrical and Computer Engineering, 2008, (33) 3: 183–189.

[11] Dong Zhe, You Zheng. Anovel extended Kalman filter for a class of nonlinear systems[J]. Progress in Natural Science, 2006, 16(9) :912-918.

[12] Dong Zhe, You Zheng. Anovel extended Kalman filter for a class of nonlinear systems[J]. Progress in Natural Science, 2006, 16(9) :912-918.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Adaptive sliding mode control for spacecraft attitude maneuvers with reduced or eliminated reaching phase

Binglong Cong, Xiangdong Liu and Zhen Chen
Key laboratory for Intelligent Control & Decision of Complex Systems, School of Automation
Beijing Institute of Technology, Beijing, China 100081
Email: {cbl, xdliu, chenzhen76}@bit.edu.cn

***Abstract*—This paper aims to present an improved adaptive sliding mode control (ASMC) design for rigid spacecraft attitude maneuvers. An adaptive scheme is proposed for the switching gain calculation when the upper bound of the system uncertainty is unknown in advance. Unlike existing ASMC design, which may result in an over-adaptation of the upper bound when the initial system trajectory is located far from the sliding surface, this paper presents a novel ASMC strategy by introducing a decay term in the sliding function to reduce or eliminate the unrelated factor in the adaptation scheme. Consequently, a lower-chattering control signal is achieved. Simulation results are presented to illustrate the effectiveness of the proposed strategy.**

***Index Terms*—attitude maneuver, adaptive sliding mode control, over-adaptation, global sliding mode, chattering suppression.**

## I. Introduction

As a subclass of variable structure control systems, sliding mode control (SMC) is a nonlinear control method that is well known for its robust performance. In the pase decades, SMC has been extensively studied in many practical control systems. SMC can offer many good properties, such as insensitivity to parameter variation, external disturbance rejection, and fast dynamic response, which make it a potential approach for spacecraft attitude control. In [1], the attitude regulation problem was studied and the sliding function was determined by solving an optimal control problem. A smoothing model-reference SMC algorithm was presented in [2], where a desired quaternion error response was predefined for the attitude control system. To reduce the static error, an integral term was added in the sliding function and modified Rodrigues parameters (MRPs) were used instead of quaternion for the non-redundancy in [3]. Moreover, in [4], a nonlinear sliding function was defined according to the properties related to the attitude kinematics.

However, for the SMC design mentioned above, a prior knowledge of the system uncertainty upper bound is required. When such a bound is unavailable in advance, conservative method is generally adopted, where the switching gain is selected sufficiently large. It is well known that the chattering level is directly determined by the switching gain. Hence, such a conservative method may aggravate the chattering problem which could excite the unmodelled dynamics and may lead to instability. Further investigations have proceeded along two lines. On the one hand, technologies are studied for the chattering reduction. Higher-order sliding mode control has been recently proposed to reduce the chattering problem while keeping the main advantages of conventional SMC ( [5]). In [6], the control chattering is reduced by low-pass filtering the control signal. In particular, in [7], three methods were presented for the chattering suppression. Nonetheless, there are no constructive conditions for the switching gain selection in those algorithms and generally a prior knowledge of the bound of the system uncertainty and/or the system states is needed.

On the other hand, attention has also been focused on eliminating the requirement of the prior knowledge of the uncertainty bound. One way is using the disturbance observer (DOB) technique, as suggested in [8] and [9]. However, the DOB based SMC algorithms usually assume that the model uncertainty is generated by a linear exogenous system [10], which is hard to satisfy due to the complexity and unpredictability of the uncertainty. The other effective approach is to integrate adaptive scheme into SMC designs. By updating the switching gain adaptively, the upper bound of model uncertainty is not required to be known in advance. At the first stage, it is generally assumed that the norm of uncertainty was bounded by a linear function of the state-norm. Correspondingly, adaptive laws were designed for the linear function parameters, as suggested in [11], [12], [13]. In particular, in [13], an ASMC algorithm was proposed for the attitude stabilization of a rigid spacecraft, where the lumped uncertainty is assumed to be bounded by a linear function of the norms of angular velocity and quaternion. Afterwards, in [14] the lumped uncertainty was assumed to be bounded by an unknown constant and consequently a simple adaptive law was proposed for the switching gain calculation. Subsequent results can be found in many other applications such as internal combustion engines ( [15]), induction servomotor ( [16]), planetary gear-type inverted-pendulum ( [17]), etc. The major problem of the ASMC algorithms mentioned above is their over-adaptation for the switching gain with respect to the uncertainty bound, which results in the serious chattering phenomena and unnecessary energy consumption.

Considering the shortcomings of the chattering suppression techniques and current ASMC design, it is necessary to put

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

forward a novel ASMC strategy, which does not need a prior knowledge of the upper bound and has the chattering suppression ability. Such a problem was recently investigated in [18], where two new methodologies for the ASMC design were proposed. Unfortunately,those two methods only reduce the switching gain in the sliding phase without considering the reaching phase. Similar to the ASMC design mentioned before, which will be called the ASMC-I algorithm in the following contents, the switching gain is also overestimated in [18] due to its failure in accounting for the initial system error in the adaptation scheme. With this in mind, this paper tries to present an improved ASMC design principle by addressing the attitude control problem for a rigid spacecraft in the presence of inertia matrix uncertainty and external disturbance. The key feature of the proposed ASMC algorithm, referred as the ASMC-II algorithm, is that a decay function is introduced to reduce or eliminate the impact of initial system error in the adaptation scheme. By this modification, the ASMC-II algorithm can give a more accurate estimation of the uncertainty bound and generate a lower-chattering control signal. A large angle attitude reorientation maneuver is employed in the simulation, where the simulation results demonstrate the effectiveness of the proposed strategy.

## II. Preliminary and problem statement

Rigid spacecraft attitude control for large angle maneuvers poses a difficult problem, including nonlinear characteristics in both the dynamics and the kinematics, modelling uncertainty, and persistent external disturbance. In this paper, we will resolve this problem in the ASMC framework. Before moving on, some notations and assumptions are presented here. Three coordinate frames are used in this paper, which are the inertia reference frame $\mathcal{F}_I$, body-fixed frame $\mathcal{F}_B$, and the desired frame $\mathcal{F}_D$. Unless otherwise specified, all the quantities are expressed in their corresponding frames. And it is assumed that the spacecraft attitude and angular velocity are available and the dynamics of actuator is neglected.

### *A. Mathematical Model*

Consider a thruster control rigid spacecraft, whose dynamics is described as follows:

$$J\dot{\omega}_b + \omega_b^\times J\omega_b = T_b + T_d \tag{1}$$

where $J \in \mathbb{R}^{3\times3}$ is the spacecraft inertia matrix, $\omega_b = [\omega_{b1} \quad \omega_{b2} \quad \omega_{b3}]^T \in \mathbb{R}^3$ is the angular velocity vector of $\mathcal{F}_B$ with respect to $\mathcal{F}_I$. The superscript $(\cdot)^\times$ on $\omega_b$ denotes the skew-symmetric matrix operator performing the cross product between two vectors, e.g.,

$$\alpha^\times\beta = \alpha\times\beta \tag{2}$$

where $\alpha$ and $\beta$ are two vectors in $\mathbb{R}^3$. $T_b = [T_{b1} \quad T_{b2} \quad T_{b3}]^T \in \mathbb{R}^3$ is the vector of control torque provided by the thrusters, $T_d = [T_{d1} \quad T_{d2} \quad T_{d3}]^T \in \mathbb{R}^3$ is the external disturbance vector, including gravitational torque, aerodynamic torque, radiation torque, and other environmental and non-environmental torques. Furthermore, the inertia matrix uncertainty is taken into account. Let $J = \hat{J} + \Delta J$ with $\Delta J$ the uncertainty caused by the change in mass properties and $\hat{J} = \mathrm{diag}(J_1, J_2, J_3)$ the nominal inertia matrix. Then the attitude dynamics is given by:

$$\hat{J}\dot{\omega}_b + \omega_b^\times\hat{J}\omega_b = T_b + T_d - \Delta J\dot{\omega}_b - \omega_b^\times\Delta J\omega_b \tag{3}$$

According to the structural feature in (3), one can merge all the elements caused by inertia matrix uncertainty and external disturbance as the lumped uncertainty, i.e., let $d = [d_1 \quad d_2 \quad d_3]^T \in \mathbb{R}^3$ with $d = T_d - \Delta J\dot{\omega}_b - \omega_b^\times\Delta J\omega_b$. Correspondingly, the attitude dynamics is rewritten as:

$$\hat{J}\dot{\omega}_b + \omega_b^\times\hat{J}\omega_b = T_b + d \tag{4}$$

From (4), it is clear that the lumped uncertainty is matched to the system. Without loss of generality, it is assumed that $d$ is bounded by an unknown upper bound, e.g., $\|d\|_\infty < d_{\max}$ with $\|\cdot\|_\infty$ the vector infinite-norm.

By introducing the shadow MRPs, the MRPs set can provide a nonsingular, bounded, minimal attitude description. Hence, MRPs are utilized in this paper instead of quaternion, whose kinematics is:

$$\dot{\sigma}_b = M(\sigma_b)\omega_b \tag{5}$$

where $\sigma_b = [\sigma_{b1} \quad \sigma_{b2} \quad \sigma_{b3}]^T \in \mathbb{R}^3$ denotes the inertial MRPs vector of $\mathcal{F}_B$ with respect to $\mathcal{F}_I$. $M : \mathbb{R}^3 \to \mathbb{R}^{3\times3}$ such that $M(\sigma_b)$ is the Jacobian matrix with $M(\sigma_b) = \dfrac{(1-\|\sigma_b\|^2)I_3 + 2\sigma_b^\times + 2\sigma_b\sigma_b^T}{4}$, $\|\cdot\|$ is the vector 2-norm and $I_3$ is the $3\times3$ identity matrix. Moreover, $M^T(\sigma_b)M(\sigma_b) = m(\sigma_b)I_3$ with $m : \mathbb{R}^3 \to \mathbb{R}$ such that $m(\sigma_b) = (1 + \|\sigma_b\|^2)^2/16$. The transition matrix from $\mathcal{F}_I$ to $\mathcal{F}_B$ in terms of MRPs is given by:

$$R(\sigma_b) = I_3 + \frac{8\sigma_b^\times\sigma_b^\times - 4(1-\|\sigma_b\|^2)\sigma_b^\times}{(1+\|\sigma_b\|^2)^2} \tag{6}$$

A typical Rest-to-Rest attitude maneuver is studied in this paper. The objective is reorienting the spacecraft from an arbitrary stationary attitude to a desired attitude with zero angular velocity. The attitude variables of the desired frame, $\mathcal{F}_D$, are denoted by $\sigma_d \in \mathbb{R}^3$ and $\omega_d \in \mathbb{R}^3$. Then, the error attitude variables are defined as follows:

$$\sigma_e = \sigma_b \oplus \sigma_d^* \tag{7}$$

$$\omega_e = \omega_b - R(\sigma_e)\omega_d \tag{8}$$

where $\sigma_e = [\sigma_{e1} \quad \sigma_{e2} \quad \sigma_{e3}]^T \in \mathbb{R}^3$, $\oplus$ is the MRPs production operator characterizing the successive rotations. For two MRPs expressed in their corresponding frames, e.g., $\sigma_1 \in \mathbb{R}^3$ and $\sigma_2 \in \mathbb{R}^3$, it is operated as follows:

$$\sigma_1 \oplus \sigma_2 = \frac{(1-\|\sigma_2\|^2)\sigma_1 + (1-\|\sigma_1\|^2)\sigma_2 - 2\sigma_1^\times\sigma_2}{1 + \|\sigma_2\|^2\|\sigma_1\|^2 - 2\sigma_2^T\sigma_1} \tag{9}$$

$\sigma_d^*$ is the inverse of $\sigma_d$, which is extracted from the inverse of $R(\sigma_d)$ and $\sigma_d^* = -\sigma_d$. $R(\sigma_e)$ and $R(\sigma_d)$ are the transition matrices from $\mathcal{F}_D$ to $\mathcal{F}_B$ and from $\mathcal{F}_I$ to $\mathcal{F}_D$, their expressions can be obtained by replacing $\sigma_b$ by $\sigma_e$ and $\sigma_d$ in (6). As $\omega_d = 0$, $\omega_e = \omega_b$. Therefore, the error attitude dynamics is expressed same as (4). With respect to the error attitude kinematics, following lemma is introduced.

*Lemma 1:* If the attitude variables pairs $(\sigma_b, \omega_b)$ and $(\sigma_d, \omega_d)$ satisfy the MRPs kinematics formulation described in (5), then the error attitude variables pair $(\sigma_e, \omega_e)$ also satisfies that MRPs kinematics formulation.

*Proof:* The proof is based on the successive rotations in terms of transition matrix. See [20] for further details. ■

Then, the system is governed by the following equations:

$$\begin{cases} \hat{J}\dot{\omega}_b &= T_b + d - \omega_b^{\times}\hat{J}\omega_b \\ \dot{\sigma}_e &= M(\sigma_e)\omega_b \end{cases} \tag{10}$$

### B. Problem Statement

Our aim can be summarized as follows: find a SMC algorithm to steer the attitude variables pair $(\sigma_b, \omega_b)$ from $(\sigma_b(0), 0)$ to $(\sigma_d, 0)$ in the presence of the lumped uncertainty, and find an adaptive law to update the estimation of the unknown $d_{\max}$ for the switching gain calculation which has the chattering suppression ability.

## III. Main Results

### A. AMSC-I Algorithm Design

In this section, we will briefly apply the ASMC-I algorithm for the attitude control problem under consideration. First, the sliding function defined in [4] is given by:

$$S = \omega_b + \lambda \frac{M^T(\sigma_e)}{m(\sigma_e)}\sigma_e \tag{11}$$

where $S = [s_1 \quad s_2 \quad s_3]^T \in \mathbb{R}^3$ and the corresponding sliding surface is determined by $S = 0$, $\lambda > 0$ is the sliding function gain. In the following derivations, $M(\sigma_e)$ and $m(\sigma_e)$ will be denoted by $M$ and $m$ for clarity.

By a left multiplication of (11) with $M$ and using the fact that $M^T M = mI_3$, one has:

$$MS = \dot{\sigma}_e + \lambda\sigma_e \tag{12}$$

When the sliding mode occurs, i.e., $S = 0$ holds, it is easy to conclude that an exponential convergence of the error MRPs, i.e., $\sigma_e(t) = \mathrm{e}^{-\lambda(t-t_r)}\sigma_e(t_r)$, can be obtained if a proper SMC law is designed, where $t_r$ is the time of arrival at the sliding surface. Such a SMC algorithm can be derived by producing a negative definite derivative of the following Lyapunov function:

$$V = \frac{1}{2}S^T\hat{J}S \tag{13}$$

On the basis of [4], one can get the following SMC algorithm:

$$T_b = \omega_b^{\times}\hat{J}\omega_b - \lambda\hat{J}\frac{\left(4M - 2\sigma_e\sigma_e^T\right)\omega_b}{1 + \|\sigma_e\|^2} - \Gamma\mathrm{sgn}(S) \tag{14}$$

where $\Gamma = \mathrm{diag}(\gamma_1, \gamma_2, \gamma_3)$ is the switching gain matrix with its elements $\gamma_i > d_{\max}$ $(i = 1, 2, 3)$ to guarantee the system stability and $\mathrm{sgn}(\cdot)$ is the sign function.

During the above derivations, the sliding function gain $\lambda$ can be determined according the desired system response in the sliding phase. In order to determine the switching gain in the absence of a prior knowledge of $d_{\max}$, the ASMC-I algorithm can be applied.

Consider the modified Lyapunov function:

$$V = \frac{1}{2}S^T\hat{J}S + \frac{1}{2c}\tilde{d}^2 \tag{15}$$

where $c > 0$ is the adaptive gain, $\tilde{d} = \hat{d} - d_{\max}$ is the estimation error with $\hat{d}$ the estimation of $d_{\max}$. According to the ASMC-I design principle in [14], the adaptive switching gain law for the attitude control is:

$$\hat{d} = c\int_0^t \|S\|_1 d\tau \tag{16}$$

where $\|\cdot\|_1$ is the vector 1-norm.

Correspondingly, the ASMC-I algorithm is given by:

$$T_b = \omega_b^{\times}\hat{J}\omega_b - \lambda\hat{J}\frac{\left(4M - 2\sigma_e\sigma_e^T\right)\omega_b}{1 + \|\sigma_e\|^2} - \hat{d}\mathrm{sgn}(S) \tag{17}$$

### B. Over-adaptation in ASMC-I Algorithm

From the ASMC-I algorithm design in Section III-A, one can see that the ASMC-I algorithm is based on conventional SMC algorithm. It is well known that the system trajectory employing the SMC algorithm consists two parts, the reaching phase and the sliding phase, as illustrated in Fig. 1.

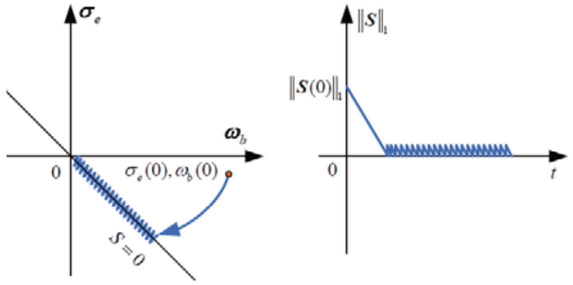


Fig. 1. System trajectory and sliding function response using SMC algorithm in (14)

Recalling the adaptive law in (16), it is obvious that the basic idea of the ASMC-I technique lies in that $d_{\max}$ can be adjusted by the deviation from the sliding surface. From (16), the integral action starts from the very beginning and any departure from the sliding surface, $S = 0$, will result in the increase of the switching gain. In other words, the switching gain adaptation depends on the initial value of the sliding function besides the lumped uncertainty. However, the initial system trajectory is generally located far from the sliding surface as shown in Fig. 1. Hence, $\hat{d}$ would increase quickly at the beginning due to a large $\|S(0)\|_1$ and the resulting $\hat{d}$ is much larger than $d_{\max}$, which leads to an over-adaptation of the switching gain and correspondingly the serious chattering problem in the ASMC-I algorithm.

Moreover, if we divide the adaptive law in (16) into two parts, i.e.,

$$\hat{d} = c\int_{0}^{t_r} \|S\|_1 d\tau + c\int_{t_r}^{t} \|S\|_1 d\tau \tag{18}$$

Then, it is obvious that the first integral term in (18) deals with the deviation in the reaching phase, which is mainly caused by the initial system error; while the second term handles the departure in the sliding phase, which is mainly affected by the lumped disturbance.

### *C. ASMC-II Algorithm Design*

To address the over-adaptation problem in the ASMC-I design, it is natural to reduce or eliminate the proportion of the first integral term in (18). With this in mind, we present the ASMC-II algorithm. First, the sliding function in (11) is modified as

$$S(t) = \omega_b - f(t)\xi + \lambda\frac{M^T}{m}\left[\sigma_e - f(t)\rho\right] \tag{19}$$

where $S(t) = [s_1(t) \quad s_2(t) \quad s_3(t)]^T \in \mathbb{R}^3$, $f(t)$ is a continuous, strictly decreasing function on $t \in [0,\infty)$ with its initial value $f(0) \in [0,1]$ and its final value $f_f = 0$, $\xi = [\xi_1 \quad \xi_2 \quad \xi_3]^T \in \mathbb{R}^3$, $\rho = [\rho_1 \quad \rho_2 \quad \rho_3]^T \in \mathbb{R}^3$ are the coefficients related to the initial system states, and $\rho = \sigma_e(0)$, $\xi = \omega_b(0)$.

According to the above definition, we can find that the initial value of the sliding function is reduced to a small value or even becomes zero by the additional decay function. Therefore, if a proper SMC algorithm is designed to achieve the sliding motion, the sliding surface related to (19) is a new kind of sliding surface, which is illustrated in Fig. 2.

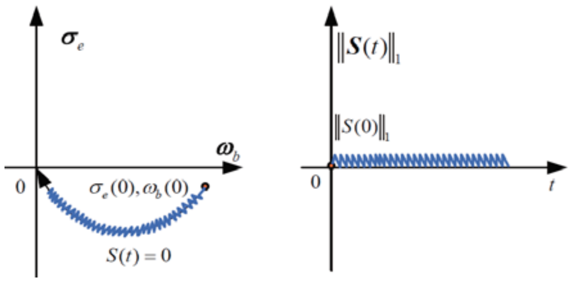


Fig. 2. System trajectory and sliding function response using sliding function in (19)

Therefore, if the sliding function in (19) is used for the switching gain adaptation, the unrelated effect of the initial system error can be reduced or eliminated in the adaptation scheme and the upper bound of the lumped uncertainty can be estimated more precisely, which is the motivation for the ASMC-II algorithm.

*Remark 1:* Actually, the sliding function defined in (19) is an extension of the time-varying sliding function investigated in [21] and [22]. In particular, in [21], it was proved that a *global sliding mode* would be achieved by using the sliding function like (19) with $f(t)$ selected as the exponential decay function $f(t) = \mathrm{e}^{-\kappa t}$ and $\kappa > 0$. However, due to the sensor noise, the initial system error can not be entirely cancelled. Furthermore, if the total time-varying sliding mode case, i.e., $S(0) = 0$, is used for the switching gain adaptation, the system cannot provide enough information for the adaptation due to the fact that there is no departure from the sliding function at the initial time, which will slow down the adaptation procedure. Therefore, in this paper, we introduce a weight in the decay function, e.g., let $f(t) = p\mathrm{e}^{-\kappa t}$ with $p \in [0,1]$ in the following derivations, which will produce an initial departure from the sliding surface purposefully to speed up the adaptive process. When $p = 0$, the sliding function in (19) turns out to be the sliding function in (11); whereas if $p = 1$, it becomes the total time-varying sliding function studied in [21].

For the attitude reorientation control problem, we have $\omega_b(0) = 0$, i.e., $\xi = 0$. For simplicity, the $\kappa$ in the decay function is selected same as the sliding function gain $\lambda$, i.e., let $\kappa = \lambda$. Then the sliding function is specified as:

$$S(t) = \omega_b + \lambda\frac{M^T}{m}\left(\sigma_e - p\mathrm{e}^{-\lambda t}\rho\right) \tag{20}$$

Then, we are ready to present the following theorem:

*Theorem 1:* For the system governed by (10), by adopting the sliding function in (20) and the ASMC-II algorithm in (21), the system trajectory will converge to the sliding function as $t \to \infty$.

$$\begin{cases} T_b = & \lambda p\mathrm{e}^{-\lambda t}\hat{J}\frac{\mathrm{d}}{\mathrm{dt}}\left(\frac{M^T\rho}{m}\right) - \lambda\hat{J}\frac{\left(4M - 2\sigma_e\sigma_e^T\right)\omega_b}{1+\|\sigma_e\|^2} \\ & +\omega_b^\times\hat{J}\omega_b - \lambda^2 p\mathrm{e}^{-\lambda t}\hat{J}\frac{M^T\rho}{m} - \hat{d}\mathrm{sgn}(S(t)) \\ \hat{d} = & c\int_0^t \|S(t)\|_1 d\tau \end{cases} \tag{21}$$

where

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{dt}}\left(\frac{M^T\rho}{m}\right) =& 8\frac{\sigma_e^T\rho M\omega_b - (\rho - \mu + \epsilon)^\times M\omega_b}{(1+\|\sigma_e\|^2)^2} \\ &- \frac{4M^T\rho\sigma_e^T M\omega_b}{(1+\|\sigma_e\|^2)m} \end{aligned}$$

where $\mu = [\sigma_{e3}\rho_2 \quad \sigma_{e1}\rho_3 \quad \sigma_{e2}\rho_1]^T \in \mathbb{R}^3$ and $\epsilon = [\sigma_{e2}\rho_3 \quad \sigma_{e3}\rho_1 \quad \sigma_{e1}\rho_2]^T \in \mathbb{R}^3$.

*Proof 1:* Considering the following Lyapunov function:

$$V = \frac{1}{2}S^T(t)\hat{J}S(t) + \frac{1}{2c}\tilde{d}^2 \tag{22}$$

The time derivative of the above Lyapunov function along the closed-loop system trajectory is:

$$\begin{aligned} \dot{V} =& S^T(t)\hat{J}\dot{S}(t) + \frac{\hat{d} - d_{\max}}{c}\dot{\hat{d}} \\ =& S^T(t)\left(d - \hat{d}\mathrm{sgn}(S(t))\right) + (\hat{d} - d_{\max})\|S(t)\|_1 \\ =& S^T(t)d - \hat{d}\|S(t)\|_1 + \hat{d}\|S(t)\|_1 - d_{\max}\|S(t)\|_1 \\ =& -\sum_{i=1}^{3}\left(d_{\max}|s_i(t)| - d_i s_i(t)\right) \le 0 \end{aligned}$$

Let $\chi = \sum_{i=1}^{3}(d_{\max}|s_i(t)| - d_i s_i(t))$ and it is obvious that $\chi$ is uniformly continuous. By integrating the above equation from zero to $t$, one has:

$$\int_0^t \dot{V} d\tau \leq -\int_0^t \chi d\tau \Rightarrow V(0) \geq \int_0^t \chi d\tau \tag{23}$$

Taking the limits as $t \to \infty$ on both sides of (23) gives:

$$\infty > V(0) \geq \lim_{t\to\infty} \int_0^t \chi d\tau \tag{24}$$

On the basis of Barbalat lemma, one can obtain

$$\lim_{t\to\infty} \chi = 0 \tag{25}$$

which implies that $\lim_{t\to\infty} S(t) = 0$.

*Remark 2:* Above proof implies that the ASMC-II algorithm can only guarantee the asymptotic stability of the sliding function but not the attitude variables, i.e., the system trajectory will converge to the sliding surface in infinite time. However, from the adaptation law in (21), one can see that the switching gain $\hat{d}$ will keep increasing if $S(t) \neq 0$. When $\hat{d}$ increases up to a value large enough to suppress the lumped uncertainty, e.g., $\hat{d} > d_{\max} + \delta$ with $\delta$ a sufficiently small positive scalar, the sliding mode will start in finite time. Similarly, denote the arrival time as $t_r$. By a left multiplication of (20) with $M$, following 3-dimensional first-order vector differential equation can be obtained:

$$\dot{\sigma}_e + \lambda\sigma_e = \lambda p e^{-\lambda t}\rho \tag{26}$$

The analytical solution for $\sigma_e$ is:

$$\sigma_e(t) = \mathrm{e}^{-\lambda(t-t_r)}(\lambda p\rho t + \sigma_e(t_r)) \tag{27}$$

It is obvious $\lim_{t\to\infty} \sigma_e(t) = 0$ and $\lim_{t\to\infty} \omega_b(t) = 0$ from (20). Hence, the attitude control system in (10) with the ASMC-II algorithm in (21) is globally asymptotically stable.

*Remark 3:* Recalling the adaptive law in (21), $\hat{d}$ will become unbounded due to the fact that the sliding function is not identically equal to zero, which may be caused by the finite switching frequency or measurement noise. For implementation in practice, the adaptive law has to be modified to get a bounded switching gain, such as the $\sigma$ modification method in [12]. Here, the approach proposed in [18] will be used, where the adaptive law in (21) is modified as

$$\hat{d} = \begin{cases} c\int_0^t \|S(t)\|_1 \mathrm{sgn}(\|S(t)\|_1 - \eta) d\tau & \text{if} \quad \hat{d} > \varrho \\ \int_0^t \varrho d\tau & \text{if} \quad \hat{d} \leq \varrho \end{cases} \tag{28}$$

where $\varrho > 0$ is a very small scalar to ensure $\hat{d}$ is positive and $\eta > 0$ is carefully chosen to deal with the trade-off in control accuracy and bounded switching gain. Further details on $\eta$ tuning can refer to [18].

## IV. Numerical Simulation

In this section, a comparison of the ASMC-I and ASMC-II algorithms is employed for a large angle attitude maneuver to test the effectiveness of the proposed strategy.

The inertia matrix for the controller design is given by $\hat{J} = \mathrm{diag}(48, 25, 61.8)$ (kg.m) and the uncertainty is 10% of the nominal value. The external disturbance is $T_d = [0.02\sin(0.01t) \quad 0.02\cos(0.01t) \quad 0.04\sin(0.01t)]^T$ (N.m). The initial attitude variables are $\sigma_b(0) = [-0.2 \quad 0.3 \quad 0.1]^T$ and $\omega_b(0) = [0 \quad 0 \quad 0]^T$ (rad/s). The desired attitude is $\sigma_d = [0.1 \quad 0.2 \quad -0.3]^T$ with the desired angular velocity $\omega_d = [0 \quad 0 \quad 0]^T$ (rad/s). For comparison, same control parameters are selected for both the ASMC-I algorithm and the ASMC-II algorithm, where $c = 1$, $\lambda = 0.25$ and the initial value of $\hat{d}$ is zero. The weight $p$ is selected as 0.8. The simulation results are shown in Fig.3–Fig.6.

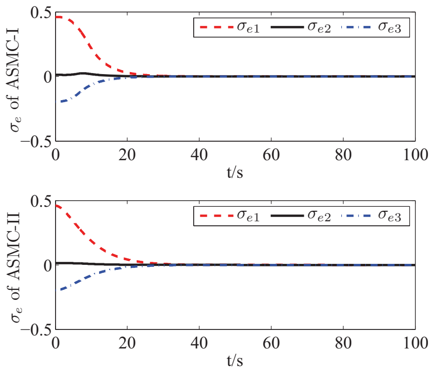


Fig. 3. Error MRPs responses of ASMC-I and ASMC-II

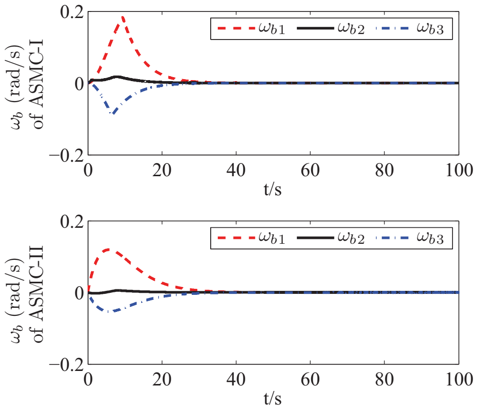


Fig. 4. Angular velocities of ASMC-I and ASMC-II

The maneuver evolutions controlled by ASMC-I and ASMC-II are compared in Fig. 3 and Fig. 4 with the corresponding control torque commands shown in Fig. 5. From Fig. 3, it is clear that both the ASMC-I and the ASMC-II can accomplish the attitude reorientation in the absence of the

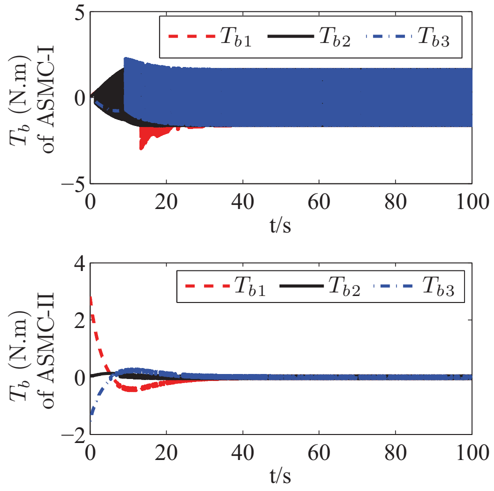


Fig. 5. Control torque commands of ASMC-I and AMSC-II

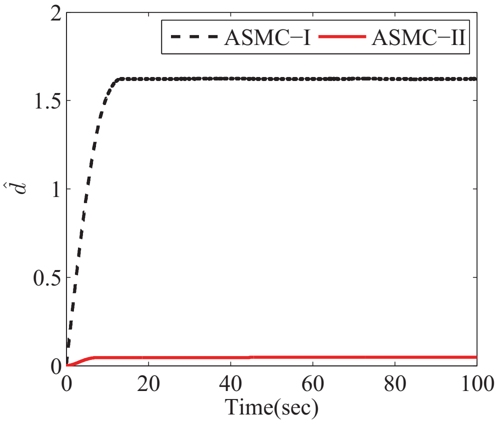


Fig. 6. Adaptive switching gains of ASMC-I and ASMC-II

prior knowledge of $d_{\max}$ and the system responses are similar. Fig. 4 shows the angular velocity response comparison, where the angular velocity controlled by the ASMC-II is smoother than that controlled by the ASMC-I. Nonetheless, there is a significant difference in the control torque commands as shown in Fig. 5. According to Fig. 5, it is obvious that the chattering in ASMC-I is much more serious than that in ASMC-II, which verified the effectiveness of the proposed strategy. Moreover, as shown in Fig. 6, the adaptive switching gain generated by ASMC-II is much smaller than the ASMC-I case, where $\hat{d} \approx 1.62$ for ASMC-I and $\hat{d} \approx 0.046$ for ASMC-II, which verified the chattering suppression ability of ASMC-II.

## V. Conclusion

The attitude control problem of a rigid spacecraft involving inertia matrix uncertainty and external disturbance has been considered. An effective solution has been presented to address the over-adaptation problem in current ASMC design. Such an improvement is achieved by reducing or eliminating the influence caused by initial system error on the switching gain adaptation. It has been shown by theoretical analysis and simulation results that the proposed strategy can produce a much smaller switching gain as compared with current ASMC design and achieve a smoother system response.

## References

[1] S.R. Vadali, “Variable structure control of spacecraft large-angle maneuvers”, *J. Guidance*, vol. 9, no. 2, pp. 235-239, 1986.

[2] S.C. Lo, Y.P. Chen, “Smooth sliding-mode control for spacecraft attitude tracking maneuvers”, *Journal of Guidance, Control, and Dynamics*, vol 18, no. 6, pp. 1345-1349, 1995.

[3] S.A. Kowalchuk, C.D. Hall, “Spacecraft attitude sliding mode controller using reaction wheels”, in *AIAA/AAS Astrodynamics Specialist Conference and Exhibit*, Honolulu, Hawaii, AIAA 2008-6260, 2008.

[4] J.L. Crassidis, “Sliding mode control using modified Rodrgiues parameters”, *Journal of Guidance, Control, and Dynamics*, vol. 19, no. 6, pp. 1381-1383, 1996.

[5] A. Levant, “Higher-order sliding modes, differentitaion and output-feedback control”, *International Journal of Control*, vol. 76, no. 9/10, pp. 924–941, 2003.

[6] M.L. Tseng, M.S. Chen, “Chattering reduction of sliding mode control by low-pass filtering the control signal”, *Asian Journal of Control*, vol. 12, no. 3, pp. 392–398, 2010.

[7] H. Lee, V. I. Utkin, “Chattering suppression methods in sliding mode control systems”, *Annual Reviews in Control*, vol. 31, no. 2, pp. 179–188, 2007.

[8] S.P. Chan, “An approach to perturbation compensation for variable structure systems”, *Automatica*, vol. 32, no. 3, pp. 469-473, 1996.

[9] M. Chen, W.H. Chen, “Sliding mode control for a class of uncertain nonlinear system based on disturbance observer”, *International Journal of Adaptive Control and Signal Proceesing*, vol. 24, no. 1, pp. 51-64, 2010.

[10] W.H. Chen, “Disturbance observer based control for nonlinear systems”, *IEEE/ASME Transactions on Mechatronics*, vol. 9, no. 4, pp. 706-710, 2004.

[11] D.S. Yoo, M.J. Chung, “A variable structure control with simple adaptation laws for upper bounds on the norm of the uncertainties”, *IEEE Transactions on Automatic Control*, vol. 7, no. 6, pp. 860–865, 1992.

[12] G. Wheeler, C.Y. Su, Y. Stepanenko, “A sliding mode controller with improved adaption laws for the upper bounds on the norm of uncertainties”, in *Proc. 1996 IEEE Workshop on Variable Structure Systems*, Tokyo, Japan, pp. 154–159, 1996.

[13] Z. Zhu, Y.Q. Xia, and M.Y. Fu, “Adaptive sliding mode control for attitude stabilization with actuator saturation”, *IEEE Transaction on Industrial Electronics*, vol. 58, no. 10, pp. 4898–4907, 2011.

[14] F.J. Lin, S.L. Chiu, “Novel sliding mode controller for synchronous motor drive”, *IEEE Transactions on Aerospace and Electronic Systems*, vol. 34, no. 2, pp. 532–542, 1998.

[15] J.S. Souder, J.K. Hedrick, “Adaptive sliding mode control of air-fuel ratio in internal combustion engines”, *Int. J. Robust Nonlinear Control*, vol. 14, no. 6, pp. 525-541, 2004.

[16] R.J. Wai, “Adaptive sliding mode for induction servomotor drive”, *IEE Proc.-Electr. Power Appl.*, vol. 147, no. 6, pp. 553-562, 2000.

[17] Y.J. Huang, T.C. Kuo and S.H. Chang, “Adaptive sliding-mode conrol for nonlinear systems with uncertain parameters”, *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, vol. 38, no. 2, pp. 534-539, 2008.

[18] F. Plestan, Y. Shtessel, et al, “New methodologies for adaptive sliding mode control”, *International Journal of Control*, vol. 83, no. 9, pp. 1907–1919, 2010.

[19] H. Schaub, J.L. Junkins, *Analytical Mechanics of Space Systems*, AIAA, Virginia, 2009.

[20] J.T.-Y. Wen, K. Kreutz-Delgado, “The attitude control problem”, *IEEE Transactions on Automatic Control*, vol. 36, no. 10, pp. 1148-1162, 1991.

[21] K.B. Park, Lee J.J., “Variable structure controller for robot manipulators using time-varying sliding surface”, in *Proc. IEEE International Conference on Robotics and Automation*, Atlanta, GA, USA, vol. 1, pp. 89–93, 1993.

[22] A. Bartoszewicz, A. Nowacka-Leverton, *Time-varying sliding modes for second and third order systems*, Berlin/Heidelberg: Springer-Verlag, 2009.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Non-fragile observer design for nonlinear switched time delay systems using delta operator

Ronghao Wang, Jianchun Xing, *IEEE member*
PLA University of Science and Technology
Nanjing, China
wrh1985@yahoo.com.cn

Ping Wang, Qiliang Yang
PLA University of Science and Technology
Nanjing, China
wp@893.com.cn

*Abstract*—**This paper considers the non-fragile observer design method for nonlinear switched time delay systems using the delta operator. Based on multiple Lyapunov function method and delta operator theory, an asymptotic stability criterion for delta operator switched system with time delay and Lipschitz nonlinearity is presented. By using the key technical lemma, a new sampling period and delay dependent design approach to the non-fragile observer is addressed. The proposed non-fragile observer can guarantee the estimated state error dynamics of delta operator time delay switched system can be asymptotically convergent for observer gain perturbations. The solution to the observer is formulated in the form of a set of linear matrix inequalities. A numerical example is employed to verify the proposed method.**

***Keywords-delta operator; non-fragile observer; Lipschitz nonlinear; switched systems; time delay***

## I. INTRODUCTION

Switched systems have attracted the interest of several scientists in the last several years. Switched systems are a class of hybrid systems consisting of subsystems and a switching law, which define a specific subsystem being activated during a certain interval of time. Switched systems exist widely in engineering and social systems, such as mechanical systems, automotive industry, aircraft and air traffic control and many other fields[1-3]. A lot of research in this direction have appeared recently[4,5]. Many important progress and remarkable achievements have been made on issues about stability and stabilization for the system. Based on common quadratic Lyapunov functions (CQLFs), a series of methods and conditions have been given for analyzing the stability of switched systems for arbitrary switching law[6, 7]. As effective tools, multiple Lyapunov function (MLF), switched Lyapunov function(SLF) and the average dwell-time approaches have been proposed to analyze the stability of switched systems, and many valuable results have been obtained for switched systems[8-10].

It is recognized that most discrete-time signals and systems are the results of sampling continuous-time signals and systems. When sampling is fast, all resulting signals and systems tend to become ill conditioned and thus difficult to deal with using the conventional algorithms. The delta operator-based algorithms are numerically better behaved under finite precision implementations for fast sampling[11]. A class of uncertain systems in delta domain has been studied and several results about robust stability for the system have been developed[12,13]. The problem of system instability in fast sampling can be solved by using delta operator model [13].

Recently, delta operator approach is used to investigate robust control for a class of uncertain switched systems, and the stabilization conditions of the delta operator switched systems are formulated in terms of a set of linear matrix inequalities (LMIs) [14]. The developed method is valid for the system. However, in actual operation, the states of the systems are not all measurable. It is necessary to design state observers for systems of this type. Several design procedures have been proposed to design state observers for switched systems[15, 16]. It is considered that the state observer gain variations could not be avoided in several applications, a kind of non-fragile observer is proposed and the design method is proved to be effective[17]. However, to date and to the best of our knowledge, the problem of the non-fragile observer design for time delay delta operator switched nonlinear systems has not been investigated, which motivated us for this study.

In this paper, we deal with the problem of the non-fragile observer design for time delay switched nonlinear systems using a delta operator, where the observer gain perturbations are assumed to be time-varying and unknown, but are norm-bounded. The aim is to design the non-fragile observer such that the estimated state error dynamics can be asymptotically convergent for observer gain perturbations. The desired non-fragile observer can be constructed by solving a set of LMIs. The remainder of the paper is organized as follows. In Section II, problem formulation and some necessary lemmas are given. In Section III, based on the MLF approach and delta operator theory, the stability analysis for a delta operator time delay switched nonlinear system is considered, and the result is dependent of time delay and will be employed to develop a non-fragile observer. A numerical example is given to illustrate the feasibility and effectiveness of the developed technique in Section IV, and concluding remarks are given in Section V.

## II. PROBLEM FORMULATION AND PRELIMINARY

Consider the following delta operator switched nonlinear system with time delay:

$$\delta x_k = A_{\sigma(k)}x_k + A_{d\sigma(k)}x_{k-n} + B_{\sigma(k)}u_k + f_{\sigma(k)}(x_k,k) \tag{1}$$

$$y_k = C_{\sigma(k)}x_k \tag{2}$$

$$x_k = \phi_k, \forall k \in \{-n,-n+1,\cdots,0\} \tag{3}$$

where $x_k \in R^p$ is the state vector at the $k$ th instant, $y_k \in R^m$

This work is supported by the Pre-Research Foundation of PLA University of Science and Technology under Grant No. 20110316.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

is the measurement output vector. $\phi_k$ is a vector-valued initial function, $f_i(\cdot,\cdot): R^p \times R \to R^p$ is an unknown nonlinear function, $n$ is the state delay of the system. $\sigma(k): Z^+ \to \underline{N} = \{1,2,\cdots,N\}$ is a switching signal. Moreover, $\sigma(k)=i$ means that the $i$ th subsystem is activated. $u_k \in R^l$ is the control input of the $i$ th subsystem at the $k$ th instants. $A_i \in R^{p\times p}, A_{di} \in R^{p\times p}, B_i \in R^{p\times l}, C_i \in R^{m\times p}$ for $i \in \underline{N}$ are known real-valued matrices with appropriate dimensions, $\delta$ denotes the delta operator, the definition can be seen in [11], ie, $\delta x_k = (x_{k+1} - x_k)/T$, where $T$ is a sample period.

We construct the following discrete-time switched system using delta operator to estimate the state of system (1)-(3):

$$\delta\hat{x}_k = A_{\sigma(k)}\hat{x}_k + A_{d\sigma(k)}\hat{x}_{k-n} + B_{\sigma(k)}u_k + f_{\sigma(k)}(\hat{x}_k,k) \quad (4)$$
$$+L_{\sigma(k)}(y_k - \hat{y}_k)$$

$$\hat{y}_k = C_{\sigma(k)}\hat{x}_k \quad (5)$$

$$\hat{x}_k = 0, \forall k \in \{-n,-n+1,\cdots,0\} \quad (6)$$

where $\hat{x}_k \in R^p$ is the estimated state vector of $x_k$, $\hat{y}_k \in R^m$ is the observer output vector, $L_i \in R^{p\times m}$ for $i \in \underline{N}$ is the observer gain.

Let $\tilde{x}_k = x_k - \hat{x}_k$ be the estimated state error. By the definition of the delta operator, we have $\delta\tilde{x}_k = \delta x_k - \delta\hat{x}_k$, then we can obtain the following error system from (1)-(6):

$$\delta\tilde{x}_k = (A_{\sigma(k)} - L_{\sigma(k)}C_{\sigma(k)})\tilde{x}_k + A_{d\sigma(k)}\tilde{x}_{k-n} + f_{\sigma(k)}(x_k,k) \quad (7)$$
$$-f_{\sigma(k)}(\hat{x}_k,k)$$

$$\tilde{x}_k = \phi_k, \forall k \in \{-n,-n+1,\cdots,0\} \quad (8)$$

If the state observer gain variations could not be avoided, a kind of non-fragile state observer will be designed as follows:

$$\delta\hat{x}_k = A_{\sigma(k)}\hat{x}_k + A_{d\sigma(k)}\hat{x}_{k-n} + B_{\sigma(k)}u_k + f_{\sigma(k)}(\hat{x}_k,k) \quad (9)$$
$$+(L_{\sigma(k)} + \Delta L_{\sigma(k)})(y_k - \hat{y}_k)$$

$$\hat{y}_k = C_{\sigma(k)}\hat{x}_k \quad (10)$$

$$\hat{x}_k = 0, \forall k \in \{-n,-n+1,\cdots,0\} \quad (11)$$

where $\Delta L_i \in R^{p\times m}$ are uncertain real-valued matrix functions representing norm-bounded parameter uncertainties.

According to system (1)-(3) and (9)-(11), the dynamic equations of error switched system for non-fragile observer can be prescribed:

$$\delta\tilde{x}_k = [A_{\sigma(k)} - (L_{\sigma(k)} + \Delta L_{\sigma(k)})C_{\sigma(k)}]\tilde{x}_k + A_{d\sigma(k)}\tilde{x}_{k-n} \quad (12)$$
$$+f_{\sigma(k)}(x_k,k) - f_{\sigma(k)}(\hat{x}_k,k)$$

$$\tilde{x}_k = \phi_k, \forall k \in \{-n,-n+1,\cdots,0\} \quad (13)$$

Without loss of generality, we make the following assumptions.

**Assumption 1** $f_i(x_k,k)$ for $i \in \underline{N}$ are nonlinear functions satisfying:

$$\|f_i(x_k,k) - f_i(\hat{x}_k,k)\| \le \|U_i(x_k - \hat{x}_k)\| \quad (14)$$

where $U_i$ are known real constant matrices.

**Assumption 2** The gain perturbations $\Delta L_i$ are of the norm-bounded form:

$$\Delta L_i = H_i F_{ik} E_i \quad (15)$$

where $H_i$, $E_i$ which denote the structure of the uncertainties are known real constant matrices with proper dimensions, and $F_{ik}$ are unknown time-varying matrices which satisfy:

$$F_{ik}^T F_{ik} \le I$$

The unknown matrices $F_{ik}$ contain the uncertain parameters in the linear part of the subsystem and the matrices $H_i$, $E_i$ specify how the unknown parameters in $F_{ik}$ affect the elements of the nominal matrices $L_i$.

The parameter uncertainty structure in equation (15) has been widely used and can represent parameter uncertainty in many physical cases. Equation (15) is called additive gain variations.

**Lemma 1** [18] Let $U$, $V$, $W$ and $X$ be real matrices of appropriate dimensions with $X$ satisfying $X = X^T$, then for all $V^T V \le I$

$$X + UVW + W^T V^T U^T < 0$$

if and only if there exists a scalar $\varepsilon > 0$ such that

$$X + \varepsilon UU^T + \varepsilon^{-1} W^T W < 0$$

**Lemma 2** [19] Consider the following system

$$x_{k+1} = f_k(x_k) \quad (16)$$

If there is a function $V: Z^+ \times R^n \to R$ such that:

(1) $V$ is a positive-definite function, decreasing and radially unbounded.

(2) $\Delta V(k,x_k) = V(k+1,x_{k+1}) - V(k,x_k) < 0$ is negative definite along the solution of (16).

then system (16) is asymptotically stable.

**Lemma 3**[20] For any constant positive semi-definite symmetric matrix $W$, two positive integers $r$ and $r_0$ satisfying $r \ge r_0 \ge 1$, the following inequality holds

$$\left(\sum_{i=r_0}^{r} x(i)\right)^T W \left(\sum_{i=r_0}^{r} x(i)\right) \le \rho \sum_{i=r_0}^{r} x^T(i) W x(i)$$

where $\rho = r - r_0 + 1$.

The objective of this paper is to design non-fragile observer gain $L_i$ for delta operator time delay switched nonlinear systems (1)-(3) such that the estimated state error dynamics is asymptotically convergent.

## III. Main Results

### A. Stability analysis

In this subsection, we investigate the stability of the following delta operator time delay switched nonlinear system

$$\delta x_k = A_{\sigma(k)}x_k + A_{d\sigma(k)}x_{k-n} + f_{\sigma(k)}(x_k,k) \quad (17)$$

where $\|f_i(x_k,k)\| \le \|U_i x_k\|$ for $i \in \underline{N}$. Define the indicator function:

$$\xi(k) = (\xi_1(k), \xi_2(k), \cdots, \xi_N(k))^T \quad (18)$$

with $i \in \{1,2,\cdots,N\}$, where

$$\xi_i(k)=\begin{cases}1 & \text{when the } i\text{th subsystem is activated} \\ 0 & \text{others}\end{cases},$$

then system (17) can be written as

$$\delta x_k=\sum_{i=1}^{N}\xi_i(k)(A_i x_k + A_{di}x_{k-n}+f_{\sigma(k)}(x_k,k))$$

$$\forall i\in\{1,2,\cdots,N\}$$

**Theorem 1** Consider system (17). If there exist symmetric positive definite matrices $P_i>0$, $R_i^{(r)}>0$, $R_i^{(s)}>0$, $Q_i^{(s)}>0$, $Q_i^{(n)}>0$ and scalar $\varepsilon_i>0$, $i\in\{1,2,\cdots,N\}$, $r=1,2,\cdots,s-1$, $s=1,2,\cdots,n-1$, such that

$$P_i<\varepsilon_i I \tag{19}$$

$$R_j^{(r+1)}<R_i^{(r)},\ r=1,2,\cdots,s-1 \tag{20}$$

$$Q_j^{(s+1)}<Q_i^{(s)},\ s=1,2,\cdots,n-1 \tag{21}$$

$$\left[\begin{array}{cc} \begin{array}{l} A_i^TP_j+P_jA_i+TA_i^TP_jA_i+\varepsilon_i(T+1)U_i^TU_i \\ +\frac{1}{T}(P_j-\frac{1}{2}P_i)-\frac{1}{n}R_i^{(s)}+Q_j^{(1)} \end{array} & A_i^TP_i \\ * & -P_i+nT^2R_j^{(1)} \rightarrow \\ * & * \end{array}\right. \tag{22}$$

$$\left.\begin{array}{c} (TA_i^T+I)P_jA_{di}+\frac{1}{n}R_i^{(s)} \\ P_iA_{di} \\ TA_{di}^TP_jA_{di}-Q_i^{(n)}-\frac{1}{n}R_i^{(s)} \end{array}\right]<0$$

$\forall i,j\in\{1,2,\cdots,N\}$, then system (17) is asymptotically stable under arbitrary switching.

**Proof** Consider the following switched Lyapunov-Krasovskii functional:

$$V(k,x_k)=V_1(k,x_k)+V_2(k,x_k)+V_3(k,x_k) \tag{23}$$

with

$$V_1(k,x_k)=\frac{1}{2}x_k^TP(\xi(k))x_k=\frac{1}{2}x_k^T\left(\sum_{i=1}^{N}\xi_i(k)P_i\right)x_k$$

$$V_2(k,x_k)=T\sum_{s=1}^{n}x_{k-s}^TQ^{(s)}(\xi(k))x_{k-s}=T\sum_{s=1}^{n}\sum_{i=1}^{N}x_{k-s}^T\xi_i(k)Q_i^{(s)}x_{k-s}$$

$$V_3(k,x_k)=T\sum_{s=1}^{n}\sum_{r=1}^{s}e_{k-r}^TR^{(r)}(\xi(k))e_{k-r}=T\sum_{s=1}^{n}\sum_{r=1}^{s}\sum_{i=1}^{N}e_{k-r}^T\xi_i(k)R_i^{(r)}e_{k-r}$$

with $P_i$, $Q_i^{(s)}$, $R_i^{(r)}$, $i\in\{1,2,\cdots,N\}$ being symmetric positive-definite matrices and $e_k=x_k-x_{k+1}$.

Then switched Lyapunov-Krasovskii functional in delta domain has the following form:

$$\delta V_1(k,x_k)=\frac{V_1(k+1,x_{k+1})-V_1(k,x_k)}{T}$$

$$=\frac{x_{k+1}^TP(\xi(k+1))x_{k+1}-x_k^TP(\xi(k))x_k}{2T}$$

$$=\frac{(x_k+T\delta x_k)^TP(\xi(k+1))(x_k+T\delta x_k)-x_k^TP(\xi(k))x_k}{2T} \tag{24}$$

$$\le x_k^T[A_i^TP_j+P_jA_i+TA_i^TP_jA_i+\varepsilon_iTU_i^TU_i+\frac{1}{T}(P_j-\frac{1}{2}P_i)]x_k$$

$$+2x_k^T[(TA_i^T+I)P_jA_{di}]x_{k-n}+x_{k-n}^T(TA_{di}^TP_jA_{di})x_{k-n}$$

$$\delta V_2(k,x_k)=\frac{1}{T}[T\sum_{s=1}^{n}x_{k-s+1}^TQ^{(s)}(\xi(k+1))x_{k-s+1}-T\sum_{s=1}^{n}x_{k-s}^TQ^{(s)}(\xi(k))x_{k-s}]$$

$$=x_k^TQ_j^{(1)}x_k-x_{k-n}^TQ_i^{(n)}x_{k-n}+\sum_{s=1}^{n-1}x_{k-s}^T(Q_j^{(s+1)}-Q_i^{(s)})x_{k-s}$$

By (21), we can obtain that

$$\delta V_2(k,x_k)<x_k^TQ_j^{(1)}x_k-x_{k-n}^TQ_i^{(n)}x_{k-n} \tag{25}$$

$$\delta V_3(k,x_k)=\frac{1}{T}[T\sum_{s=1}^{n}\sum_{r=1}^{s}e_{k-r+1}^TR^{(r)}(\xi(k+1))e_{k-r+1}-T\sum_{s=1}^{n}\sum_{r=1}^{s}e_{k-r}^TR^{(r)}(\xi(k))e_{k-r}]$$

$$=ne_k^TR_j^{(1)}e_k+\sum_{s=2}^{n}\sum_{r=1}^{s-1}e_{k-r}^TR_j^{(r+1)}e_{k-r}-\sum_{s=2}^{n}\sum_{r=1}^{s-1}e_{k-r}^TR_i^{(r)}e_{k-r}-\sum_{s=1}^{n}e_{k-s}^TR_i^{(s)}e_{k-s}$$

By (20), we can obtain that

$$\begin{aligned}\delta V_3(k,x_k)&\le ne_k^TR_j^{(1)}e_k-\frac{1}{n}(\sum_{s=1}^{n}e_{k-s})^TR_i^{(s)}(\sum_{s=1}^{n}e_{k-s})\\&=-\frac{1}{n}(x_{k-n}-x_k)^TR_i^{(s)}(x_{k-n}-x_k)+nT^2\delta x_k^TR_j^{(1)}\delta x_k\end{aligned} \tag{26}$$

If $\delta V(k,x_k)<0$ holds under arbitrary switching signal, it follows that this has to hold for special configuration $\xi_i(k)=1$, $\xi_{h\neq i}(k)=0$, $\xi_j(k+1)=1$, $\xi_{g\neq j}(k+1)=0$ and for all $x_k\in R^p$. By (24), (25) and (26), we have

$$\begin{aligned}\delta V(k,x_k)&=\delta V_1(k,x_k)+\delta V_2(k,x_k)+\delta V_3(k,x_k)\\&\le x_k^T[A_i^TP_j+P_jA_i+TA_i^TP_jA_i+\varepsilon_iTU_i^TU_i\\&+\frac{1}{T}(P_j-\frac{1}{2}P_i)]x_k+x_{k-n}^T(TA_{di}^TP_jA_{di})x_{k-n}\\&+2x_k^T[(TA_i^T+I)P_jA_{di}]x_{k-n}+x_k^TQ_j^{(1)}x_k+nT^2\delta x_k^TR_j^{(1)}\delta x_k\\&-x_{k-n}^TQ_i^{(n)}x_{k-n}-\frac{1}{n}(x_{k-n}-x_k)^TR_i^{(s)}(x_{k-n}-x_k)\end{aligned} \tag{27}$$

Notice that

$$0=-2\delta x_k^TP_i(\delta x_k-A_ix_k-A_{di}x_{k-n}-f_i(x_k,k)) \tag{28}$$

Combining (27) and (28) leads to

$$\delta V(k,x_k)\le\eta^T\left[\begin{array}{cc} \begin{array}{l} A_i^TP_j+P_jA_i+TA_i^TP_jA_i+\varepsilon_i(T+1)U_i^TU_i \\ +\frac{1}{T}(P_j-\frac{1}{2}P_i)-\frac{1}{n}R_i^{(s)}+Q_j^{(1)} \end{array} & A_i^TP_i \\ * & -P_i+nT^2R_j^{(1)} \rightarrow \\ * & * \end{array}\right.$$

$$\left.\begin{array}{c} (TA_i^T+I)P_jA_{di}+\frac{1}{n}R_i^{(s)} \\ P_iA_{di} \\ TA_{di}^TP_jA_{di}-Q_i^{(n)}-\frac{1}{n}R_i^{(s)} \end{array}\right]\eta$$

where $\eta = \begin{bmatrix} x_k^T & \delta x_k^T & x_{k-n}^T \end{bmatrix}^T$. From (22), we have $\delta V(k, x_k) < 0$. By Lemma 2, the system (17) is asymptotically stable. The proof is completed.

**Remark 1** When time delay $n \equiv 0$ and $f_{\sigma(k)}(x_k, k) \equiv 0$, system (17) becomes as follows:

$$\delta x_k = \tilde{A}_{\sigma(k)} x_k$$

where $\tilde{A}_{\sigma(k)} = A_{\sigma(k)} + A_{d\sigma(k)}$. From (22), we know that the matrix inequality imply that the following inequality holds:

$$\tilde{A}_i^T P_j + P_j \tilde{A}_i + T \tilde{A}_i^T P_j \tilde{A}_i + \frac{1}{T}(P_j - P_i) < 0$$

The above inequality is just a sufficient condition of stability for the delta operator switched system without time delay, which can be expressed in [14].

**Remark 2** When sample period $T = 0$ and $f_{\sigma(k)}(x_k, k) \equiv 0$, system (17) becomes a continuous-time switched linear system as follows:

$$\dot{x}(t) = A_{\sigma(t)} x(t) + A_{d\sigma(t)} x(t-d) \tag{29}$$

We can obtain sufficient condition of stability for system (29) by Theorem 1.

**Corollary 1** Consider system (29). If there exist symmetric positive definite matrices $P > 0$, $R > 0$, $Q > 0$, such that

$$\begin{bmatrix} A_i^T P + P A_i - \frac{1}{d} R + Q & A_i^T P & P A_{di} + \frac{1}{d} R \\ * & -2P & P A_{di} \\ * & * & -Q - \frac{1}{d} R \end{bmatrix} < 0 \tag{30}$$

$\forall i \in \{1, 2, \cdots, N\}$, then system (29) is asymptotically stable under arbitrary switching.

**Remark 3** Let $\bar{A}_{\sigma(k)} = A_{\sigma(k)} + I$. When sample period $T = 1$ and $f_{\sigma(k)}(x_k, k) \equiv 0$, system (17) becomes a discrete-time switched linear system as follows:

$$x_{k+1} = \bar{A}_{\sigma(k)} x_k + A_{d\sigma(k)} x_{k-n} \tag{31}$$

We can obtain sufficient condition of stability for system (31) by Theorem 1.

**Corollary 2** Consider system (31). If there exist symmetric positive definite matrices $P_i > 0$, $R_i^{(r)} > 0$, $R_i^{(s)} > 0$, $Q_i^{(s)} > 0$, $Q_i^{(n)} > 0$, $i \in \{1, 2, \cdots, N\}$, $r = 1, 2, \cdots, s-1$, $s = 1, 2, \cdots, n-1$, such that

$$R_j^{(r+1)} < R_i^{(r)},\ r = 1, 2, \cdots, s-1 \tag{32}$$

$$Q_j^{(s+1)} < Q_i^{(s)},\ s = 1, 2, \cdots, n-1 \tag{33}$$

$$\begin{bmatrix} \begin{matrix} \bar{A}_i^T P_j + P_j \bar{A}_i + \bar{A}_i^T P_j \bar{A}_i \\ -P_j - P_i - \frac{1}{n} R_i^{(s)} + Q_j^{(1)} \end{matrix} & \bar{A}_i^T P_i & \bar{A}_i^T P_j A_{di} + \frac{1}{n} R_i^{(s)} \\ * & -2P_i + n R_j^{(1)} & P_i A_{di} \\ * & * & A_{di}^T P_j A_{di} - Q_i^{(n)} - \frac{1}{n} R_i^{(s)} \end{bmatrix} < 0 \tag{34}$$

$\forall i, j \in \{1, 2, \cdots, N\}$, then system (31) is asymptotically stable.

*B. Observer Design*

Now we consider system (7)-(8). The following theorem presents sufficient conditions for the existence of asymptotic stability of system (7)-(8).

**Theorem 2** Consider system (7)-(8), if there exist a set of matrices $X_i > 0$, $V_i^{(s)} > 0$, $V_i^{(n)} > 0$, $Z_i^{(r)} > 0$, $Z_i^{(s)} > 0$, $V_{ij}^{(s)} > 0$, $V_{ij}^{(n)} > 0$, $Z_{ij}^{(r)} > 0$, $Z_{ij}^{(s)} > 0$, $\varepsilon_i > 0$, $i \in \{1, 2, \cdots, N\}$, $r = 1, 2, \cdots, s-1$, $s = 1, 2, \cdots, n-1$, and $W_i$ such that the following LMIs have feasible solutions

$$X_i > \varepsilon_i^{-1} I \tag{35}$$

$$Z_{ij}^{(r+1)} < Z_i^{(r)},\ r = 1, 2, \cdots, s-1 \tag{36}$$

$$V_{ij}^{(s+1)} < V_i^{(s)},\ s = 1, 2, \cdots, n-1 \tag{37}$$

$$\left[\begin{matrix} -\frac{1}{2T} X_i + V_{ij}^{(1)} - \frac{1}{n} Z_i^{(s)} & (A_i X_i - W_i)^T & \frac{1}{n} Z_i^{(s)} \\ * & -X_i + n T^2 Z_{ij}^{(1)} & A_{di} X_i \\ * & * & -V_i^{(n)} - \frac{1}{n} Z_i^{(s)} \\ * & * & * \\ * & * & * \end{matrix}\right. \rightarrow \tag{38}$$

$$\left.\begin{matrix} T(A_i X_i - W_i)^T + X_i & X_i U_i^T \\ 0 & 0 \\ T X_i A_{di}^T & 0 \\ -T X_j & 0 \\ * & -\frac{1}{\varepsilon_i (T+1)} I \end{matrix}\right] < 0$$

where $i, j \in \{1, 2, \cdots, N\}$, then the state observer gain $L_i$ satisfying the following equations:

$$L_i C_i = W_i X_i^{-1} \tag{39}$$

can guarantees system (7)-(8) is asymptotically stable under arbitrary switching.

**Proof** Denote

$$M_{ij} = \left[\begin{matrix} \begin{matrix} (A_i - L_i C_i)^T P_j + P_j (A_i - L_i C_i) + T (A_i - L_i C_i)^T P_j (A_i - L_i C_i) \\ + \varepsilon_i (T+1) U_i^T U_i + \frac{1}{T}(P_j - \frac{1}{2} P_i) - \frac{1}{n} R_i^{(s)} + Q_j^{(1)} \end{matrix} & (A_i - L_i C_i)^T P_i \\ * & -P_i + n T^2 R_j^{(1)} \\ * & * \end{matrix}\right. \rightarrow$$

$$\left.\begin{matrix} [T (A_i - L_i C_i)^T + I] P_j A_{di} + \frac{1}{n} R_i^{(s)} \\ P_i A_{di} \\ T A_{di}^T P_j A_{di} - Q_i^{(n)} - \frac{1}{n} R_i^{(s)} \end{matrix}\right]$$

By Theorem 1, if $M_{ij} < 0$, then the system (7)-(8) is asymptotically stable. By Schur Complement, we have $M_{ij} < 0$ is equivalent to

$$\begin{bmatrix} -\frac{1}{2T}P_i+Q_j^{(1)}-\frac{1}{n}R_i^{(s)} \\ +\varepsilon_i(T+1)U_i^TU_i & (A_i-L_iC_i)^TP_i & \frac{1}{n}R_i^{(s)} & T(A_i-L_iC_i)^T+I \\ * & -P_i+nT^2R_j^{(1)} & P_iA_{di} & 0 \\ * & * & -Q_i^{(n)}-\frac{1}{n}R_i^{(s)} & TA_{di}^T \\ * & * & * & -TP_j^{-1} \end{bmatrix}<0 \tag{40}$$

Using $diag\left\{P_i^{-1} \quad P_i^{-1} \quad P_i^{-1} \quad I\right\}$ to pre- and post- multiply the left term of (40) respectively, and denoting $X_i=P_i^{-1}$, $W_i=L_iC_iX_i$, $Z_i^{(s)}=P_i^{-1}R_i^{(s)}P_i^{-1}$, $V_i^{(n)}=P_i^{-1}Q_i^{(n)}P_i^{-1}$, $V_{ij}^{(1)}=P_i^{-1}Q_j^{(1)}P_i^{-1}$, $Z_{ij}^{(1)}=P_i^{-1}R_j^{(1)}P_i^{-1}$, we can obtain

$$\left[\begin{matrix} -\frac{1}{2T}X_i+V_{ij}^{(1)}-\frac{1}{n}Z_i^{(s)} \\ +\varepsilon_i(T+1)X_iU_i^TU_iX_i & X_i(A_i-L_iC_i)^T \\ * & -X_i+nT^2Z_{ij}^{(1)} \\ * & * \\ * & * \end{matrix}\right. \rightarrow \tag{41}$$

$$\left.\begin{matrix} \frac{1}{n}Z_i^{(s)} & TX_i(A_i-L_iC_i)^T+X_i \\ A_{di}X_i & 0 \\ -V_i^{(n)}-\frac{1}{n}Z_i^{(s)} & TX_iA_{di}^T \\ * & -TX_j \end{matrix}\right]<0$$

Using Schur Complement again, we can obtain (38). By the MFL method, $M_{ij}<0$, $\forall i,j\in\{1,2,\cdots,N\}$ can guarantee that the system (7)-(8) is asymptotically stable under arbitrary switching. The proof is completed.

The procedure of observer design of system (7)-(8) can be summarized as follows:

**The Procedure of Observer Design**

**Step 1.** If there exist a set of positive definite symmetric matrices $X_i>0$, $V_i^{(s)}>0$, $V_i^{(n)}>0$, $Z_i^{(r)}>0$, $Z_i^{(s)}>0$, $V_{ij}^{(s)}>0$, $V_{ij}^{(n)}>0$, $Z_{ij}^{(r)}>0$, $Z_{ij}^{(s)}>0$, $\varepsilon_i>0$, $i\in\{1,2,\cdots,N\}$, $r=1,2,\cdots,s-1$, $s=1,2,\cdots,n-1$ and $W_i$, LMIs (35)-(38) have feasible solutions, then go to Step 2.

**Step 2.** If the equations $L_iC_i=W_iX_i^{-1}$ have feasible solutions for unknown matrices $L_i$, then the observer can be constructed as (4)-(6), where $L_i$ are the observer gain.

**Remark 4** For the matrices equations (39), by the matrix theory we can know that the sufficient and necessary condition for the equations having feasible solutions is

$$rank[C_i]=rank[C_i^T\vdots(W_iX_i^{-1})^T]$$

However, when $rank[C_i]\neq rank[C_i^T\vdots(W_iX_i^{-1})^T]$, the equations don not exist feasible solutions. In particular, if $rank[C_i]=p$, that is to say the output matrices $C_i$ is column filled rank matrices, then the equations (39) have feasible solutions. But $rank[C_i]\neq p$ implies that the equations (39) may have not feasible solutions. When there do not exist feasible solutions for the equations (39), we can construct the matrices $B_i$ such that $rank[C_i]=rank[C_i^T\vdots B_i^T]$. Let $W_iX_i^{-1}=B_i$, substituting $W_i=B_iX_i$ to (38), if there exists feasible solutions for LMIs (35)-(38), then the observer gain $L_i$ can be obtain from the equations $L_iC_i=B_i$. Specially, when the matrices $B_i$ are chosen as $C_i$ and LMIs (35)-(38) have feasible solutions, $L_i$ can be constructed as $L_i=I$ obviously, where $I$ denotes the identity matrix.

### C. *Non-fragile Observer Design*

Now consider systems (12)-(13), if there exist additive gain variations in state observer, we can obtain the following result.

**Theorem 3** Consider system (12)-(13), if there exist a set of matrices $X_i>0$, $V_i^{(s)}>0$, $V_i^{(n)}>0$, $Z_i^{(r)}>0$, $Z_i^{(s)}>0$, $V_{ij}^{(s)}>0$, $V_{ij}^{(n)}>0$, $Z_{ij}^{(r)}>0$, $Z_{ij}^{(s)}>0$, $\varepsilon_i>0$, $\mu_i>0$, $i\in\{1,2,\cdots,N\}$, $r=1,2,\cdots,s-1$, $s=1,2,\cdots,n-1$ and $W_i$ such that the following LMIs hold:

$$X_i>\varepsilon_i^{-1}I \tag{42}$$

$$Z_{ij}^{(r+1)}<Z_i^{(r)},\ r=1,2,\cdots,s-1 \tag{43}$$

$$V_{ij}^{(s+1)}<V_i^{(s)},\ s=1,2,\cdots,n-1 \tag{44}$$

$$\left[\begin{matrix} -\frac{1}{2T}X_i+V_{ij}^{(1)}-\frac{1}{n}Z_i^{(s)} & (A_iX_i-W_i)^T & \frac{1}{n}Z_i^{(s)} \\ * & -X_i+nT^2Z_{ij}^{(1)}+\mu_iH_iH_i^T & A_{di}X_i \\ * & * & -V_i^{(n)}-\frac{1}{n}Z_i^{(s)} \\ * & * & * \\ * & * & * \\ * & * & * \end{matrix}\right. \rightarrow \tag{45}$$

$$\left.\begin{matrix} T(A_iX_i-W_i)^T+X_i & X_iU_i^T & -X_iC_i^TE_i^T \\ \mu_iTH_iH_i^T & 0 & 0 \\ TX_iA_{di}^T & 0 & 0 \\ -TX_j+\mu_iT^2H_iH_i^T & 0 & 0 \\ * & -\frac{1}{\varepsilon_i(T+1)}I & 0 \\ * & * & -\mu_iI \end{matrix}\right]<0$$

where $i\in\{1,2,\cdots,N\}$, then the non-fragile state observer gain $L_i$ satisfying the following equations:

$$L_iC_i=W_iX_i^{-1} \tag{46}$$

can guarantees system (12)-(13) is asymptotically stable under arbitrary switching.

**Proof** Denote

$$T_{ij}=\left[\begin{matrix} -\frac{1}{2T}X_i+V_{ij}^{(1)}-\frac{1}{n}Z_i^{(s)} & X_i[A_i-(L_i+\Delta L_i)C_i]^T & \frac{1}{n}Z_i^{(s)} \\ * & -X_i+nT^2Z_{ij}^{(1)} & A_{di}X_i \\ * & * & -V_i^{(n)}-\frac{1}{n}Z_i^{(s)} \\ * & * & * \\ * & * & * \end{matrix}\right. \rightarrow$$

$$\begin{matrix} TX_i[A_i-(L_i+\Delta L_i)C_i]^T+X_i & X_iU_i^T \\ 0 & 0 \\ TX_iA_{di}^T & 0 \\ -TX_j & 0 \\ * & -\dfrac{1}{\varepsilon_i(T+1)}I \end{matrix}\right]$$

Substituting (15) into the above equation leads to

$$T_{ij}=T_{ij}^{(1)}+T_i$$

where

$$T_{ij}^{(1)}=\left[\begin{matrix} -\dfrac{1}{2T}X_i+V_{ij}^{(1)}-\dfrac{1}{n}Z_i^{(s)} & [(A_i-L_iC_i)X_i]^T & \dfrac{1}{n}Z_i^{(s)} \\ * & -X_i+nT^2Z_{ij}^{(1)} & A_{di}X_i \\ * & * & -V_i^{(n)}-\dfrac{1}{n}Z_i^{(s)} \rightarrow \\ * & * & * \\ * & * & * \end{matrix}\right.$$

$$\left.\begin{matrix} T[(A_i-L_iC_i)X_i]^T+X_i & X_iU_i^T \\ 0 & 0 \\ TX_iA_{di}^T & 0 \\ -TX_j & 0 \\ * & -\dfrac{1}{\varepsilon_i(T+1)}I \end{matrix}\right]$$

$$T_i=\begin{bmatrix} 0 & -(H_iF_{ik}E_iC_iX_i)^T & 0 & -T(H_iF_{ik}E_iC_iX_i)^T & 0 \\ * & 0 & 0 & 0 & 0 \\ * & * & 0 & 0 & 0 \\ * & * & * & 0 & 0 \\ * & * & * & * & 0 \end{bmatrix}$$

$$=\begin{bmatrix} 0 \\ H_i \\ 0 \\ TH_i \\ 0 \end{bmatrix}F_{ik}\begin{bmatrix}-E_iC_iX_i & 0 & 0 & 0 & 0\end{bmatrix}+\left(\begin{bmatrix} 0 \\ H_i \\ 0 \\ TH_i \\ 0 \end{bmatrix}F_{ik}\begin{bmatrix}-E_iC_iX_i & 0 & 0 & 0 & 0\end{bmatrix}\right)^T$$

According to Lemma 1, we have

$$T_{ij}\le T_{ij}^{(1)}+\mu_i\begin{bmatrix} 0 \\ H_i \\ 0 \\ TH_i \\ 0 \end{bmatrix}\begin{bmatrix} 0 \\ H_i \\ 0 \\ TH_i \\ 0 \end{bmatrix}^T \tag{47}$$

$$+\mu_i^{-1}\begin{bmatrix}-E_iC_iX_i & 0 & 0 & 0 & 0\end{bmatrix}^T\begin{bmatrix}-E_iC_iX_i & 0 & 0 & 0 & 0\end{bmatrix}$$

Denote $Z_{ij}$ is the right term of inequality (47), by Schur Complement, $Z_{ij}<0$ are equivalent to

$$\left[\begin{matrix} -\dfrac{1}{2T}X_i+V_{ij}^{(1)}-\dfrac{1}{n}Z_i^{(s)} & [(A_i-L_iC_i)X_i]^T & \dfrac{1}{n}Z_i^{(s)} \\ * & -X_i+nT^2Z_{ij}^{(1)}+\mu_iH_iH_i^T & A_{di}X_i \\ * & * & -V_i^{(n)}-\dfrac{1}{n}Z_i^{(s)} \rightarrow \\ * & * & * \\ * & * & * \\ * & * & * \end{matrix}\right.$$

$$\left.\begin{matrix} T[(A_i-L_iC_i)X_i]^T+X_i & X_iU_i^T & -X_iC_i^TE_i^T \\ \mu_iTH_iH_i^T & 0 & 0 \\ TX_iA_{di}^T & 0 & 0 \\ -TX_j+\mu_iT^2H_iH_i^T & 0 & 0 \\ * & -\dfrac{1}{\varepsilon_i(T+1)}I & 0 \\ * & 0 & -\mu_iI \end{matrix}\right]<0 \tag{48}$$

Denote $W_i=L_iC_iX_i$, (48) is equivalent to (45). By Theorem 1, we conclude that system (12)-(13) is asymptotically stable. The proof is completed.

If the state observer gain variations could not be avoided, then the non-fragile observer for system (1)-(3) can be designed according to the following procedure.

**The Procedure of Non-fragile Observer Design**

**Step 1.** If there exist a set of positive definite symmetric matrices $X_i>0$, $V_i^{(s)}>0$, $V_i^{(n)}>0$, $Z_i^{(r)}>0$, $Z_i^{(s)}>0$, $V_{ij}^{(s)}>0$, $V_{ij}^{(n)}>0$, $Z_{ij}^{(r)}>0$, $Z_{ij}^{(s)}>0$, $\varepsilon_i>0$, $\mu_i>0$, $i\in\{1,2,\cdots,N\}$, $r=1,2,\cdots,s-1$, $s=1,2,\cdots,n-1$ and $W_i$, LMIs (42)-(45) have feasible solutions, then go to Step 2.

**Step 2.** If the equations $L_iC_i=W_iX_i^{-1}$ have feasible solutions for unknown matrices $L_i$, then the observer can be constructed as (9)-(11), where $L_i$ are the observer gain.

## IV. NUMERICAL EXAMPLE

In this section, we present an example to illustrate the effectiveness of the proposed approach. Consider system (1)-(3) with parameters as follows:

$$A_1=\begin{bmatrix} 0.1 & 0.2 \\ 0.3 & -0.1 \end{bmatrix},\ A_2=\begin{bmatrix} -0.1 & 0 \\ 0.1 & 0.2 \end{bmatrix},$$

$$A_{d1}=\begin{bmatrix} 0.1 & 0.1 \\ 0 & 0.2 \end{bmatrix},\ A_{d2}=\begin{bmatrix} 0 & 0.1 \\ 0 & 0 \end{bmatrix},$$

$$E_1=\begin{bmatrix} -0.1 & 0 \\ 0.1 & 0 \end{bmatrix},\ E_2=\begin{bmatrix} 0.1 & 0.1 \\ 0 & 0.1 \end{bmatrix},$$

$$U_1=\begin{bmatrix} 0.1 & 0 \\ 0 & 0 \end{bmatrix},U_2=\begin{bmatrix} 0 & 0 \\ -2 & 0 \end{bmatrix}$$

$$H_1=H_2=\begin{bmatrix} 0.2 & 0.1 \\ 0 & 0 \end{bmatrix},$$

$$C_1=C_2=\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

and the uncertain time-varying parameter matrices $F_{1,k} = F_{2,k} = \begin{bmatrix} \sin k & 0 \\ 0 & \sin k \end{bmatrix}$. The sampling period $T = 0.5$, and the time delay $n = 2$, then using the Matlab LMI Control Toolbox to solve the LMIs in (42) to (45), we obtain the solution as follows:

$$X_1 = \begin{bmatrix} 76.8668 & 0.1346 \\ 0.1346 & 84.1706 \end{bmatrix}, X_2 = \begin{bmatrix} 36.2793 & -0.0023 \\ -0.0023 & 76.7278 \end{bmatrix}$$

$$W_1 = \begin{bmatrix} 90.1949 & 18.2174 \\ 23.3830 & 64.1468 \end{bmatrix}, W_2 = \begin{bmatrix} 28.0449 & -1.0833 \\ 3.6168 & 76.6194 \end{bmatrix}$$

Therefore, by Theorem 3 a desired non-fragile observer gain can be solved from the following equations:

$$L_1 C_1 = W_1 X_1^{-1} \text{ and } L_2 C_2 = W_2 X_2^{-1}$$

The we can obtain that

$$L_1 = \begin{bmatrix} 1.1730 & -0.9584 \\ 0.3029 & 0.4587 \end{bmatrix}, L_2 = \begin{bmatrix} 0.7730 & -0.7871 \\ 0.0998 & 0.8988 \end{bmatrix}.$$

## V. Conclusion

In this paper, we focus on both stability and non-fragile observer design for delta operator time delay switched nonlinear systems. By using the MLF method and choosing Lyapunov-Krasovskii functional in delta domain appropriately, sufficient conditions for the existence of non-fragile observer have been given in terms of LMIs. The result is dependent of sampling period and time delay and can be developed for non-fragile observer problems. Future work will focus on using the proposed design method to solve the problem of stabilization based on non-fragile observer for time delay switched nonlinear systems via delta operator approach.

## Acknowledgment

This work is supported by the Pre-Research Foundation of PLA University of Science and Technology under Grant No. 20110316. The authors would like to thank the editor and anonymous reviewers for their valuable comments, which helped to significantly improve the quality and presentation of this paper.

## References

[1] P. Varaiya, "Smart cars on smart roads: problems of control," IEEE Transactions on Automatic Control, vol.38, no. 2, pp. 195–207, 1993.

[2] W. Wang and R. W. Brockett, "Systems with finite communication bandwidth constraints − part I: state estimation problems," IEEE Transactions on Automatic Control, vol. 42, no. 9, pp. 1294–1299, 1997.

[3] C. Tomlin, G. J. Pappas and S. Sastry, "Conflict resolution for air traffic management: a study in multi-agent hybrid systems," IEEE Transactions on Automatic Control, vol. 43, no. 4, pp. 509–521, 1998.

[4] D. Cheng, L. Guo, Y. Lin and Y. Wang, "Stabilization of switched linear systems," IEEE Transactions on Automatic Control, vol. 50, no. 5, pp. 661–666, 2005.

[5] J. Daafouz, P. Riedinger and C. Iung, "Stability analysis and control synthesis for switched systems: a switched Lyapunov function approach," IEEE Transactions on Automatic Control, vol. 47, no. 11, pp. 1883–1887, 2002.

[6] D. Liberzon and Morse A S. "Basic Problems in stability and design of switched systems," IEEE Control System Magazine, vol. 19, no. 5, pp. 59–70, 1999.

[7] D. Cheng, C. Martin and J. P. Xiang. "An algorithm for common quadratic Lyapunov function," Proceedings of the 3rd World Congress on Intelligent Control and Automation, pp. 2965–2969, 2000.

[8] C. De Persis , R. De Santis and A. S. Morse, "Nonlinear switched systems with state dependent dwell-time," Proceedings of IEEE Conference on Decision and Control, pp. 4419–4424, 2002.

[9] H. Lin and P. J. Antsaklis. "Stability and stabilizability of switched linear systems: a survey of recent results," IEEE Transactions on Automatic Control, vol. 54, no. 2, pp. 308–322, 2009.

[10] X. Lin, H. Du and S. Li, "Finite-time boundedness and $L_2$-gain analysis for switched delay systems with norm-bounded disturbance," Applied Mathematics and Computation, vol. 217, no. 2, pp. 5982–5993, 2011.

[11] G. C. Goodwin and R. H. Middleton, "Rapprochement between continuous and discrete model reference adaptive control," Automatica, vol. 22, no. 2, pp. 199–207, 1986.

[12] S. Chen , R. H. Istepanian, J. Wu and J. Chu, "Comparative study on optimizing closed loop stability bounds of finite-precision controller structures with shift and delta operators," Systems and Control Letters, vol. 40, no. 3, pp. 153–163, 2000.

[13] W. Ebert, "Optimal filtered predictive control a delta operator approach," Systems and Control Letters, vol. 42, no. 1, pp. 69–80, 2001.

[14] Z. R. Xiang and R. H. Wang, "Robust H-infinite control for a class of uncertain switched systems using delta operator," Transactions of the Institute of Measurement and Control, vol. 32, no. 3, pp. 331–344, 2010.

[15] Z. R. Xiang and W. M. Xiang, "Observer design for a class of switched nonlinear systems," Control and Intelligent Systems, vol. 36, no. 4, pp. 318–322, 2008.

[16] S. Pettersson, "Observer design for switched systems using multiple quadratic Lyapunov functions," Proceedings of the 2005 IEEE International Symposium on Intelligent Control, pp. 262–267, 2005.

[17] Z. R. Xiang, R. H. Wang and B. Jiang, "Nonfragile observer for discrete-time switched nonlinear systems with time delay," Circuits Syst Signal Process, vol. 30, pp. 73–87, 2011.

[18] C. H. Lien, "Non-fragile guaranteed cost control for uncertain neutral dynamic systems with time-varying delays in state and control input," Chaos, Solitons & Fractals, vol. 31, no. 26, pp. 889–899, 2007.

[19] M. Vidyasagar, "Nonlinear systems analysis," Prentice-Hall, 1993.

[20] X. Jiang, Q. Han and X. Yu, "Stability criteria for linear discrete time systems with interval-like time-varying delay," American Control Conference, pp. 2817-2822, 2005.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# *Mean-Square Exponential Stabilization of Packet-Based Networked Systems with Time-Varying Transmission Delays, Packet Losses and Input Missing*

Bin Tang, Defeng He, and Yun Zhang
School of Automation
Guangdong University of Technology
Guangzhou, China
tangbin316@163.com

*Abstract*—**This paper is concerned with the mean-square exponential stabilization of packet-based networked systems with time-varying transmission delays, packet losses and input missing. Based on a multivariate i.i.d. model of input delays and a Bernoulli model of input missing, the packet-based networked system is formulated as a switching system with multiple subsystems of different input delays. The sufficient fast-switching conditions are directly established for the closed-loop mean-square exponential stability via an appropriate Lyapunov-Krasovskii approach different from average dwell time approach. The resulting controller design method can be obtained by cone complementarity linearization (CCL). Numerical example is also given to substantiate the effectiveness of our results.**

***Keywords—Networked systems; packet-based; transmission delays; packet losses; input missing***

## I. INTRODUCTION

Networked control systems (NCSs) have attracted much attention in recent years for its advantages over conventional control systems, such as low installation cost, less power consumption, simple maintenance and flexible configuration [1]-[3]. But transmission delays, packet losses, and input missing are common phenomena in NCSs, which often degrade the system performance and even lead to the closed-loop instability.

Transmission delays and packet losses are time-varying and random in nature, and often taken as input delays of NCSs [2], [3]. Bernoulli process is widely used to model discrete transmission delays and packet losses [4]-[6]. By transforming continuous input delays as discrete interval-distributed variables, Bernoulli process is also an appropriate model to explore the binary interval-distributed property of continuous input delays [7]-[9]. Recently, the multivariate independent and identically distributed process (i.i.d) has been proposed as a more realistic and general model to formulate discrete and continuous random input delays [10]-[13]. Based on i.i.d. models, NCSs with time-varying transmission delays and packet losses were essentially switching systems. NCSs with discrete i.i.d. input delays have attracted much research attention and many results have been reported in the literatures, see [10] and [11] for overall system performance conditions dependent and independent on average dwell time of switching subsystems, respectively. But for continuous i.i.d. input delays, most of results are additionally dependent on average dwell time of switching subsystems to guarantee closed-loop mean-square stability, except [7]-[9] where continuous binary i.i.d. input delays, i.e. continuous Bernoulli input delays, were considered. Unfortunately, the results of [7]-[9] could not be extended to the case of continuous multivariate i.i.d. input delays.

Input missing, resulted from control failure, actuator failure or energy-saving consideration, is a common event in NCSs and has an innegligible effect on system performance [14]-[16]. So it is practically important to take it into account in the modeling, analysis and synthesis of NCSs. Input missing is often modeled as Bernoulli process. But all results in [14]-[16] are derived via the average dwell time approach and are essentially slow switching conditions, which would reduce the performance room in system synthesis.

In this paper, the mean-square exponential stabilization problem is considered for packet-based networked systems with time-varying transmission delays, packet losses and input missing. It is known that packet-based scheme is an effective way to compensate time-varying transmission delays and packet losses [12], [13]. But the existing results are all dependent on average dwell time of subsystems. Here input delays are modeled as a continuous multivariate i.i.d. process, and input missing is modeled as a Bernoulli process. Then the NCS is described as a switching system with multiple subsystems of different input delays. A new Lyapunov-Krasovskii approach is proposed for the packet-based networked systems, which takes the distribution information of input delays into account both in constructing the Lyapunov-Krasovskii functional and bounding its infinitesimal. Based on this approach, the fast-switching conditions are directly derived for the closed-loop mean-square

This work was supported in part by the National Natural Science Foundation of China under grants 61104105, U0735003 and 60974047, and the Natural Science Foundation of Guangdong Province under grant 9451009001002702.

exponential stability as well as the resulting controller design method. Numerical examples show the effectiveness and advantage of our results.

*Notation:* The superscript '$T$' denotes matrix transposition. $\mathbb{N}$ and $\mathbb{R}^n$ denote the sets of positive integers and $n\times n$ real matrices, respectively. The notation $P>0$ ($\geq 0$) means that $P$ is real symmetric and positive definite (semi-definite). $I$ and $0$ denote an identity matrix and a zero matrix of appropriate dimensions, respectively. For given positive integers $n_1$ and $n_2$, $I_{n_1}$ denotes the $n_1\times n_1$ identity matrix, and $0_{n_1\times n_2}$ the $n_1\times n_2$ zero matrix. $\mathbb{E}\{\bullet\}$ denotes the mathematical expectation. The space of functions $\phi:[-\bar{\eta},0]\to\mathbb{R}^n$, which are absolutely continuous functions on $[-\bar{\eta},0]$, have a finite $\lim_{\theta\to 0^-}\phi(\theta)$, and have square integrable first-order derivatives is denoted by $\Gamma$ with the norm

$$\|\phi\|_\Gamma = \max_{\theta\in[-\bar{\eta},0]}|\phi(\theta)| + [\int_{-\bar{\eta}}^{0}|\dot{\phi}(s)|^2\, ds]^{1/2}$$

Denote $x_t(\theta)=x(t+\theta)$ where $\theta\in[-\bar{\eta},0]$.

## II. Problem Formulation

Packet-based networked system is plotted in Figure 1. Assume that: 1) the sensor is clock-driven while the controller and the zero-order hold (ZOH) are event-driven, and all of them are connected through network; 2) Feedback and control data packets are transmitted in one packet, and there exist time-varying transmission delays and packet losses in the channels from the sensor to the controller and from the controller to the ZOH; 3) The ZOH suffers from input missing; 4) The controller and the ZOH can determine if a data packet is new, only new data packets are accepted, and $u(t)=0$ before the first updating instant.

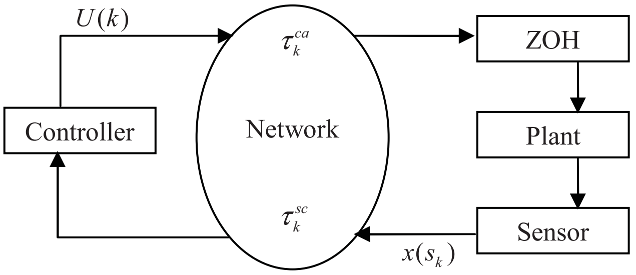


Figure 1. Block diagram of pakcet-based networked systems

Let $t_0$ denotes the initial instant of NCS, $k\in\mathbb{N}$ the number of new data packet arriving at the ZOH, $t_k$ the corresponding arriving instant, $s_k$ the associated sensor's sampling instant, and $\tau_k^{sc}$ and $\tau_k^{ca}$ are the transmission delays from the sensor to the controller and from the controller to the ZOH, respectively. It is seen that $t_k - s_k = \tau_k^{sc}+\tau_k^{ca}$. In the packet-based NCS, when the controller receives a new state feedback $x(s_k)$, it computes a sequence $U(k)$ of controls associated with different input delays, and then sends the sequence in one data packet to the ZOH. The ZOH selects different controls from the received new data packet according to the real input delays of the closed-loop systems, and then performs them.

For $t\in[t_k,t_{k+1})$, let $\tau(t)=t-s_k$ denotes the time-varying input delays induced by transmission delays and packet losses. It is assumed that there exist scalars $\bar{\tau}>0$ and $\underline{\tau}\geq 0$ such that

$$\underline{\tau}\leq\tau(t)<\bar{\tau}\ ,\ t\in[t_k,t_{k+1}) \tag{1}$$

To implement the packet-based framework, the distributed interval $[\underline{\tau},\bar{\tau})$ of input delays interval is divided as

$$[\underline{\tau},\bar{\tau})=\sum\nolimits_{i=1}^{m}[\bar{\tau}_{i-1},\bar{\tau}_i) \tag{2}$$

where $\bar{\tau}_0=\underline{\tau}$ and $\bar{\tau}_m=\bar{\tau}$, and the controller adopts different feedback gains for $\tau(t)$ in different subintervals $[\bar{\tau}_{i-1},\bar{\tau}_i)$, i.e. $\tau(t)\in[\bar{\tau}_{i-1},\bar{\tau}_i)$. Let $\tau_i(t)$ denote the input delays $\tau(t)\in[\bar{\tau}_{i-1},\bar{\tau}_i)$. It is seen that $\tau(t)\in[\underline{\tau},\bar{\tau})$ are now represented as $m$ discrete variables $\tau_i(t)\in[\bar{\tau}_{i-1},\bar{\tau}_i)$.

The plant considered in this paper is of the form

$$\dot{x}(t)=Ax(t)+Bu(t) \tag{3}$$

where $x(t)\in\mathbb{R}^{n_x}$ and $u(t)\in\mathbb{R}^{n_u}$ are state vector and control input vector, respectively, and $A$ and $B$ are constant matrices with appropriate dimensions. In packet-based NCS with input missing, the input is represented as

$$u(t)=\alpha(t)\sum\nolimits_{i=1}^{m}\beta_i(t)u_i(k)\ ,\ t\in[t_k,t_{k+1}) \tag{4}$$

where $u_i(k)$ is a control signal of the sequence $U(k)=[u_1(k), u_2(k),\cdots,u_m(k)]$ and is given as

$$u_i(k)=K_i x(s_k) \tag{5}$$

where $x(s_k)$ is the state feedback, and $K_i$ is the feedback gain associated with $\tau_i(t)$. The random variable $\beta_i(t)$ is defined as follows

$$\beta_i(t)=\begin{cases}1 & \tau(t)\in[\bar{\tau}_{i-1},\bar{\tau}_i)\\ 0 & \tau(t)\notin[\bar{\tau}_{i-1},\bar{\tau}_i)\end{cases} \tag{6a}$$

with the probability $\bar{\beta}_i$ given as

$$\bar{\beta}_i=\Pr\{\tau(t)\in[\bar{\tau}_{i-1},\bar{\tau}_i)\}=\mathbb{E}\{\beta_i(t)\}=\lim_{t\to\infty}\frac{1}{t}\int_{t_1}^{t}\beta_i(s)ds \tag{6b}$$

It is seen that $\{\beta(t)\}$ is actually a continuous multivariate i.i.d. process with respect to $t$, where $\beta(t)=col\{\beta_1(t),\beta_2(t),\cdots,\beta_m(t)\}$. The random variable $\alpha(t)\in\{1,0\}$ denotes if the input missing occurs, where 1 and 0 correspond to the situations with and without input missing, respectively, and the related probability is computed as

$$\bar{\alpha}=\Pr\{\alpha(t)=1\}=\mathbb{E}\{\alpha(t)\}=\lim_{t\to\infty}\frac{1}{t}\int_{t_1}^{t}\alpha(s)ds \tag{7}$$

So $\{\alpha(t)\}$ is a continuous Bernoulli process with respect to $t$. Then the closed-loop packet-based NCS in (3)-(5) is formulated as

$$\dot{x}(t) = Ax(t) + B\alpha(t)\sum\nolimits_{i=1}^{m}\beta_i(t)K_i x(t-\tau_i(t)) \tag{8}$$

It is seen that (8) is essentially a switching system.

The purpose of this paper is to develop the fast-switching conditions of the mean-square exponential stability for the packet-based NCS with time-varying transmission delays, packet losses and input missing basing on (6)-(8). Before giving our main results, we introduce the following definitions.

*Definition 1:* The closed-loop system (3)-(5) is said to be mean square exponentially stable if for any finite initial condition $x_{t_1}$, there exist scalars $\varepsilon > 0$ and $\lambda > 0$ such that

$$\mathrm{E}\{\| x(t) \|^2\} \le \varepsilon e^{-\lambda(t-t_1)} \| x_{t_1} \|_{\Gamma}^2$$

*Definition 2:* The infinitesimal operator is defined as

$$\mathcal{L}V(t,x_t,\dot{x}_t) = \lim_{\vartheta\to 0^+}\frac{1}{\vartheta}\mathrm{E}\{V(t+\vartheta, x_{t+\vartheta}, \dot{x}_{t+\vartheta}) - V(t,x_t,\dot{x}_t)\}$$

where $\mathrm{X}_t = \{x_t\}$.

## III. Main Results

To derive the fast-switching conditions of the mean-square exponential stability of (3)-(5), we construct the following Lyapunov-Krasovskii functional via delay-decomposition approach:

$$V(t,x_t,\dot{x}_t) = V_1(t,x_t) + V_2(t,x_t) + V_3(t,\dot{x}_t) \tag{9a}$$

$$V_1(t,x_t,\dot{x}_t) = x^T(t)Px(t) \tag{9b}$$

$$\begin{aligned} V_2(t,x_t,\dot{x}_t) = {} & \sum_{i=1}^{m_0}\int_{t-\underline{\tau}_i}^{t-\underline{\tau}_{i-1}} e^{\lambda(s-t)}x^T(s)Q_{0,i}x(s)ds \\ & + \sum_{i=1}^{m}\int_{t-\overline{\tau}_i}^{t-\overline{\tau}_{i-1}} e^{\lambda(s-t)}x^T(s)Q_{2,i}x(s)ds \end{aligned} \tag{9c}$$

$$\begin{aligned} V_3(t,\dot{x}_t,\dot{x}_t) = {} & \sum_{i=1}^{m_0}(\underline{\tau}_i - \underline{\tau}_{i-1})\int_{-\underline{\tau}_i}^{-\underline{\tau}_{i-1}}\int_{t+\theta}^{t} e^{\lambda(s-t)}\dot{x}^T(s)R_0\dot{x}(s)dsd\theta \\ & + \sum_{i=1}^{m}(\overline{\tau}_i - \overline{\tau}_{i-1})\int_{-\overline{\tau}_i}^{-\overline{\tau}_{i-1}}\int_{t+\theta}^{t} e^{\lambda(s-t)}\dot{x}^T(s)R_{2,i}\dot{x}(s)dsd\theta \end{aligned} \tag{9d}$$

where $[0,\underline{\tau}) = \bigcup\nolimits_{i=1}^{m_0}[\underline{\tau}_{i-1},\underline{\tau}_i)$, $P > 0$, $Q_{0,i} > 0$ ($i = 1,\ 2,\cdots,m_0$), $R_0 > 0$, $Q_{2,i} > 0$, $R_{2,i} > 0$ ($i = 1,2,\cdots,m$). It is seen from (9) that (8) is not simply deemed as a system with multiple delays $\tau_i(t)$ ($i = 1,2,\cdots,m$) and $\tau_i(t)$ are still taken as a case of $\tau(t)$ belong to $[\overline{\tau}_{i-1},\overline{\tau}_i)$, which is more accurate than the method in [11].

*Theorem 1:* Given (1), (2), (6), (7), $[0,\underline{\tau}) = \bigcup\nolimits_{i=1}^{m_0}[\underline{\tau}_{i-1},\underline{\tau}_i)$, $\rho_i = \overline{\beta}_i \big/ \sum\nolimits_{j=i}^{m}\overline{\beta}_j$ for $\overline{\beta}_i \ne 0$ and $\rho_i = 0$ for $\overline{\beta}_i = 0$ and $K_i$ ($i = 1,2,\cdots,m$), the closed-loop system (3)-(5) is mean-square exponentially stable with a decay rate $\lambda$, if there exist matrices $P > 0$, $Q_{0,i} > 0$ ($i = 1,2,\cdots,m_0$), $Q_{2,i} > 0$, $R_0 > 0$, $R_{2,i} > 0$ ($i = 1,2,\cdots,m$) of appropriate dimensions such that the following inequalities hold:

$$\Omega_1 + \Omega_2 + \Omega_3 + \Omega_4 + \Pi\Lambda_2\Pi^{\mathrm{T}} < 0 \tag{10}$$

for $\sigma_i = 0$ and 1 with $i = 1,2,\cdots,m$, where

$$\Omega_1 = \begin{bmatrix} PA + A^TP + \lambda P & 0_{n_x\times m_0 n_x} & \Omega_{11} \\ * & & \\ * & & 0_{(m_0+2m)n_x\times(m_0+2m)n_x} \end{bmatrix}$$

$$\Omega_{11} = \overline{\alpha}PB\begin{bmatrix} \overline{\beta}_1K_1 & 0 & \overline{\beta}_2K_2 & 0 & \cdots & \overline{\beta}_mK_m & 0 \end{bmatrix}$$

$$\begin{aligned} \Omega_2 = diag\{ & Q_{0,1}, -e^{-\lambda\underline{\tau}_1}Q_{0,1} + e^{-\lambda\underline{\tau}_1}Q_{0,2},\cdots, \\ & -e^{-\lambda\underline{\tau}_{m_0-1}}Q_{0,m_0-1} + e^{-\lambda\underline{\tau}_{m_0-1}}Q_{0,m_0}, \\ & -e^{-\lambda\underline{\tau}_{m_0}}Q_{0,m_0} + e^{-\lambda\overline{\tau}_0}Q_{2,1}, 0, -e^{-\lambda\overline{\tau}_1}Q_{2,1} + e^{-\lambda\overline{\tau}_1}Q_{2,2},\cdots, \\ & 0, -e^{-\lambda\overline{\tau}_{m-1}}Q_{2,m-1} + e^{-\lambda\overline{\tau}_{m-1}}Q_{2,m}, 0, -e^{-\lambda\overline{\tau}_m}Q_{2,m}\} \end{aligned}$$

$$\begin{aligned} \Omega_3 = {} & \Omega_{31} + \left[\begin{array}{cc} 0_{(m_0+2m)n_x\times n_x} & \Omega_{32} \\ \hline \multicolumn{2}{c}{0_{n_x\times(m_0+2m+1)n_x}} \end{array}\right] + \left[\begin{array}{cc} 0_{(m_0+2m)n_x\times n_x} & \Omega_{32} \\ \hline \multicolumn{2}{c}{0_{n_x\times(m_0+2m+1)n_x}} \end{array}\right]^T \\ & + \left[\begin{array}{cc} 0_{(m_0+2m-1)n_x\times 2n_x} & \Omega_{33} \\ \hline \multicolumn{2}{c}{0_{2n_x\times(m_0+2m+1)n_x}} \end{array}\right] + \left[\begin{array}{cc} 0_{(m_0+2m-1)n_x\times 2n_x} & \Omega_{33} \\ \hline \multicolumn{2}{c}{0_{2n_x\times(m_0+2m+1)n_x}} \end{array}\right]^T \end{aligned}$$

$$\begin{aligned} \Omega_{31} = diag\{ & -e^{-\lambda\underline{\tau}_1}R_0, -e^{-\lambda\underline{\tau}_1}R_0 - e^{-\lambda\underline{\tau}_2}R_0,\cdots, \\ & -e^{-\lambda\underline{\tau}_{m_0-1}}R_0 - e^{-\lambda\underline{\tau}_{m_0}}R_0, -e^{-\lambda\underline{\tau}_{m_0}}R_0 \\ & -e^{-\lambda\overline{\tau}_1}(\overline{\beta}_1\rho_1 - \sigma_{2,1}\overline{\beta}_1\rho_1 + 1)R_{2,1}, -3e^{-\lambda\overline{\tau}_1}\overline{\beta}_1\rho_1R_{2,1}, \\ & -e^{-\lambda\overline{\tau}_1}(\sigma_{2,1}\overline{\beta}_1\rho_1 + 1)R_{2,1} - e^{-\lambda\overline{\tau}_2}(\overline{\beta}_2\rho_2 - \sigma_{2,2}\overline{\beta}_2\rho_2 + 1)R_{2,2}, \\ & -3e^{-\lambda\overline{\tau}_2}\overline{\beta}_2\rho_2R_{2,2},\cdots, -e^{-\lambda\overline{\tau}_{m-1}}(\sigma_{2,m-1}\overline{\beta}_{m-1}\rho_{m-1} + 1)R_{2,m-1} \\ & -e^{-\lambda\overline{\tau}_m}(\overline{\beta}_m\rho_m - \sigma_{2,m}\overline{\beta}_m\rho_m + 1)R_{2,m}, \\ & -3e^{-\lambda\overline{\tau}_m}\overline{\beta}_m\rho_mR_{2,m}, -e^{-\lambda\overline{\tau}_m}(\sigma_{2,m}\overline{\beta}_m\rho_m + 1)R_{2,m}\} \end{aligned}$$

$$\begin{aligned} \Omega_{32} = diag\{ & e^{-\lambda\underline{\tau}_1}R_0,\cdots, e^{-\lambda\underline{\tau}_{m_0}}R_0, \\ & e^{-\lambda\overline{\tau}_1}(2 - \sigma_{2,1})\overline{\beta}_1\rho_1R_{2,1}, e^{-\lambda\overline{\tau}_1}(1 + \sigma_{2,1})\overline{\beta}_1\rho_1R_{2,1},\cdots, \\ & e^{-\lambda\overline{\tau}_m}(2 - \sigma_{2,m})\overline{\beta}_m\rho_mR_{2,m}, e^{-\lambda\overline{\tau}_m}(1 + \sigma_{2,m})\overline{\beta}_m\rho_mR_{2,m}\} \end{aligned}$$

$$\begin{aligned} \Omega_{33} = diag\{ & \underbrace{0,\cdots,0}_{m_0}, e^{-\lambda\overline{\tau}_1}(1 - \overline{\beta}_1\rho_1)R_{2,1}, 0,\cdots, \\ & e^{-\lambda\overline{\tau}_{m-1}}(1 - \overline{\beta}_{m-1}\rho_{m-1})R_{2,m-1}, 0, e^{-\lambda\overline{\tau}_m}(1 - \overline{\beta}_m\rho_m)R_{2,m}\} \end{aligned}$$

$$\Pi = \begin{bmatrix} \sqrt{\overline{\alpha}\overline{\beta}_1}A^T & \sqrt{\overline{\alpha}\overline{\beta}_2}A^T & \cdots & \sqrt{\overline{\alpha}\overline{\beta}_m}A^T \\ 0_{m_0n_x\times n_x} & 0_{m_0n_x\times n_x} & \cdots & 0_{m_0n_x\times n_x} \\ \begin{bmatrix} \sqrt{\overline{\alpha}\overline{\beta}_1}K_1^TB^T \\ 0 \end{bmatrix} & 0_{2n_x\times n_x} & \cdots & 0_{2n_x\times n_x} \\ 0_{2n_x\times n_x} & \begin{bmatrix} \sqrt{\overline{\alpha}\overline{\beta}_2}K_2^TB^T \\ 0 \end{bmatrix} & \cdots & 0_{2n_x\times n_x} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{2n_x\times n_x} & 0_{2n_x\times n_x} & \cdots & \begin{bmatrix} \sqrt{\overline{\alpha}\overline{\beta}_m}K_m^TB^T \\ 0 \end{bmatrix} \end{bmatrix}$$

$$\Lambda_2 = diag\{\Lambda_1 \quad \Lambda_1 \quad \cdots \quad \Lambda_1\}$$

$$\Lambda_1 = \sum_{i=1}^{m_0}(\underline{\tau}_i - \underline{\tau}_{i-1})^2 R_0 + \sum_{i=1}^{m}(\overline{\tau}_i - \overline{\tau}_{i-1})^2 R_{2,i}$$

$$\Omega_4 = \begin{bmatrix} (1-\overline{\alpha})A^T\Lambda_1 A & 0_{n_x\times(m_0+2m)n_x} \\ * & 0_{(m_0+2m)n_x\times(m_0+2m)n_x} \end{bmatrix}.$$

*Proof:* Along the trajectories of the closed-loop system (8), it follows from (9) that

$$\mathcal{L}V(t,x_t,\dot{x}_t) = \mathcal{L}V_1(t,x_t) + \mathcal{L}V_2(t,x_t) + \mathcal{L}V_3(t,\dot{x}_t) \qquad (11)$$

where

$$\begin{aligned}\mathcal{L}V_1(t,x_t) &= -\lambda V_1(t,x_t) + x^T(t)\lambda Px(t) \\ &\quad +2x^T(t)P\{Ax(t) + \overline{\alpha}\sum_{i=1}^{m}\overline{\beta}_i(k)BK_i x(t-\tau_i(t))\}\end{aligned}$$

$$\begin{aligned}\mathcal{L}V_2(t,x_t) &= -\lambda V_2(t,x_t) + \mathbb{E}\{\sum_{i=1}^{m_0} x^T(t-\underline{\tau}_{i-1})e^{-\lambda\underline{\tau}_{i-1}}Q_{0,i}x(t-\underline{\tau}_{i-1}) \\ &\quad -x^T(t-\underline{\tau}_i)e^{-\lambda\underline{\tau}_i}Q_{0,i}x(t-\underline{\tau}_i)\big|\mathrm{X}_t\} \\ &\quad +\mathbb{E}\{\sum_{i=1}^{m} x^T(t-\overline{\tau}_{i-1})e^{-\lambda\overline{\eta}_{i-1}}Q_{2,i}x(t-\overline{\tau}_{i-1}) \\ &\quad -x^T(t-\overline{\tau}_i)e^{-\lambda\overline{\eta}_i}Q_{2,i}x(t-\overline{\tau}_i)\big|\mathrm{X}_t\}\end{aligned}$$

$$\begin{aligned}\mathcal{L}V_3(t,\dot{x}_t) &= -\lambda V_3(t,\dot{x}_t) + \mathbb{E}\{\dot{x}^T(t)\Lambda_1\dot{x}(t)\big|\mathrm{X}_t\} \\ &\quad +\mathbb{E}\{\sum_{i=1}^{m_0} -(\underline{\tau}_i - \underline{\tau}_{i-1})\int_{t-\underline{\tau}_i}^{t-\underline{\tau}_{i-1}} \dot{x}^T(s)e^{\lambda(s-t)}R_0\dot{x}(s)ds\big|\mathrm{X}_t\} \\ &\quad +\mathbb{E}\{\sum_{i=1}^{m} -(\overline{\tau}_i - \overline{\tau}_{i-1})\int_{t-\overline{\eta}_i}^{t-\overline{\eta}_{i-1}} \dot{x}^T(s)e^{\lambda(s-t)}R_{2,i}\dot{x}(s)ds\big|\mathrm{X}_t\}.\end{aligned}$$

With the closed-loop system in (8), it follows that

$$\begin{aligned}&\mathbb{E}\left\{\dot{x}^T(t)\Lambda_1\dot{x}(t)\big|\mathrm{X}_t\right\} \\ &= (1-\overline{\alpha})x^T(t)A^T\Lambda_1 Ax(t) \\ &+\sum_{i=1}^{m}\overline{\alpha}\overline{\beta}_i\left\{x^T(t)A^T\Lambda_1 Ax(t) + 2x^T(t)A^T\Lambda_1 BK_i x(t-\tau_i(t))\right. \qquad (12) \\ &\left.+x^T(t-\tau_i(t))K_i^T B^T\Lambda_1 BK_i x(t-\tau_i(t))\right\} \\ &= \xi^T(t)\Omega_4\xi(t) + \xi^T(t)\Pi\Lambda_2\Pi^{\mathrm{T}}\xi(t)\end{aligned}$$

where

$$\begin{aligned}\xi(t) = col\{&x(t), x(t-\underline{\tau}_1), \cdots, x(t-\underline{\tau}_{m_0}), \\ &x(t-\tau_1(t)), x(t-\overline{\tau}_1), \cdots, x(t-\tau_m(t)), x(t-\overline{\tau}_m)\}.\end{aligned}$$

By Jensen's inequality, it follows that for $i = 1,2,\cdots,m_0$

$$\begin{aligned}&\mathbb{E}\{-(\underline{\tau}_i - \underline{\tau}_{i-1})\int_{t-\underline{\tau}_i}^{t-\underline{\tau}_{i-1}} e^{\lambda(s-t)}\dot{x}^T(s)R_0\dot{x}(s)ds\big|\mathrm{X}_t\} \\ &\le e^{-\lambda\underline{\tau}_i}\begin{bmatrix} x(t-\underline{\tau}_{i-1}) \\ x(t-\underline{\tau}_i)\end{bmatrix}^T \begin{bmatrix} -R_0 & R_0 \\ R_0 & -R_0\end{bmatrix}\begin{bmatrix} x(t-\underline{\tau}_{i-1}) \\ x(t-\underline{\tau}_i)\end{bmatrix}.\end{aligned} \qquad (13)$$

In this paper the bounding of the other integrals associated with $R_{2,i}$ in (11) takes the distribution of $\tau(t)$ into account. For this purpose, the following decomposition is made:

$$\begin{aligned}&\mathbb{E}\{-(\overline{\tau}_i - \overline{\tau}_{i-1})\int_{t-\overline{\tau}_i}^{t-\overline{\tau}_{i-1}} \dot{x}^T(s)e^{\lambda(s-t)}R_{2,i}\dot{x}(s)ds\big|\mathrm{X}_t\} \\ &= \rho_i(k)\mathbb{E}\{-(\overline{\tau}_i - \overline{\tau}_{i-1})\int_{t-\overline{\tau}_i}^{t-\overline{\tau}_{i-1}} \dot{x}^T(s)e^{\lambda(s-t)}R_{2,i}\dot{x}(s)ds\big|\mathrm{X}_t\} \\ &+(1-\rho_i(k))\mathbb{E}\{-(\overline{\tau}_i - \overline{\tau}_{i-1})\int_{t-\overline{\tau}_i}^{t-\overline{\tau}_{i-1}} \dot{x}^T(s)e^{\lambda(s-t)}R_{2,i}\dot{x}(s)ds\big|\mathrm{X}_t\}\end{aligned} \qquad (14)$$

for $i = 1,2,\cdots,m$. In (14), the first decomposing term is related to the input delays $\tau_i(t)$, and the second one is related to other input delays $\tau_j(t)$ with $j = i+1,\cdots,m$.

As seen in [17], different ways are needed in bounding

$$-\int_{t-\overline{\tau}_i}^{t-\overline{\tau}_{i-1}} \dot{x}^T(s)e^{\lambda(s-t)}R_{2,i}\dot{x}(s)ds$$

with respect to $\tau_i(k,t) \in [\overline{\tau}_{i-1}, \overline{\tau}_i)$ and $\tau_i(k,t) \notin [\overline{\tau}_{i-1}, \overline{\tau}_i)$, respectively. Based on this idea, it follows from Jensen's inequality and the technique similar to that of [17] that

$$\begin{aligned}&\rho_i(k)\mathbb{E}\{-(\overline{\tau}_i - \overline{\tau}_{i-1})\int_{t-\overline{\tau}_i}^{t-\overline{\tau}_{i-1}} \dot{x}^T(s)e^{\lambda(s-t)}R_{2,i}\dot{x}(s)ds\big|\mathrm{X}_t\} \\ &\le \zeta_i^T(t)e^{-\lambda\overline{\tau}_i}\rho_i(k) \\ &\times\left\{\overline{\beta}_i(k)\begin{bmatrix} -(2-\sigma_i)R_{2,i} & (2-\sigma_i)R_{2,i} & 0 \\ (2-\sigma_i)R_{2,i} & -3R_{2,i} & (1+\sigma_i)R_{2,i} \\ 0 & (1+\sigma_i)R_{2,i} & -(1+\sigma_i)R_{2,i}\end{bmatrix}\right. \\ &\left.+(1-\overline{\beta}_i(k))\begin{bmatrix} -R_{2,i} & 0 & R_{2,i} \\ 0 & 0 & 0 \\ R_{2,i} & 0 & -R_{2,i}\end{bmatrix}\right\}\zeta_i(t) \\ &= \zeta_i^T(t)e^{-\lambda\overline{\tau}_i}\rho_i(k)\begin{bmatrix} \Upsilon_i & (2-\sigma_i)\overline{\beta}_i R_{2,i} & (1-\overline{\beta}_i)R_{2,i} \\ * & -3\overline{\beta}_i R_{2,i} & (1+\sigma_i)\overline{\beta}_i R_{2,i} \\ * & * & -(\sigma_i\overline{\beta}_i+1)R_{2,i}\end{bmatrix}\zeta_i(t)\end{aligned} \qquad (15)$$

and

$$\begin{aligned}&(1-\rho_i(k))\mathbb{E}\{-(\overline{\tau}_i - \overline{\tau}_{i-1})\int_{t-\overline{\tau}_i}^{t-\overline{\tau}_{i-1}} \dot{x}^T(s)e^{\lambda(s-t)}R_{2,i}\dot{x}(s)ds\big|\mathrm{X}_t\} \\ &\le \zeta_i^T(t)e^{-\lambda\overline{\tau}_i}(1-\rho_i(k))\begin{bmatrix} -R_{2,i} & 0 & R_{2,i} \\ 0 & 0 & 0 \\ R_{2,i} & 0 & -R_{2,i}\end{bmatrix}\zeta_i(t).\end{aligned} \qquad (16)$$

where $\zeta_i^T(t) = [x^T(t-\overline{\tau}_{i-1}), x^T(t-\tau_i(t)), x^T(t-\overline{\tau}_i)]$, $\Upsilon_i = -(\overline{\beta}_i -\sigma_i\overline{\beta}_i + 1)R_{2,i}$, $\sigma_i = (\tau_i(t) - \overline{\tau}_{i-1})/(\overline{\tau}_i - \overline{\tau}_{i-1})$. It is seen that (15) further takes the distribution of $\tau_i(t)$ into account. From (15) and (16), we have the following

$$\begin{aligned}&\mathbb{E}\{-(\overline{\tau}_i - \overline{\tau}_{i-1})\int_{t-\overline{\tau}_i}^{t-\overline{\tau}_{i-1}} \dot{x}^T(s)e^{\lambda(s-t)}R_{2,i}\dot{x}(s)ds\big|\mathrm{X}_t\} \\ &= \zeta_i^T(t)e^{-\lambda\overline{\tau}_i} \\ &\times\left\{\overline{\beta}_i\rho_i\begin{bmatrix} -(1-\sigma_i)R_{2,i} & (2-\sigma_i)R_{2,i} & -R_{2,i} \\ (2-\sigma_i)R_{2,i} & -3R_{2,i} & (1+\sigma_{2,i})R_{2,i} \\ -R_{2,i} & (1+\sigma_i)R_{2,i} & -\sigma_i R_{2,i}\end{bmatrix}\right.\end{aligned}$$

$$+\begin{bmatrix} -R_{2,i} & 0 & R_{2,i} \\ 0 & 0 & 0 \\ R_{2,i} & 0 & -R_{2,i} \end{bmatrix}\zeta_i(t)\Bigg\}. \tag{17}$$

If inequality (10) holds with $\sigma_i = 0$ and 1 for $i = 1, 2, \cdots, m$, it follows from (11)-(13), (17) that

$$\begin{aligned} &\mathfrak{L}V(t, x_t, \dot{x}_t) + \lambda V(t, x_t, \dot{x}_t) \le \xi^T(t)(\Omega_1 + \Omega_2 + \Omega_3 \\ &+\Omega_4 + \Pi\Lambda_2\Pi^{\mathrm{T}})\xi(t) < 0 \end{aligned} \tag{18}$$

which means that $\mathbb{E}\{V(t, x_t, \dot{x}_t)\} \le e^{-\lambda(t-t_1)}\mathbb{E}\{V(t_1, x_{t_1}, \dot{x}_{t_1})\}$. So it can be concluded by Definition 1 that the closed-loop system (3)-(5) is mean-square exponentially stable under the given conditions of Theorem 1. The proof of Theorem 1 is completed.

*Theorem 2:* Given (1), (2), (6), (7), $[0, \underline{\tau}) = \bigcup_{i=1}^{m_0}[\underline{\tau}_{i-1}, \underline{\tau}_i)$, $\rho_i = \bar{\beta}_i \big/ \sum_{j=i}^{m}\bar{\beta}_j$ for $\bar{\beta}_i \ne 0$ and $\rho_i = 0$ for $\bar{\beta}_i = 0$ ( $i = 1, 2, \cdots, m$ ), the closed-loop system (3)-(5) is mean-square exponentially stabilizable with a decay rate $\lambda$, if there exist matrices $X > 0$, $\tilde{Q}_{0,i} > 0$ $(i = 1, 2, \cdots, m_0)$, $\tilde{Q}_{2,i} > 0$, $\tilde{R}_0 > 0$, $\tilde{R}_{2,i} > 0$, $Y_i$ $(i = 1, 2, \cdots, m)$ of appropriate dimensions such that the following inequalities hold:

$$\begin{bmatrix} \tilde{\Omega}_1 + \tilde{\Omega}_2 + \tilde{\Omega}_3 & \tilde{\Omega}_4\Phi_1 & \tilde{\Pi}\Phi_2 \\ * & -\tilde{\Lambda}_1 & 0 \\ * & * & -\tilde{\Lambda}_2 \end{bmatrix} < 0 \tag{19}$$

for $\sigma_i = 0$ and 1 with $i = 1, 2, \cdots, m$, where

$$\tilde{\Omega}_1 = \begin{bmatrix} AX + XA^T + \lambda X & 0_{n_x \times m_0 n_x} & \tilde{\Omega}_{11} \\ * & & \\ * & & 0_{(m_0+2m)n_x \times (m_0+2m)n_x} \end{bmatrix}$$

$$\tilde{\Omega}_{11} = \bar{\alpha}B\begin{bmatrix} \bar{\beta}_1 Y_1 & 0 & \bar{\beta}_2 Y_2 & 0 & \cdots & \bar{\beta}_m Y_m & 0 \end{bmatrix}$$

$$\begin{aligned} \tilde{\Omega}_2 = diag\{&\tilde{Q}_{0,1}, -e^{-\lambda\underline{\tau}_1}\tilde{Q}_{0,1} + e^{-\lambda\underline{\tau}_1}\tilde{Q}_{0,2}, \cdots, \\ &-e^{-\lambda\underline{\tau}_{m_0-1}}\tilde{Q}_{0,m_0-1} + e^{-\lambda\underline{\tau}_{m_0-1}}\tilde{Q}_{0,m_0}, \\ &-e^{-\lambda\underline{\tau}_{m_0}}\tilde{Q}_{0,m_0} + e^{-\lambda\bar{\tau}_0}\tilde{Q}_{2,1}, 0, -e^{-\lambda\bar{\tau}_1}\tilde{Q}_{2,1} + e^{-\lambda\bar{\tau}_1}\tilde{Q}_{2,2}, \cdots, \\ &0, -e^{-\lambda\bar{\tau}_{m-1}}\tilde{Q}_{2,m-1} + e^{-\lambda\bar{\tau}_{m-1}}\tilde{Q}_{2,m}, 0, -e^{-\lambda\bar{\tau}_m}\tilde{Q}_{2,m}\} \end{aligned}$$

$$\begin{aligned} \tilde{\Omega}_3 = \tilde{\Omega}_{31} &+ \left[\begin{array}{cc} 0_{(m_0+2m)n_x \times n_x} & \tilde{\Omega}_{32} \\ \hline \multicolumn{2}{c}{0_{n_x \times (m_0+2m+1)n_x}} \end{array}\right] + \left[\begin{array}{cc} 0_{(m_0+2m)n_x \times n_x} & \tilde{\Omega}_{32} \\ \hline \multicolumn{2}{c}{0_{n_x \times (m_0+2m+1)n_x}} \end{array}\right]^T \\ &+ \left[\begin{array}{cc} 0_{(m_0+2m-1)n_x \times 2n_x} & \tilde{\Omega}_{33} \\ \hline \multicolumn{2}{c}{0_{2n_x \times (m_0+2m+1)n_x}} \end{array}\right] + \left[\begin{array}{cc} 0_{(m_0+2m-1)n_x \times 2n_x} & \tilde{\Omega}_{33} \\ \hline \multicolumn{2}{c}{0_{2n_x \times (m_0+2m+1)n_x}} \end{array}\right]^T \end{aligned}$$

$$\begin{aligned} \tilde{\Omega}_{31} = diag\{&-e^{-\lambda\underline{\tau}_1}\tilde{R}_0, -e^{-\lambda\underline{\tau}_1}\tilde{R}_0 - e^{-\lambda\underline{\tau}_2}\tilde{R}_0, \cdots, \\ &-e^{-\lambda\underline{\tau}_{m_0-1}}\tilde{R}_0 - e^{-\lambda\underline{\tau}_{m_0}}\tilde{R}_0, -e^{-\lambda\underline{\tau}_{m_0}}\tilde{R}_0 \\ &-e^{-\lambda\bar{\tau}_1}(\bar{\beta}_1\rho_1 - \sigma_{2,1}\bar{\beta}_1\rho_1 + 1)\tilde{R}_{2,1}, -3e^{-\lambda\bar{\tau}_1}\bar{\beta}_1\rho_1\tilde{R}_{2,1}, \\ &-e^{-\lambda\bar{\tau}_1}(\sigma_{2,1}\bar{\beta}_1\rho_1 + 1)\tilde{R}_{2,1} - e^{-\lambda\bar{\tau}_2}(\bar{\beta}_2\rho_2 - \sigma_{2,2}\bar{\beta}_2\rho_2 + 1)\tilde{R}_{2,2}, \\ &-3e^{-\lambda\bar{\tau}_2}\bar{\beta}_2\rho_2\tilde{R}_{2,2}, \cdots, -e^{-\lambda\bar{\tau}_{m-1}}(\sigma_{2,m-1}\bar{\beta}_{m-1}\rho_{m-1} + 1)\tilde{R}_{2,m-1} \\ &-e^{-\lambda\bar{\tau}_m}(\bar{\beta}_m\rho_m - \sigma_{2,m}\bar{\beta}_m\rho_m + 1)\tilde{R}_{2,m}, \\ &-3e^{-\lambda\bar{\tau}_m}\bar{\beta}_m\rho_m R_{2,m}, -e^{-\lambda\bar{\tau}_m}(\sigma_{2,m}\bar{\beta}_m\rho_m + 1)\tilde{R}_{2,m}\} \end{aligned}$$

$$\begin{aligned} \tilde{\Omega}_{32} = diag\{&e^{-\lambda\underline{\tau}_1}\tilde{R}_0, \cdots, e^{-\lambda\underline{\tau}_{m_0}}\tilde{R}_0, \\ &e^{-\lambda\bar{\tau}_1}(2 - \sigma_{2,1})\bar{\beta}_1\rho_1\tilde{R}_{2,1}, e^{-\lambda\bar{\tau}_1}(1 + \sigma_{2,1})\bar{\beta}_1\rho_1\tilde{R}_{2,1}, \cdots, \\ &e^{-\lambda\bar{\tau}_m}(2 - \sigma_{2,m})\bar{\beta}_m\rho_m\tilde{R}_{2,m}, e^{-\lambda\bar{\tau}_m}(1 + \sigma_{2,m})\bar{\beta}_m\rho_m\tilde{R}_{2,m}\} \end{aligned}$$

$$\begin{aligned} \tilde{\Omega}_{33} = diag\{&\underbrace{0, \cdots, 0}_{m_0}, e^{-\lambda\bar{\tau}_1}(1 - \bar{\beta}_1\rho_1)\tilde{R}_{2,1}, 0, \cdots, \\ &e^{-\lambda\bar{\tau}_{m-1}}(1 - \bar{\beta}_{m-1}\rho_{m-1})\tilde{R}_{2,m-1}, 0, e^{-\lambda\bar{\tau}_m}(1 - \bar{\beta}_m\rho_m)\tilde{R}_{2,m}\} \end{aligned}$$

$$\tilde{\Omega}_4 = \left[\begin{array}{c} \sqrt{1-\bar{\alpha}}XA^T, \cdots, \sqrt{1-\bar{\alpha}}XA^T \\ \hline 0_{(m_0+2m)n_x \times (1+m)n_x} \end{array}\right]$$

$$\tilde{\Lambda}_2 = diag\{\underbrace{\tilde{\Lambda}_1, \tilde{\Lambda}_1, \cdots, \tilde{\Lambda}_1}_{m}\}$$

$$\tilde{\Lambda}_1 = diag\{X\tilde{R}_0^{-1}X, X\tilde{R}_{2,1}^{-1}X, \cdots, X\tilde{R}_{2,m}^{-1}X\}$$

$$\Phi_2 = diag\{\underbrace{\Phi_1, \Phi_1, \cdots, \Phi_1}_{m}\}$$

$$\Phi_1 = diag\{\sqrt{\textstyle\sum_{i=1}^{m_0}(\underline{\tau}_i - \underline{\tau}_{i-1})^2}, \bar{\tau}_1 - \bar{\tau}_0, \cdots, \bar{\tau}_m - \bar{\tau}_{m-1}\}$$

$$\tilde{\Pi} = [\tilde{\Pi}_1, \tilde{\Pi}_2, \cdots, \tilde{\Pi}_m]$$

$$\tilde{\Pi}_i = \begin{bmatrix} \underbrace{\begin{matrix} \sqrt{\bar{\alpha}\bar{\beta}_i}XA^T & & \sqrt{\bar{\alpha}\bar{\beta}_i}XA^T \\ 0_{(m_0+2i-2))n_x \times n_x} & \cdots & 0_{(m_0+2i-2))n_x \times n_x} \\ \sqrt{\bar{\alpha}\bar{\beta}_i}Y_i^T B^T & & \sqrt{\bar{\alpha}\bar{\beta}_i}Y_i^T B^T \\ 0_{(2m-2i+1)n_x \times n_x} & & 0_{(2m-2i+1)n_x \times n_x} \end{matrix}}_{1+m} \end{bmatrix}.$$

Furthermore, the controller gains are given by $K_i = Y_i X^{-1}$.

*Proof:* Applying the well known Schur Lemma and congruence transformation to (10) yields Theorem 2.

Inequality (19) is nonlinear for the existing of $X\tilde{R}_{2,i}^{-1}X$, $i = 1, 2, \cdots, m$, and can not be directly solved by LMI toolbox. The feasible problem of nonlinear inequality in Theorem 2 can be transformed by CCL algorithm into a nonlinear convex optimization problem subject to LMI constraints, which can be directly solved by LMI toolbox.

## IV. NUMERICAL EXAMPLE

Consider the linear system in [12] with

$$A = \begin{bmatrix} -1 & 0 & -0.5 \\ 1 & -0.5 & 0 \\ 0 & 0 & 0.5 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Let $m_0 = 1$, $m = 6$, $\underline{\tau} = 0$, $\bar{\tau}_i - \bar{\tau}_{i-1} = (\bar{\tau} - \underline{\tau})/m$, $\bar{\beta}_i = 1/m$ with $i = 1, 2, \cdots, m$. It is obtained by Theorem 2 that:

*Case 1:* Given $\bar{\alpha} = 1$, $\lambda = 0$, the maximum value of $\bar{\tau}$ is 2.64 with the following feedback gains

$$K_1 = [-0.0015, 0.0002, -0.5285]$$
$$K_2 = [-0.0005, -0.0001, -0.5309]$$
$$K_3 = [-0.0002, -0.0002, -0.5287]$$
$$K_4 = [0.0000, -0.0003, -0.5235]$$
$$K_5 = [0.0002, -0.0003, -0.5172]$$
$$K_6 = [0.0008, -0.0005, -0.5132].$$

*Case 2:* Given $\overline{\alpha} = 0.8$, $\lambda = 0$, the maximum value of $\overline{\tau}$ is $1.70$ with the following feedback gains

$$K_1 = [-0.0045, -0.0031, -0.9652]$$
$$K_2 = [0.0068, -0.0068, -0.9844]$$
$$K_3 = [0.0103, -0.0078, -0.9693]$$
$$K_4 = [0.0122, -0.0082, -0.9269]$$
$$K_5 = [0.0140, -0.0085, -0.8719]$$
$$K_6 = [0.0167, -0.0092, -0.8386].$$

*Case 3:* Given $\overline{\alpha} = 0.8$, $\overline{\tau} = 1.0$, the maximum value of $\lambda$ is $0.44$ with the following feedback gains

$$K_1 = [0.0033, -0.0021, -1.4844]$$
$$K_2 = [0.0095, -0.0024, -1.5318]$$
$$K_3 = [0.0111, -0.0024, -1.5150]$$
$$K_4 = [0.0116, -0.0023, -1.4373]$$
$$K_5 = [0.0117, -0.0022, -1.3199]$$
$$K_6 = [0.0126, -0.0021, -1.2355].$$

## V. CONCLUSION

This paper gives the fast-switching conditions for packet-based NCSs with time-varying transmission delays, packet losses and input missing, which are independent on average dwell time of subsystems. With a multivariate i.i.d. model of input delays and a Bernoulli model of input missing, the NCS is formulated as a switching system with multiple subsystems of different input delays. Appropriate Lyapunov-Krasovskii functional is constructed by a delay decomposition approach. By using the bounding techniques based on Jensen's inequality as in [17], the distribution of input delays is taken into account when bounding the integral terms of the infinitesimal of the functional. The resulting controller design method can be solved by CCL algorithm. Numerical example shows the effectiveness of our results.

## REFERENCES

[1] P. Antsaklis and J. Baillieul, "Special issue on technology of networked control systems," Proceedings of the IEEE, vol. 95, no. 1, pp. 5-8, 2007.

[2] H. J. Gao, T. W. Chen, and J. Lam, "A new delay system approach to network-based control," Automatica, vol. 44, no. 1, pp. 39-52, 2008.

[3] X. F. Jiang, Q. L. Han, S. R. Liu, and A. K. Xue, "A new H-infinity stabilization criterion for networked control systems," IEEE Transactions on Automatic Control, vol. 53, no. 4, pp. 1025-1032, 2008.

[4] F. W. Yang, Z. D. Wang, Y. S. Hung, and M. Gani, "H-infinity control of networked systems with random communication delays," IEEE Transactions on Automatic Control, vol. 51, no. 3, pp. 511-518, 2006.

[5] R. N. Yang, P. Shi, G. P. Liu, and H. J. Gao, "Network-based feedback control for systems with mixed delays based on quantization and dropout compensation," Automatica, vol. 47, no. 12, pp. 2805-2809, 2011.

[6] R. N. Yang, P. Shi, and G. P. Liu, "Filtering for discrete-time networked nonlinear systems with mixed random delays and packet dropouts," IEEE Transactions on Automatic Control, vol. 56, no. 11, pp. 2655-2660, 2011.

[7] C. Peng, D. Yue, E. G. Tian, and Z. Gu, "A delay distribution based stability analysis and synthesis approach for networked control systems," Journal of the Franklin Institute, vol. 346, no. 4, pp. 349-365, 2009.

[8] C. Lin, Z. D. Wang, and F. W. Yang, "Observer-based networked control for continuous- time systems with random sensor delays," Automatica, vol. 45, no. 2, pp. 578-584, 2009.

[9] C. Peng and T. C. Yang, "Communication-delay-distribution-dependent networked control for a class of T-S fuzzy systems," IEEE Transactions on Fuzzy Systems, vol. 18, no. 2, pp. 326-335, 2010.

[10] W. A. Zhang and L. Yu, "Modeling and control of networked control systems with both network-induced delay and packet-dropout," Automatica, vol. 44, no. 12, pp. 3206-3210, 2008.

[11] H. J. Gao, X. Y. Meng, and T. W. Chen, "Stabilization of networked control systems with a new delay characterization," IEEE Transactions on Automatic Control, vol. 53, no. 9, pp. 2142-2148, 2008.

[12] Y. B. Zhao, G. P. Liu, and D. Rees, "Modeling and stabilization of continuous-time packet-based networked control systems," IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics, vol. 39, no. 6, pp. 1646-1652, 2009.

[13] R. Wang, B. Wang, G. P. Liu, W. Wang, and D. Rees, "Hinfinity controller design for networked predictive control systems based on the average dwell time approach," IEEE Transactions on Circuits ans Systems-II: Express Briefs, vol. 57, no. 4, pp. 310-314, 2010.

[14] X. M. Sun, G. P. Liu, D. Rees, and W. Wang, "Stability of systems with controller failure and time-varying delay," IEEE Transactions on Automatic Control, vol. 53, no. 10, pp. 2391-2396, 2008.

[15] W. A. Zhang and L.Yu, "Stabilization of sampled-data control systems with control inputs missing," IEEE Transactions on Automatic Control, vol. 55, no. 2, pp. 447-452, 2010.

[16] X. M. Sun, G. P. Liu, W. Wang, and D. Rees, "L2 gain of systems with input delays and controller temporary failure: zero-order hold model," IEEE Transactions on Control Systems Technology, vol. 19, no. 3, pp. 699-706, 2011.

[17] J. Sun, G. P. Liu, J. Chen, and D. Rees, "Improved delay-range-dependent stability criteria for linear systems with time-varying delays," Automatica, vol. 46, no. 2, pp. 466-470, 2010.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Overview on Computer Forensics Tools

Raza Hasan
Computer Engineering Department
Sir Syed University of Engineering & Technology
Karachi, Pakistan
raza_6@hotmail.com

Salman Mahmood
Computer Engineering Department
Sir Syed University of Engineering & Technology
Karachi, Pakistan
salmanm@ssuet.edu.pk

Akshyadeep Raghav
School of Computing
Staffordshire University
Stafford, UK
akshyadeep@gmail.com

*Abstract***— Different tools are used to aid the investigation process. The need of specialized software is required for the acquisition and examination of data gathered from the crime scene. To abide by chain of custody proper crime scene reconstruction or image is acquired from the original source that can be admissible to the court. This paper focuses on the various hardware and software tools that are widely used during a Computer Forensics Investigation.**

***Keywords-component; Computer; Forensics; Investigation; Hardware; Software; Crime; Tool;***

## I. INTRODUCTION

A computer crime is defined as a criminal act in which people commit the offence using the digital knowledge stored in the computer system. To investigate the computer based crime a new field of specialization - forensic computing has been developed, which is the process of computer investigation and analysis technique to gather evidence in a manner that is legally acceptable [1].

Computers and internet continue to spread and occupy our lives by increasing the potential of harm caused by it through increased number of computer crimes. To deal with this rise, new and advance methods of investigations are required. Electronic evidence in the form of data or information of investigative importance is stored or transmitted digitally by means of electronic devices. Electronic evidence by its nature is very fragile. It can be damaged, destroyed or altered by improper handling and examination. For this reason, special precautions or set of rules have to be followed in acquiring, analyzing and reporting this type of evidence, failure to do so may result an inaccurate conclusion. Electronic evidence poses special challenges regarding its admissibility in court [2].

An important tool used by investigators to safeguard evidence, is called chain of custody. Essentially, this means accounting for those who has touched a given piece of evidence, when they touch it and what they did to evidence. It's a way of demonstrating that evidence hasn't been damaged or tampered with while in the care of the investigator. In the book, Criminalistics: An Introduction to Forensic Science, Richard Saferstein notes in 7th Edition (July 31, 2000), "Failure to substantiate the evidence's chain of custody may lead to serious questions regarding the authenticity and integrity of the evidence and the examinations rendered upon it (pg. 48)." As one would imagine, changes to the chain of custody can quickly ruin a case [3].

The paper is organized as follows: Section II describes the scenerio on which the investigation focuses. Section III describes the Hardware tools used in the investigation. Section IV describes the Software tools used in the investigation. Section V concludes the paper.

## II. SCENERIO

The use of computer forensics may be able to show a pattern of activity that will aid in a lawsuit. Through the use of e-mails, deleted data, and other items found during an investigation, the computer forensics specialist can piece together a history of the prior employee's computer usage and this type of information is very useful in a court.

For example, in an Advertising company they have recently hired a new salesman. Six months after his hire, he leaves the company and forms a competing interest, sending letters or client contacts to another competitive company. The organisation might think this a bit odd and contact an attorney to consider filing a suit. What has occurred is a virtual theft; the salesman stole a copy of your client database. Note that this is a VIRTUAL theft, since you were not deprived of any property (he didn't delete it, just copied it) you will likely not be able to prosecute him criminally.

In order to prove him guilty under law the use of Computer Forensics could be put to practise. The Computer Forensics Tools are of two kinds Hardware tools and Software tools. Data recovery is as much art as it is science. Using industry standard tools, computer data once thought to be lost is restored either in full or at least in part.

When a file is deleted, the space it occupied on the hard drive is not initially overwritten. Additionally, there are snippets of data, previous versions of documents, and other content that may be scattered throughout the hard drive. Computer Forensics experts can recover the lost

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

data. Sometimes this is a simple undelete whereas other times it takes a considerable amount of effort to piece the file back together. In addition to recovering deleted files, we can also break password protected files (such as a Word document for example) and in some cases encrypted files [4].

In order to carry out these procedures of retrieval of data a Computer Forensics expert would use various Hardware and Software Tools, which would be discussed in next section.

## III. HARDWARE TOOLS

A Computer Forensics expert should be aware and familiar with the inside of a computer system. One should know the inside and outside of the system before they could work on the tools to retrieve data. They should have a good knowledge of the hard drives and their settings. There are many hardware tools that could be used by a Forensics expert, this section would be addressing and discussing about FRED in this chapter. FRED is the most common hardware device used by most investigators. FRED stands for Forensic Recovery of Evidence Device. The FRED families of forensic workstations are highly integrated, flexible and modular forensic platforms and now include DI's exclusive ***UltraBay Write Protected Imaging Bay***. There are a lot of versions of FRED like FRED SR, FRED L, etc... [5]

### A. *FRED System*

FRED systems are optimized for stationary laboratory acquisition and analysis. Simply remove the hard drive(s) from the suspect system and plug them into FRED and acquire the digital evidence. FRED will acquire data directly from IDE/EIDE/ ATA/SATA/ATAPI/SCSI I/SCSI II/SCSI III hard drives and storage devices and save forensic images to DVD, CD or hard drives. FRED systems also acquire data from floppies, 100/250/750 MB ZIP cartridges, CD-ROM, DVD-ROM, Compact Flash, Micro Drives, Smart Media, Memory Stick, Memory Stick Pro, xD Cards, Secure Digital Media and Multimedia Cards. Furthermore, with the optional tape drive FRED is capable of archiving to or acquiring evidence from 4mm DAT tapes. With the RAID option FRED has an incredible 1.6 TB (1600 GB) of internal RAID storage. All FRED systems include the UltraBay, custom front panel connections, and removable drive trays so there is no need to open up the processing system to install drives or crawl around the back of the unit to attach devices. Fig 1 illustrates a FRED system; its estimate cost would be $5999.00 [5].

Figure 1. FRED System [5]

#### 1) *The UltraBay II*

The UltraBay II can be used to acquire a forensically sound image of IDE, SATA, SCSI, USB and Firewire using your choice of Forensic Imaging software. Furthermore, drives may be connected/ removed from the UltraBay II without having to shut down the workstation or leaving the GUI. The UltraBay II is exclusively available with Digital Intelligence FRED systems and is not available separately or from any other source. Fig 2 illustrates an UltraBay [5].

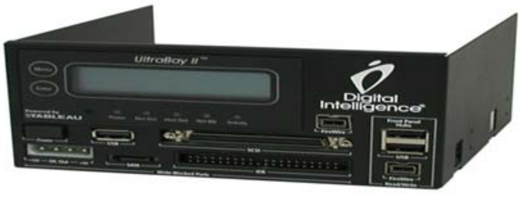


Figure 2. UltraBay II

FRED systems come with two high capacity hard drives. One of these drives is used for your forensic acquisition and processing tools and the other drive as a work drive for restoring and processing digital evidence. With multiple boot menu options FRED can be booted into data acquisition mode and PDBlock loaded automatically, write protecting the suspect hard drive.

Another boot option can be configured to place the FRED in data analysis mode with full access to your forensic analysis tools. FRED systems even come with Linux 9.1 Professional pre-configured! Both hard drives are supplied in removable trays with front panel switches for master/slave configuration [5].

FRED systems have inbuilt network functionality. All FRED systems can be connected directly to a network (10/100/1000 Mb Ethernet) for use as a standard workstation or file server when not processing or acquiring data.

The FRED Systems are usually stationary system used in the Forensics labs. There are other portable devices like the FRED – L, Ultrakit etc. FRED –L is the first laptop member of the FRED family. It as got a price tag of $4999. Though the specifications are less compared to the FRED system usually, FRED-L comes complete with an UltraKit for the ultimate mobile field forensic acquisition kit [5].

#### 2) *FRED – L*

The FRED-L forensic laptop and the included UltraKit work together to quickly, efficiently, and securely image IDE, SATA, and SCSI hard drives in a forensically sound manner. FRED-L is built on the very latest and fastest Intel Core i7-2720QM (2.2GHz, 6MB L3 Cache) Processor with up to 8 GB RAM, built-in FireWire 1394a, USB 2.0, Wireless 802.11 a/b/g/n, and Gigabit (10/100/1000 Mb/s) Ethernet support. This support is provided completely via integrated laptop components and is in no way reliant on add-on or auxiliary cards or devices.

The FRED-L has inbuilt network functionality like FRED systems. FRED-L also has the ability to connect directly to a

10Mb, 100Mb, or even Gigabit Ethernet networks for use as a standard laptop when not processing or acquiring data. FRED-L also includes integrated 802.11b/g wireless capabilities. With the addition of Network Analysis software (Packet Analyzer), FRED-L can also be used to monitor network traffic and communications at the crime scene [5]. Fig 3 illustrates the FRED-L System.

Figure 3. FRED - L

### 3) *Ultrakit III*

The UltraKit is portable kit which contains a complete family of hardware write blockers for use in acquiring a forensically sound image of virtually any hard drive you may encounter (eSATA IDE / SATA, UltraBlock SCSI, UltraBlock USB and an UltraBlock Forensic Card Reader). The UltraKit contains all the write blockers, cables, adapters, and power supplies necessary for use in acquiring images in the field using a standard laptop with FireWire or USB support. Fig 4 illustrates the Ultrakit which comes along with the FRED-L system. An Ultrakit would cost approximately $1369.

Figure 4. Ultrakit

The UltraKit consists of a Write Protected UltraBlock-IDE, UltraBlock-SATA, UltraBlock-SCSI, and a Write Enabled UltraBlock-IDE. FRED-L is designed for use "On Location" at electronic crime scenes. Remove the hard drive(s) from the suspect system and attach them to the appropriate write blocker in the UltraKit. You can then use the FRED-L system to quickly and efficiently create your image file(s) on the acquisition drive attached to the Read/Write UltraBlock. Using the Read/Write UltraBlock device allows you to utilize faster, larger, less costly desktop drives to receive your forensic images. No more worrying about the problems encountered trying to configure parallel devices on suspect equipment in order to use external backup devices. No worries about installing a SCSI adapter into a suspect’s computer [5].

With multiple boot menu options, FRED-L is not limited to use as a Forensic Imaging tool. FRED-L can be booted into DOS 6.22, Windows 98 (Standalone DOS), or Windows XP and Windows 7 and will support any forensic tools which run within those environments. FRED-L also comes complete with a fully configured installation of Suse Linux 9.1 Professional. Capable of configuration with the fastest Intel Centrino Pentium-M mobile processors (2 GHz and beyond), and with an impressive memory capacity (up to 2 GB), FRED-L is also a very formidable processing platform.

There are many more hardware devices that are used for investigation purposes like UltraBlock Forensics card reader, Image MASSter Solo, FastBloc, Acard, etc... , Each Hardware device as its own functionalities and depending on the investigation scenario the hardware’s is used. Hardware required for computer forensics include workstations and blockers such as write blockers needed to prevent contamination of evidence [6].

## IV. SOFTWARE TOOLS

The Software Tools used in Computer Forensics is usually based on the type of investigation that is carried out. If it is a data recovery investigation then Data Recovery Tools are used, each software tools as its own purpose and its own results. Computer forensics software tools would be characterised into Data Recovery Tools, Partition Tools, Disk Clone Tools, Recovery Tools, Testing Tools, RAM Test utility, System Speed Test, Hard Disk Tools, System Information Tools, Dos Tools and Other Tools. Each Tool as a variety of software’s below is few examples of and some description on them. The tools are listed below according to the category of the job.

### A. *Stealth™ Suite*

The Stealth™ Suite is used to assess activity on a computer hard disk drive without the user needing a forensic background. This set of tools helps identify whether or not a targeted computer system was used to access inappropriate information [7] [8].

### B. *Computer Incident Response Suite*

These suites of tools are often used in corporate and government investigations and security risk reviews. This suite is optimized for the lowest cost forensic platform for DOS and Windows processing, DOS. Many of the tools also have version that can be run on a Windows OS. This should be one of your first forensic toolsets. It also makes an excellent set of tools to cross-validate your findings before you go before the court or the board [7] [8].

### C. *Data Elimination Suite™*

This Suite allows you to remove information from a drive and cross-validate that the information has been removed.

This is our most popular suite of software tools for the high assurance government or corporate environment. This suite of tools has been tested and certified by the US Department of

Defence. It eliminates classified data 'leakage' and verifies that the data was properly eliminated [7] [8].

### D. *TextSearch Suite*

TextSearch NT and TextSearch Plus have both been upgraded. TextSearch NT is used to process Windows NT/2000/XP-based computer systems from a DOS command line. The upgraded program provides the same popular interface and features as TextSearch Plus but it identifies many compressed and graphics files using the file header signature, giving the investigator a listing of files that could store information in a compressed or graphic format.

Also included in this suite is HexSearch. This tool provides a similar interface as TextSearch Plus while allowing the user to search for hexadecimal strings, such as file headers, non-printing characters, and more [7] [8].

### E. *NTI Secure ToolKit*

This software is used to secure sensitive files stored on portable and desktop computers. Because it uses NIST tested and approved AES 256 encryption, it qualifies for government use with classified 'Secret' level data. This software exceeds commercial security requirements and it is much easier to use than PGP. It includes a management tool so that corporate information is not lost to the corporation. An export license may be required for locations outside the United States [7] [8].

### F. *SafeBack 3.0*

The industry standard for making evidence grade bit-stream backups of hard drives has gotten even better with version 3.0 [7] [8].

### G. *Guidance Software Encase*

Most of the software’s are packed in suites. EnCase Forensic has become the industry standard tool for uncovering, analyzing and presenting forensic data. Used by investigators in law enforcement, government, small businesses, consulting firms and corporations, EnCase Forensic provides a robust way to authenticate, search and recover computer evidence rapidly and thoroughly.

Computer evidence recovered with EnCase has been admitted into thousands of court proceedings in several countries and jurisdictions, and the EnCase software has been validated by the courts in several published decisions. [CGFI] The following are the advanced features of EnCase:

- Extracts messages from Microsoft PST files.
- Spans multiples Redundant Array of Inexpensive Disk (RAID) volumes.
- Supports NTFS compression and Access Control List (ACL) of files.
- Provides advanced language support.

Several software vendors have recently introduced computing investigation tools that work in Windows. The command line DOS tools you explored in the previous section require a strong understanding of MS-DOS and the various file systems. Because GUI (Graphical User Interface) forensics tools do not require the same level of knowledge, they can simplify computer forensics investigations. These GUI tools have also simplified training for beginning examiners in computer forensics. However DOS forensics should also be known because there are rare cases when the GUI tool would miss out critical evidence and this could be got using a DOS tool [7][8][9].

## V. Conclusion

This paper focuses on the most essential and widely used Hardware tools, there are many more tools used but the main focus was on mainly the ones that are quite common. Also, discussed on the Software tools used but not in depth, as the paper focuses on mainly the ones that are quite common in a Computer Forensics Investigation as there are other softwares as well but due to the scope of this paper it was inadequate to cover all the software tools, given a brief description of a few software’s that are used for evidence collection.

The paper attempts to give the audience to increase the level of understanding with the wide range of tools used for Computer Forensics Investigations.

## Acknowledgment

Special thanks to all other faculty members for helping during the research and giving their valuable suggestion and guidance.

## References

[1] Hasan, R.; Raghav, A.; Mahmood, S.; Hasan, M.A.; , "Artificial Intelligence Based Model for Incident Response," *Information Management, Innovation Management and Industrial Engineering (ICIII), 2011 International Conference on* , vol.3, no., pp.91-93, 26-27 Nov.2011.
doi:10.1109/ICIII.2011.307
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6114714&isnumber=6114683

[2] J. Ashcroft, “*Electronic Crime Scene Investigation: a Guide for LawEnforcement*”, 2001.

[3] J. Rory, “Practical Handbook For Private Investigators”, ISBN 0-8493-0290-0 Timothy E. Wright, 2000, Field Guide, 2001.

[4] J. Nerlinger, Jr., Jatero Consulting & Development, URL:http://www.jatero.com, retrieved 23 December 2011.

[5] Digital Intelligence, URL:http://www.digitalintelligence.com, retrieved 23 January 2012.

[6] Bill Nelson, Amelia Phillips, Frank Enfinger, Chris Steuart, *Computer Forensics and Investigation*, Cengage Learning, 2010 ISBN: 1435498836, 9781435498839.

[7] J. Wiles, K. Cardwell, A. Reyes,*"The Best Damn Cybercrime and Digital Forensics Book Period"*, 1st Edition, Syngress, 26 Nov 2007, ISBN: 9781597492287.

[8] D. Shinder; M. Cross;, "*Scene of the Cybercrime*", Syngress, 2 edition, June 20, 2008, ISBN: 978-1597492768.

[9] A. Phillips; B. Nelson; F. Enfinger; C. Steuart, "*Guide to Computer Forensics and Investigations*”, Course Technology, 2nd edition, September 28, 2009 , ISBN:978-0619217068.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Adaptive Mutation Based Particle Swarm Optimization Algorithm

Jiyang Dai, Jin Ying
Nondestructive Test Key Laboratory of Ministry Education
Nanchang Hangkong University
Nanchang, China
E-mail: djiyang@163.com, 783026713@qq.com

***Abstract*—In this paper an adaptive mutation based PSO (AMBPSO) is presented for improvement of deficiencies of standard PSO， which is modified by the combination of dynamic adjustment of the inertia weights, the update of position and velocity of each particle by means of randomly adaptive mutation, and the limit of the update for the change in a reasonable range. The optimization results of two standard test functions show that these modifications can enhance particles' activity to improve the algorithm's search precision and convergence speed and to keep away from easily immerging in local minima efficiently compared with standard PSO and general PSO.**

***Keywords-particle swarm optimization*; *swarm intelligent optimization*; *adaptive mutation*; *modified algorithm***

## I. INTRODUCTION

Particle Swarm Optimization (PSO) is a technique modeling swarm intelligence that was developed by Kennedy and Eberhart [1] in 1995, who were inspired by the natural flocking and swarming behavior of birds. Compared with other swarm intelligent optimization algorithms such as Genetic Algorithms [2], it has more intuitive simplicity, less parameter settings and higher efficiency. It is a recently invented high performance optimizer that possesses several highly desirable attributes and especially that is an effective global optimization search algorithm for dealing with the complex nonlinear optimization problem with difficulty to solve by the traditional optimization methods [3-4].

Many advances in PSO development elevated its capabilities to handle a wide class of complex engineering and science optimization problems. Summaries of recent advances in these areas are presented in [5] and [6]. The existed PSO algorithms mainly include PSO with inertia weight, PSO with contraction factor, the binary PSO, the niche PSO and hybrid PSO, etc. Different variants of the PSO algorithm were proposed but the most standard one is the global version of PSO (Gbest model) introduced by Shi and Eberhart [7], where the whole population is considered as a single neighborhood throughout the optimization process. Although a key attractive feature of the PSO approach is its simplicity as it involves only two model equations, it has some disadvantages, such as less sensitive to environmental variations, and easily immerging in local minima frequently [8, 9]. Many improved techniques effectively increased the PSO's local disadvantages to a certain extent, but these problems need to be further resolved [10-14].

In this paper an adaptive mutation-based PSO (AMBPSO) is presented for improvement of the above deficiencies of standard PSO which is modified by the combination of dynamic adjustment of the inertia weights, update of position and velocity of each particle by means of randomly adaptive mutation, and the limit of the update for the change in a reasonable range. These modifications can enhance particles' activity to improve the algorithm's search precision and convergence speed and to keep away from easily immerging in local minima efficiently.

The paper is structured as follows. Section 2 analyses the standard PSO algorithm. Section 3 details an adaptive mutation based PSO algorithm. The optimization calculation and analysis of two standard functions using AMBPSO algorithm are given in Section 4. Conclusions are given in Section 5.

## II. STANDARD PARTICLE SWARM OPTIMIZATION ALGORITHM

### *A. Basic Concept of PSO*

In a standard PSO, a great number of particles move around in a multidimensional space, with each particle memorizing its position vector and velocity vector as well as the time at which it reached its highest level of fitness. The inspiration underlying the development of this algorithm was the social behavior of animals, such as the flocking of birds and the schooling of fish, and the swarm theory. The advantages of PSO are that it involves no evolution operators, such as crossover and mutation operators, and it does not require the adjustment of too many free parameters. PSO begins with a random population and searches for optima by continually updating this population. Moreover, each potential solution is assigned a randomized velocity. The potential solutions, called particles, are the "flown" through the problem space. Related particles can share data at the best-fitness time. The velocity of each particle is updated according to the best positions reached by all particles through iterations, and the best positions are determined by the related particles over the course of multiple generations.

It is Similar to other intelligent optimization strategies that there are individuals called as "particles", all of which

This work was supported in part by the National Natural Science Foundation of China (6116405) and the Aeronautical Science Foundation of China (2011ZA56021)

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

constitute a solution group in particle swarm optimization (PSO) algorithm. As they are flown through the searching space, particles are attracted towards the best solution found by the given particle's neighbors and by the particle itself. Each particle has a position vector and a velocity vector. Suppose that the searching space is $N$-dimensional, and $M$ particles form the colony. The $i$th particle has an N-dimensional position vector $\boldsymbol{X}_i$ ($i$= 1, 2, ⋯, $M$), which means that the $i$th particle is located at $\boldsymbol{X}_i$= $(x_{i1}, x_{i2}, \cdots, x_{iN})^{\mathrm{T}}$ in the searching space. The position of each particle is a potential result. The $i$th particle's "flying" velocity is also an N-dimensional vector, represented by $V_i = (v_{i1}, v_{i2}, \cdots, v_{iN})^{\mathrm{T}}$ ($i$= 1, 2, ⋯, $M$). The best particle among all particles can be represented by $x_d^{gbest}$ ($d$=1, 2, ⋯, $N$) whose fitness is *gbest*. Each particle also keeps track of the value of its best position, which is represented by $x_{id}^{Pbest}$ ($d$=1, 2, ⋯, $N$) corresponding to fitness $Pbset_i$.

In the standard PSO algorithm, the velocity and position of each particle are adjusted as follows:

$$v_{id}^{k+1} = wv_{id}^{k} + c_1 r_1 (x_{id}^{Pbest} - x_{id}^{k}) + c_2 r_2 (x_{d}^{gbest} - x_{id}^{k}) \qquad (1)$$

$$x_{id}^{k+1} = x_{id}^{k} + v_{id}^{k+1} \qquad (2)$$

where $r_1$, $r_2$ ∈(0, 1) are uniformly distributed random numbers between 0 and 1; $c_1$ and $c_2$ are learning rates; and $w$ denotes the inertia weight. $k$ ($k$=1, 2, ⋯, $D$) is the current iteration, and $D$ is the maximum number of iterations.

PSO randomly initializes the flock of birds throughout the searching space; every bird is called a "particle". During each individual generation, each particle adjusts its velocity vector, based on its best solution $x_{id}^{Pbest}$ and that of its neighbors $x_d^{gbest}$, as calculated by means of Eq.(1) and (2). The second part $c_1 r_1 (x_{id}^{Pbest} - x_{id}^{k})$ of Eq. (1) refers to the "cognitive component", which reflects the distance at which a particle is located from the best solution. The particle determines this distance on its own. On the other hand, the last part $c_2 r_2 (x_d^{gbest} - x_{id}^{k})$ of Eq.(1) refers to the "social component", which reflects the distance between a particle and the best solution found by its neighbors. Eq. (1) is used to calculate the particle's new velocity on the basis of its previous velocity, the distance between its current position and its own best previous position, and the collaborative interaction between the particles. This stage features cooperation among all particles in sharing information. Finally, the particle updates its position by means of Eq. (2). The best global position, determined by taking the condition of all particles into consideration, is defined as the particle with the minimum fitness.

The role of inertia weight $w$ is to make a compromise between local optima and global optima. Larger $w$ is beneficial to a wide range of search and smaller w to precise search in a small area. In general, $w$ is set as a function with linearly decreasing in the process of iterations.

$$w = w_{\max} - (w_{\max} - w_{\min})k / D \qquad (3)$$

where $w_{max}$ and $w_{min}$ are the initial and ultimate inertia weights, respectively.

### *B. Parameter Analysis*

The PSO is very simple to operate that involves fewer parameters. The main parameters need to be set as follows: Target Dimensional $N$, Population Size $m$, Inertia Weight $w$, Velocity Coefficient $c_1$, $c_2$, Max Iteration $D$. The parameters of the algorithm and meaning in the general principles of the selection are summarized below [2]:

- The target dimensional $N$, the number of parameters in the PSO need to be optimized, according to the definite corresponding target search space dimension.
- The population size m is generally selected from 10 to 40 suitable for solving most of optimization problems. But according to specific or very complex optimization problem, the population size can be taken between 100 and 200.
- Inertia weight w is generally selected between 0.35 and 0.9 in the practical optimization problem, where the initial and final inertia weights are usually selected as 0.9 and 0.35, respectively.
- Acceleration coefficient $c_1$ and $c_2$ are usually selected between 0 and 4, which are the main parameters for adjustment of particles "Cognitive item (Cognitive Term)" and "Social items (Social Term)". If $c_1$ = 0, then particles do not have cognitive abilities. This moment the particles have the ability to interact and can be expanded to new search space, but the particles in the iteration process are easy to fall into the local minimum; If $c_2$ = 0, the particles do not have the social information sharing capability. This time the PSO will develop random search; If $c_1 = c_2 = 0$, the particles will keep the initial velocity until they reach maximum iterating times to end;
- Maximum iteration $D$ as the threshold value control evolution of the PSO, can reflect the evolution of the algorithm efficiency, and too much iteration will cause a tremendous waste of the algorithm, or too small iteration may not be able to reach the goal of the optimization.

### *C. Implementation Steps*

After determining target search dimension $N$, population size $M$, the maximum number of iterations $D$ and the learning rates $c_1$ and $c_2$, the implemented steps of the PSO are as follows:

1) Initialize $v_{id}^{0}$ and $x_{id}^{0}$ .

2) Calculate the fitness of each particle, and store the best fitness of the individual (*pbest*) and its corresponding position $x_{id}^{Pbest}$, the best fitness of group (*gbest*) and its corresponding position $x_d^{gbest}$, the average value $x_{id}^{mbest}$ of positions of all the best particles relative to their own.

3) Update velocities $v_{id}^{k+1}$ and positions $x_{id}^{k+1}$ according Eq. (1) and (2).

4) Repeat steps 2–3 until a termination criterion is satisfied.

## III. Adaptive Mutation Based PSO

### A. Ideas of Improvement for PSO

Although the PSO has many advantages such as simple operation, fewer parameters, fast convergence speed, but it also has many disadvantages to be overcome, such as:

- In spite of the fact that particles can fly towards the optimal solution in search space according to all members of the group and their own experiences, the improper choice of inertia weight will lead to the lack of precise search ability of particles;
- In the process of particles' searching towards the goal of optimal solution, when a particle is close to the optimal value, its search speed will be small. As a result, the search diversity will be lost and a particle is easy to fall into the local minimum.

In order to overcome the above deficiencies, an adaptive mutation-based PSO (AMBPSO) algorithm is proposed for the improvement of the search accuracy and convergency of the standard PSO algorithm.

Firstly, the inertia weight is modified as a function with gradient descent:

$$w = (w_{\max} - w_{\min}) / k^H + w_{\min} \tag{4}$$

where $H$ is a given positive number. If $k$=1, then $w$=$w_{max}$ and if $k$→∞, then $w$=$w_{min}$.

Secondly, we sort fitness values of all particles in each iteration. According to the ranking, the particles in the top 50% are preserved while the rest in the colony is modified by means of adaptive mutation.

$$v_{bd}^k = v_{gd}^k (1 + \beta r) \tag{5}$$

$$x_{bd}^k = x_{gd}^k (1 + \alpha r) \tag{6}$$

where,

$x_{gd}^k$ and $v_{gd}^k$ are position and velocity of the particles in the top 50% of fitness ranking respectively.

$x_{bd}^k$ and $v_{bd}^k$ are position and velocity of the particles in the rest of 50% worse particles in the colony respectively, which will be modified by adaptive mutation.

$r$ is a uniformly distributed Gaussian random number with a range of [0, 1].

$\alpha$ and $\beta$ are two given positive numbers.

Meanwhile, each particle's position and velocity in AMBPSO algorithm are limited into the given boundaries. If the positions or velocities are out of the given boundaries, they will be replaced by the given boundaries.

Finally, a mutation factor $P$ is introduced as follows:

$$P = x_{id}^{mbest} - x_{d}^{gbest} \tag{7}$$

where $x_{id}^{mbest}$ is a mean value of the positions of the best particles obtained in the $i$th iteration, and $x_d^{gbest}$ is the global position value with the particles of the best fitness obtained so far.

The inertia weight coefficient can be dynamically adjusted on the basis of the following equation:

$$H = H(1 + \gamma r) \tag{8}$$

where $\gamma$ is a positive number and $r$ is a uniformly distributed Gaussian random number with a range of [0, 1].

It is necessary to select the appropriate mutation thresholds $\rho$ according to the actual situation. If $P$ is less than or equal to $\rho$, then $H$ is modified by Eq.(8). The aim on the introduction of mutations factor $P$ is to improve the mutant ability of AMBPSO. Because $x_{id}^{mbest}$ can comprehensively reflect the activities of the individual particles and $x_d^{gbest}$ can refmutationall activities of all particles, the mutations factor $P$ is adopted not only to effectively suppress the AMBPSO premature to early fall into local minima, but also to efficiently ensure the AMBPSO to jump out of the local minima. As a result, it is critical to select an appropriate mutation threshold ρ. A large number of experiments show that, the proposed AMBPSO algorithm can effectively make up for deficiencies on convergence and premature of the standard PSO to improve the search performances of PSO.

### B. Implementation Steps of the AMBPSO

In summary, The AMBPSO algorithm can be described in general as follows:

1) Determine target search dimension $N$, population size $M$, the maximum number of iterations $D$ and the learning rates $c_1$ and $c_2$.

2) Initialize $v_{id}^0$ and $x_{id}^0$ .

3) Calculate the fitness of each particle, and store the best fitness of the individual (*pbest*) and its corresponding position $x_{id}^{Pbest}$, the best fitness of group (*gbest*) and its corresponding position $x_d^{gbest}$, the average value $x_{id}^{mbest}$ of positions of all the best particles relative to their own.

4) Adaptively mutate 50% worse particles $v_{bd}^k$ and $x_{bd}^k$ according to Eq. (5) and (6).

5) Judge mutation factor $P$ according to Eq. (7), dynamically adjust the inertia weight coefficient $H$ according to Eq. (8) and finally calculate the inertia weight $w$ in terms of Eq. (4).

6) Update velocities $v_{id}^{k+1}$ and positions $x_{id}^{k+1}$ according Eq. (1) and (2).

7) Repeat steps 3–6 until a termination criterion is satisfied.

## IV. Calculation and Analysis of Standard Test Functions

In order to validate the performances of the improved particle swarm optimization algorithm, the AMBPSO, the Standard PSO (SPSO) and the PSO are used to test two types of standard functions. The comparison is made from three aspects of the convergence rate, the optimal fitness value and the search precision of the three PSO algorithms. The results on the above three aspects of algorithm performances for standard test functions can indicate which algorithms are much better. Different types of standard test functions are utilized for different types of optimization problems [1].

Two types of standard test functions are used to validate the three PSO algorithms, which are defined as follows:

1) The first test function is chosen as Generalized Rastrigin function expressed as

$$f(x)=\sum_{i=1}^{n}[x_i^2-10\cos(2\pi x_i)+10]$$

Search range: $-5.12 \le x_i \le 5.12$

Global optimal value: $\min(f)=f(0,\cdots,0)=0$

2) The second test function is chosen as Ackley function expressed as

$$f(x)=-20\,\mathrm{e}^{(-0.2\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2})}-\mathrm{e}^{(\frac{1}{n}\sum_{i=1}^{n}\cos 2\pi x_i)}+20+e$$

Search range: $-32 \le x_i \le 32$

Global optimal value: $\min(f)=f(0,\cdots,0)=0$

There are a great number of local minima contained in the selected test functions due to the effect of the cosine functions in the above standard test functions. If the selected intelligence group optimization algorithms have an inadequate activity for departing from local minimum values, the global optimal values cannot be achieved. These types of test functions are used to better validate the performance that the AMBPSO can make up for deficiency that both of the PSO and the SPSO are easy to get into local minima.

In order to show convincingly that the proposed novel PSO algorithm is more superior to other PSO algorithms by comparing AMBPSO with SPSO and PSO, the selection of control parameters is based on a consistency principle, that is, the basic control parameters of the three optimization algorithms are set as the same in this paper. The parameters are set as follows:

Population size $m$=40, Learning rates $c_1$=$c_2$=2.05, Initial inertia weight $w_{max}$=0.9, Terminal inertia weight $w_{min}$=0.35, Maximal velocity $v_{max}$=1, Minimal velocity $v_{min}$= -1, Maximal position $x_{max}$=10, Minimal position $x_{min}$=0.05, Inertia weight coefficient $H$=0.5096, Mutation Threshold $\rho$=0.0001, Mutation factor of inertia weight coefficient mutation $\gamma$=0.02, Mutation factor of position $\alpha$=0.0005, Mutation factor of velocity $\beta$=0.0005, Maximal iteration numbers *maxgen*=100/500/1000.

Initial speeds and positions in the improved particle swarm optimization algorithm are generated randomly in terms of actual optimization problems. At the same time, the numerical experiments are done on three different maximum iteration numbers. The results are illustrated by only using the fitness graphs at *maxgen*=100.

The curves of the best fitness variations with iterations generated by three PSO algorithms as to the two test functions are shown respectively in Fig.1 and Fig.2. The dimensions of the two selected functions are both set as $n$=20. In order to avoid contingency, 50 times of optimization experiments are done for all the algorithms to analyze and compare the performances of the three PSO algorithms. From Fig. 1 and 2, it can be obviously shown that The AMBPSO is of higher convergence speed than the other two algorithms for both of the selected test functions during the process of iterations.

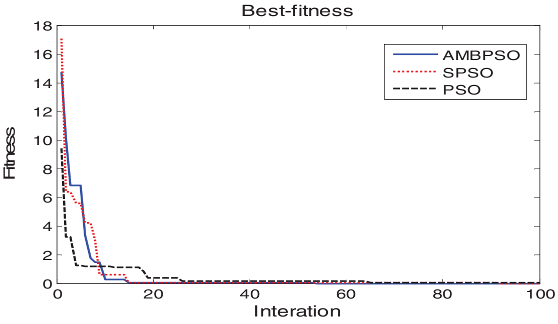


Fig. 1 Fitness curves of Generalized Rastrigin function

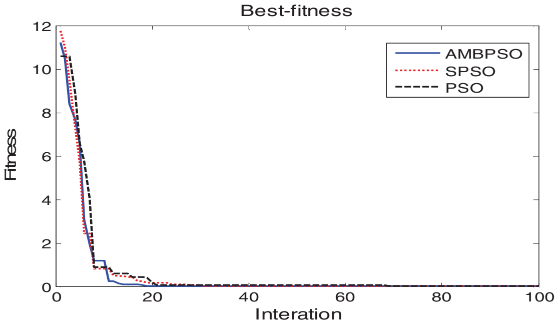


Fig. 2 Fitness curves of Ackley function

As shown in Table 1 are root mean square errors between real optimal values and calculated values by the three PSO algorithms as to the two given test functions at three different terminal iterations.

TABLE I. RMS ERRORS OF OPTIMAL VALUES FOR EACH TEST FUNCTION

| | **Generalized Rastrigin** | | | **Ackley** | | |
|---|---|---|---|---|---|---|
| | AMBPSO | SPSO | PSO | AMBSPO | SPSO | PSO |
| 100 | $3.07\times10^{-4}$ | $4.20\times10^{-3}$ | $3.86\times10^{-2}$ | $6.22\times10^{-6}$ | $2.80\times10^{-3}$ | $1.92\times10^{-2}$ |
| 500 | $7.83\times10^{-9}$ | $1.14\times10^{-4}$ | $6.60\times10^{-3}$ | $8.64\times10^{-10}$ | $7.34\times10^{-5}$ | $4.30\times10^{-3}$ |
| 1000 | $8.88\times10^{-13}$ | $4.44\times10^{-7}$ | $1.21\times10^{-4}$ | $3.14\times10^{-13}$ | $1.52\times10^{-9}$ | $9.24\times10^{-5}$ |

From Table 1, it can be shown that the search precision increases with the increment of the maximal number of iterations given in advance, and that under the same terminal condition the AMBPSO algorithm is of much higher precision and faster convergency rate than the other two PSO algorithms for whichever standard test functions.

## V. CONCLUSIONS

In this paper an adaptive mutation-based PSO (AMBPSO) is presented for improvement of the deficiencies of standard PSO which is modified by the combination of dynamic adjustment of the inertia weights, update of position and velocity of each particle by means of randomly adaptive mutation, and the limit of the update for the change in a reasonable range. The optimization results of two standard test functions show that these modifications can enhance particles' activity to improve the algorithm's search precision and convergence speed and to keep away from easily immerging in local minima efficiently.

## REFERENCES

[1] J. Kennedy and R. Eberhart, Particle swarm optimization, in Proc. IEEE Int. Conf. Neural Network, 1995, vol. 4, pp. 1942-1948.

[2] Zhen JI, Huilian LIAO, Q. H. WU. Paritcle Swarm Optimization algorithm and application. Beijing: Science, 2009.

[3] Ying Song, Zengqiang Chen, Zhuzhi Yuan et al. New Chaotic PSO-Based Neural Network Predictive Control for Nonlinear Process. IEEE Transactions on Neural Networks, 2007, vol. 8, no. 2, pp. 595-601.

[4] Guo-Dong Li, Shiro Daisuke Yamaguchi et al. An Optimal Grey PID Control System. IEEJ Transactions on Electrical and Electronic Engineering, 2009, vol. 4, no. 4, pp. 570-577.

[5] H. Xiaohui, S. Yuhui, and R. Eberhart, Recent advances in particle swarm, Proc. Congr. Evol. Comput., 2004, vol. 1, pp. 90–97.

[6] R. C. Eberhart and Y. Shi, Guest editorial, IEEE Trans. Evol. Comput. (Special Issue on Particle Swarm Optimization), 2004, vol. 8, no. 3, pp. 201-203.

[7] Y. Shi and R. Eberhart, A modified particle swarm optimizer, Proc. IEEE World Congr. Comput. Intell., 1998, pp. 69-73.

[8] M. R. AlRashidi, M. E. El-Hawary, A survey of particle swarm optimization applications in electric power systems, IEEE Transaction on Evolutionary Computation. 2009, vol. 13, no. 4, pp. 913-918.

[9] Senthil Arumugam M, Rao MVC, Chandramohan A. A new and improved version of particle swarm optimization algorithm with global-local best parameters. Knowledge and Information Systems, 2008, vol. 16, no. 3, pp. 15-26.

[10] H. M. Soliman, M. A. Awadallah, M. Nadim Emira et al. Robust Controller Design For Active Suspensions Using Particle Swarm Optimization[J]. International Joual of Modeling, Identification and Control, 2008, vol. 5, no. 1, pp. 66-76.

[11] Zhen JI, Huilian LIAO, Yiwei WANG, Q. H. WU. A Novel Intelligent Particle Optimizer for Global Optimization of Multimodal Functions. IEEE Congress on Evolutionary Computation(CEC 2007), 2007, pp. 3272-3275.

[12] M. Senthil Arumugam, M.V.C.Rao. Novel hybrid approach for Real Coded Genetic Algorithm to Compute the Optimal Control of a single stage Hybrid Manufacturing Systems, International Journal of Computational Intelligence, 2005, vol 1, no. 3, pp. 231- 249.

[13] Omkar SN, Mudigere D, Narayana Naik G, et al. Vector evaluated particle swarm optimization for multiobjective design optimization of composite structures. Computers and Structures, 2008, vol. 86, no. 12, pp. 45-56.

[14] Shi B, Li YX, Yu XH, et al. A modified particle swarm optimization and radial basis function neural network hybrid algorithm model and its application. WRI Global Congress on Intelligent Systems(GCIS 2009), 2009, vol. 1, pp. 134-138.

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# Control for High Heat Chips' Cooling Based on Power Consumption and Temperature Signals

Jian Wang
School of Electronics and Information Engeering
University of Science and Technology of Suzhou, USTS
Suzhou, China
wangjiansuzhou@sina.com

Chuan-yang Liu
School of Electronics and Information Engeering
University of Science and Technology of Suzhou, USTS
Suzhou, China

*Abstract*—**For CPU chips' cooling task, existing mono-variable control system has the defects of delay, fluctuation and high energy consumption. With the trend of multi-core and high frequency work pulse in CPU chips, traditional system does not fit for future cooling task. Several approaches such as chips' dissipated power nonlinear feed-forward, core temperature tracking and surface temperature feed-back are applied to improve control system. Simulations prove that comparing with the traditional out-core temperature based mono-input-output proportion tracking strategy, the improved system can bring better quality and energy-saving results. Also, the new system has parameters adaptability.**

***Keywords-CPU chip; temperature; cooling; control; power dissipation; usage rate; feed-forward***

## I. INTRODUCTION

In future space mission, more and more high heat integrated circuit chips need to be cooled [1]. Promoting the quality of high heat IC chips' cooling control is just a new study focus. CPU chips are a typical kind of high heat chips and their cooling control is the key of guaranteeing their stability and dynamic properties [2]. Traditional cooling control systems are mono-variable tracking system, and they apply proportional algorithm as control strategy [3-5]. That is, the discrete values of out-core temperature in chip are multiplied by a coefficient and then are used to adjust cooling actuator's driving power.

CPU chip is a body with inside heat sources. Its heat generating mechanism [6-9] makes temperature in chip is a nonstationary random process with delay and many influence factors. This is the bottle-neck for traditional control system to fit for future CPU chips' cooling task. Concretely speaking, uncertain deviation between out-core temperature measuring values and real temperature [7], single system input variable and simple tracking control strategy, and so on, are main reasons. So, it is urgent to find new control system to cope with future challenges. Unluckily, trends of more cores and higher frequency of main pulse of CPU chip, future scheduling strategy of operating system, development of CPU internal structures and techniques, even increasing of cooling actuator's efficiency, will make CPU chip's power, heat and temperature more complex than expected.

To promote the transient and steady state quality of chip's surface temperature, this paper suggests a new system which combines several approaches such as new signals and their processing, more control algorithms and more complicated strategies. In the new system, CPU usage rate which indicates the chip's dissipated power in advance [9-11], and CPU core temperature which is more accurate and more sensitive and no delay, are used as inputs. And, power feed-forward, core temperature tracking and surface temperature feed-back strategies are added. Numerical simulation proved our new method is effective and advanced.

## II. CHIPS' COOLING AND TRADITIONAL CONTROL SYSTEMS

Traditional cooling control system (seeing Fig. 1) uses analogue variables as inputs. It applies proportional strategy in Fig. 2 to determine the excitation voltage $V_{out}$ of the cooling actuator, usually one or two air fan, as follows:

$$V_{out} = k_p T(t) \tag{1}$$

It is obvious that the object itself is not complex. That is, the chip's heat transfer delay and the air fan's cooling time are

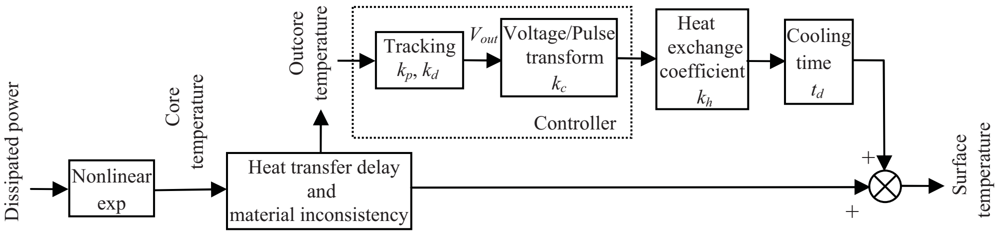


Figure 1. CPU chips' cooling model of tracking control system.

common properties in control systems. But there is peculiar complexity resulting from the nonstationary random and anisotropic inconsistency in temperature field. As being seen in Fig. 3, not only is there nonstationary random in out-core temperature and core temperature, but at least 1 second of time delay and at least 1°C of temperature variance in them as well. Besides, the uncertainty of heat resulted from CPU workloads and the variant cooling time influenced by heat exchanging efficiency of actuator, etc., bring the chips' cooling with difficulties.

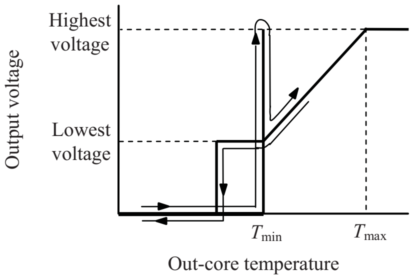


Figure 2. Traditional proportional tracking strategy.

Generally speaking, the traditional proportional tracking strategy cannot further improve the cooling control quality, and cannot face future challenges.

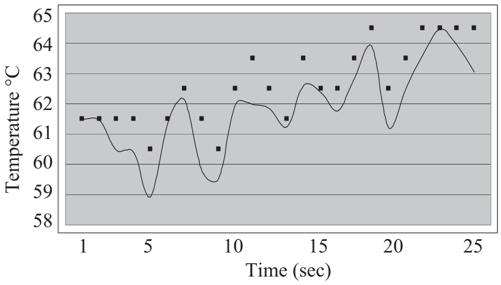


Figure 3. CPU core temp. & out-core temp.

## III. Model of Control System with Dissipated Power Feed-Forward and Multi-Temperature PID

### A. Improved Model of Controller

We noticed that chip's power consumption is in advance to its temperature. On the basis of existing tracking system, we introduce usage feed-forward to reflect the action of CPU's power consumption to temperature. Besides, we substitute out-core temperature with core temperature as input variable, and add algorithms of tracking and feed-back. Certainly, we take short window mean value from sequences of core temperature and usage. Fig. 4 is the improved control system model. In Eq. (2) are several algorithms in the new model such as usage sequence nonlinear feed-forward, core temperature tracking, surface temperature feed-back, etc. In that equation, $k_u$ is usage feed-forward coefficient and the linear item, $k_r$, is added to make up the attenuation with usage increasing. The $k_p$, $k_i$ and $k_d$ are proportional, integral and deviational coefficients of PID algorithm, respectively. The $k_f$ is coefficient of different value feed-back.

$$\begin{aligned}
&k_u \left\{1 - \exp\left(- \mathrm{E}\left[U_n^{L,l}\right]\right) + k_r \mathrm{E}\left[U_n^{L,l}\right]\right\} \\
&k_p\, \mathrm{E}\left[T_n^{L,l}\right] + k_d \frac{\left\{\mathrm{E}\left[T_n^{L,l}\right] - \mathrm{E}\left[T_{n-1}^{L,l}\right]\right\}}{l} \\
&k_i \sum_{i=1}^{n} \left\{T_o(n) - \frac{T_{\max} + T_{\min}}{2}\right\} \\
&k_f \left\{T_o(n-1) - \mathrm{E}\left[T_n^{L,l}\right]\right\}
\end{aligned} \tag{2}$$

In Eq. (2), $\mathrm{E}\left[U_n^{L,l}\right]$ and $E\left[T_n^{L,l}\right]$ are short window mean value of usage sequence and core temperature sequence, respectively.

### B. Nonlinear Relation between Usage Rate and Core Temperature

The CPU usage measures microprocessor's workloads. It reflects the power consumption of chip directly and determines the amount of heat generated in the chip [7, 10-11]. For a CPU chip, suppose it gains heat $Q_{gained}$ from its environment, $Q_{transfer}$ is the amount of heat transferred to it during a period of time and $Q_{component}$ is the amount of heat produced by running programs during a period of time. This can be expressed as [10]:

$$\begin{aligned}
&Q_{gained} = Q_{transfer} + Q_{component} = \\
&k\left(T_{out} - T_{source}\right)t + \left[P_{base} + \left(P_{\max} - P_{base}\right)u\right]t
\end{aligned} \tag{3}$$

where $T_{source}$ is core temperature and $T_{out}$ is air temperature, $P_{base}$ is the power consumption when CPU is idle and $P_{\max}$ is the consumption when chip is fully utilized, $u$ is usage rate.

Since $\Delta T = \frac{1}{mc}\Delta Q$, then

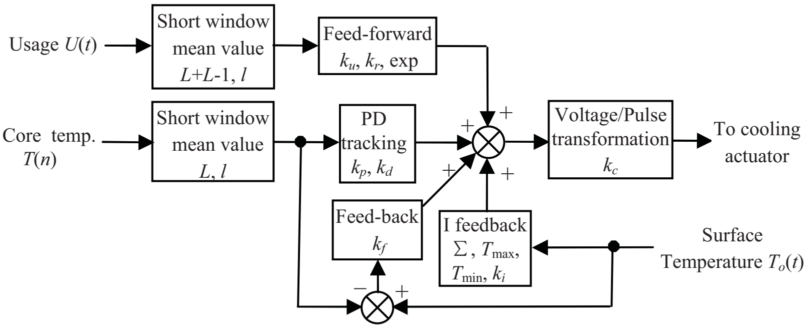

Usage $U(t)$
Short window mean value $L$+$L$-1, $l$
Feed-forward $k_u$, $k_r$, exp
Core temp. $T(n)$
Short window mean value $L$, $l$
PD tracking $k_p$, $k_d$
+
+
+
+
Voltage/Pulse transformation $k_c$
To cooling actuator
Feed-back $k_f$
I feedback $\Sigma$, $T_{\max}$, $T_{\min}$, $k_i$
Surface Temperature $T_o(t)$
−
+

Figure 4. Improved model of CPU chip cooling controller.

$$\frac{dQ_{gained}}{dt} + k\left(T_{source} - T_{out}\right) = P_{base} + \left(P_{\max} - P_{base}\right)u$$

$$mc\frac{d\left(T_{source} - T_{out}\right)}{dt} + k\left(T_{source} - T_{out}\right) = \qquad (4)$$

$$P_{base} + \left(P_{\max} - P_{base}\right)u$$

where $m$ is the mass of chip, $c$ is its heat capacity, and $k$ is related to $m$ and $c$.

In Fig. 5 is statistic result of real out-core temperature measuring values. When usage is larger than 10%, the action of dissipated power to out-core temperature can be seen easily. Through fitting, we get the relation between steady temperature values and constant usage values as Eq. (5), where $T_{base}$ is the basic temperature corresponding to zero usage, $k_u$ and $k_r$ are constants. It must be pointed out that Eq. (5) just coincides with the solution form of Eq. (4).

$$T(\infty) = T_{base} + k_u\left(1 - e^{-k_r u}\right) \qquad (5)$$

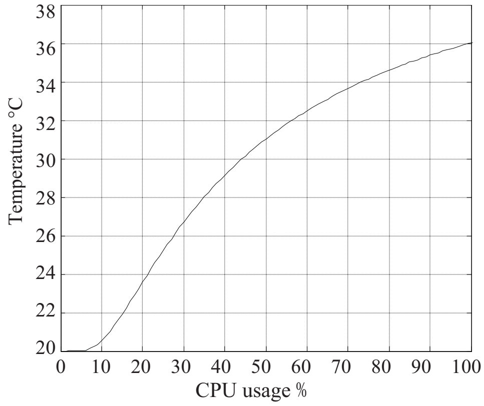


Figure 5. Steady temperature under different usage.

### C. Short Window Mean Value of Core Temperature and Usage Sequences

The discrete sequence of core temperature $\{T_n\}$ is a nonstationary random one with big sampling cycle. It can be proved that the short window mean value of the sequence $\left\{E\left[T_n^{L,l}\right]\right\}$ contains the same heat as $\{T_n\}$. And, the waveform of the former is smoother than that of the latter. This means short window mean value is more fittable for being applied to cooling control than discrete values.

**Brief proof:**

Let frame length $L$ is 2, frame width $l$ is 1, then $\left\{E\left[T_n^{L,l}\right]\right\}$ contains heat

$Q_{aver} = \sum_{i=1}^{n}\int_{i-1}^{i} T_i dt = \frac{T_1}{2} + \sum_{i=1}^{n}\int_{i}^{i+1}\frac{T_i + T_{i+1}}{2}dt + \frac{T_n}{2}$. And $\{T_n\}$ contains heat $Q_{disc} = \frac{T_1}{2} + \sum_{i=1}^{n}\int_{i}^{i+1}\frac{T_i + T_{i+1}}{2}dt$ , if $n$ is large enough, $Q_{aver} \approx Q_{disc}$. Furthermore, if $T_i$+1>$T_i$, then $T_{i+1} > \frac{T_i + T_{i+1}}{2}$, if $T_i$+1<$T_i$, then $T_{i+1} < \frac{T_i + T_{i+1}}{2}$. **Proof end**

As for usage sequence, $U(n)$, we make use of its short window mean value $\left\{E\left[U_n^{L,l}\right]\right\}$, too.

## IV. Simulation and Analysis

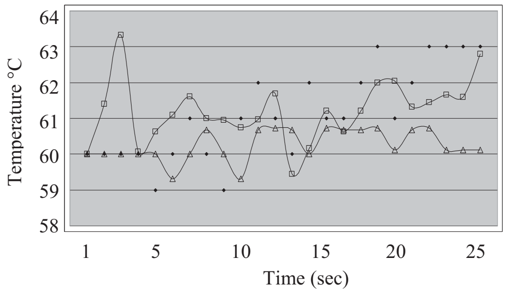
### A. Selection of Parameters

For simplicity, we suppose the power dissipated by cooling driving actuator is all used for cooling. According to experiment experiences, let cooling time, $t_d$, is 1 second, and time constant of heat transfer, $\tau$, is 0.25 second. That means all heat generated from source is taken away to environment by cooling actuator in 1 second.

All parameters' values are taken as follows: $L$=2, $l$=1, $\tau$=0.25s, $t_d$=1s, ($T_{\min}$, $T_{\max}$)=(50°C, 70°C), ($U_{\min}$, $U_{\max}$)=(8V, 12V), $k_p$=0.1, $k_i$=0.01, $k_d$=0.01, $k_f$=0.01, $k_c$=1.2, $k_h$=8, $k_u$=0.1, $k_r$=1.0.

### B. Simulation Results and Analysis

*1) Comparisons between core temperature and out-core temperature as input variable.* The simulation results of pure proportional tracking control (seeing Fig. 1) using out-core temeprature or core temeprature as inputs are showed in Fig. 6, respectively. Obviously, owing to the uncertainty and delay of out-core temeprature, its result is much worse than that of core temeprature. Simulation results indicate that the larger or smaller is $k_p$, the larger is the surface temperature's fluctuation amplitude. This is because the pure proportional tracking strategy can not eliminate the disturbance of uncertain deviation in out-core temperature values. In real application, the effect of $k_p$ is the same as that of heat exchange coefficient of cooling actuator, this means that traditional strategy does not fit for cooling actuator with lower heat exchange efficiency, and not for actuator with higher heat exchange efficiency, either. It is needed to point out that out-core temperature used in our numerical calculating has smaller random deviation. Actual random deviation is bigger and its influence is more serious.



■ ---- outcore temperature as input variable

▲ ---- core temperature as input variable

Figure 6. Control results by traditional system.

*2) Comparisons of control qualities made by different improved strategies.* For improved control system in Fig. 4, we use core temperature in Fig. 2 as input variable to do simulations. At the beginning, the controller only with PD core temperature tracking algorithm make surface temperature having 1°C steady state deviation. Then, after surface temperature feed-back being added, controller eliminates the deviation. Or, after usage nonlinear feed-forward being added, the overshoot of surface temperature is lower. Lastly, once surface temperature feed-back and usage feed-forward are added simultaniously, the result is best both in transient and steady state.(seeing Fig. 7)

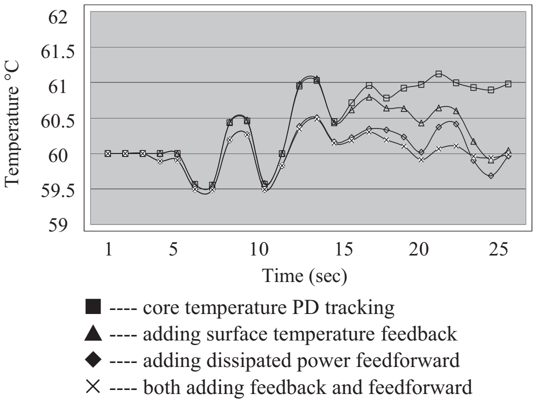


Figure 7. Control results by improved system.

*3) Comparison of power dissipated by cooling actuator.* For the combination strategy of core temperature tracking and surface temperature feed-back, the numerical calculating results of using two different values spent different energy. Control system using discrete values of core temperature makes the actuator spend 2.46053 Joule of energy, whereas system using short window mean values of core temperature makes the actuator spend 1.76581 Joule of energy. In our example, the former spends 0.69472 Joule more energy than the latter during 20 seconds.

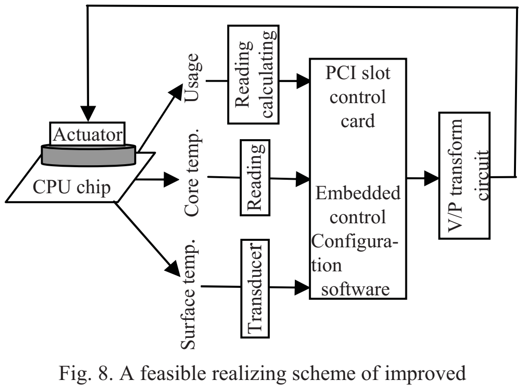


Fig. 8. A feasible realizing scheme of improved chip's cooling control system.

*4) Discussions about feasibility and practicability.* Simulating work indicates that the parameters in new system model such as $k_p$, $k_i$, $k_d$, $k_f$, $k_u$, $k_r$, $L$ and $l$ can be adjusted in wide range to cope with different values of $\tau$, $t_d$, $k_c$, $k_h$. For instance, whether the cooling actuator is air fan or liquid pump or semeconductor thermoelectric patch, i.e., whether the actuators' heat exchange efficiency is high or low, simulating shows our model can give excellent control quality. For example, under the circumstance of CPU core temperature behaving different random properties owing to different operating systems, or different main pulses, or different user softwares, new system can adapt them by using corresponding $L$ and $l$.

## V. Conclusion

This paper studied how to improve CPU chips' cooling control system and suggested several new methods in two aspects: one is of input signals and their processing; the other is of control strategies. Numerical calculating and simulating verified that usage sequence which indicates CPU power and core temperature sequence which is more accurate and more instant can avoid the disadvantages of out-core temperature values. Also, it was verified that power feed-forward and surface temperature feed-back can enhance system's functions and improve its performance. The improved control system has advantages in control quality, energy saving and adaptability. Our improved system can be easily realized by making use of existing techniques. Actually, it does be our coming work. Fig. 8 shows one feasible scheme of our improved control system.

## References

[1] Jong M. Park, Allan T. Evans, K. Rasmussen, et al. A Microvalve with Integrated Sensors and Customizable Normal State for Low-temperature Operation. Journal of Microelectromechanical Systems, Vol. 18, No.4, August 2009. 868-877

[2] R. Jayaseelan, T. Mitra. Temperature Aware Scheduling for Embedded Processors. Journal of Low Power Electronics, American Scientific Publisher, 5(3), Oct. 2009

[3] Matt Smith. Measuring temperatures on computer chips with speed and accuracy, Analog

[4] M. Moonat. Using the On-Chip Thermal Diode on Analog Devices Processors. ANALOG DEVICES: Technical notes on using Analog Devices DSPs, processors and development tools Rev 2, Aug. 2, 2010

[5] David Hanrahan. Fan-Speed Control Techniques in PCs. Analog Dialogue 34-4 (2000)

[6] Ed. Grochowski, Murali Annavaram. Energy per Instruction Trends in Intel Microprocessors. Microarchitecture Reserch Lab, Intel Corporation

[7] E. Rotem, J. Hermerding, C. Aviad, et al. Temperature Measurement in the Intel Core Duo Processor. Intel Document Published August 2008, 45nm Desktop Dual Core Processors Intel Core 2 Duo processor E8000 and E7000 series

[8] G. Paci, F. Poletti, L. Benini, et al. Exploring Temperature-aware Design in Low-power MPSoCs.

[9] W. Wu, L. Jin, J. Yang, et al. Tan. Efficient Power Modeling and Software Thermal Sensing for Run time Temperature Monitoring. University of California at River side

[10] Taliver Heath, Ana Paula Centeno, Pradeep George, et al. Mercury and Freon: Temperature Emulation and Management for Server Systems [J]. ASPLOS'06 October 21-25 2006, San Jose, California, USA.

[11] Kevin Skadron, Mircea R. Stan, Wei Huang, et al. Temperature-Aware Microarchitecture: Extended Discussion and Results. REPORT CS-2003-08 APRIL 2003.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# *Command Switching Strategy Based Safety Protection Control for Aeroengines*

Chao Chen
State Key Laboratory of Synthetical Automation for Process Industries
Northeastern University
Shenyang 110819, P. R. China
amos.orchid@yahoo.com.cn

Xi-Ming Sun
School of Control Science and Engineering
Dalian University of Technology
Dalian 116024, P. R. China
sunxm@dlut.edu.cn

*Abstract*—**Based on the simplified hypersonic air-breathing propulsion model, this paper studies the output regulation/ safety protection multi-objective switching control problem focusing on the safety boundaries existing during the working progressing. Command switching strategy based on the safety margin is researched. In the safe region with a sufficiently large safety margin, the regulation loop is active and the regulated output is controlled to track the reference signal as quickly as possible. But the protected loop will be switched on and the protected output will be forced to escape from dangers once the safety boundaries approach. A dynamical state feedback controller and a protection controller work in turn in a hysteresis switching way to guarantee the asymptotic tracking with certain safety performance. The conditions under which the asymptotic tracking could be guaranteed are given and the control parameters could be calculated by solving optimal problems. It is pointed out in this paper that the designing of the regulation controller and the protection controller could be implemented separately, and the control parameters could be optimized to get certain optimal performance index using numerical method. Finally, simulation researches are performed to verify the effectiveness of the given methods, which also indicate that the commands switching control can improve both safety margin and the dynamical performance indices than the single controller.**

***Keywords-command switching control; safety protection; aeroengine; safety margin; optimization; tracking***

## I. INTRODUCTION

There have been dramatically large amounts of attentions attracted by the modeling and control of hypersonic air-breathing propulsion in the recent years [1], especially since the successful flight experiment of the X-51 aircraft in late May 2010. Usually, a hypersonic air-breathing vehicle is driven by a scramjet engine, and is with the integrated airframe/ scramjet configuration which causes strong couplings among flight dynamics, aerodynamics, propulsion and control [2]. The coupled dynamics results in various engine safety boundaries during the working progressing of a hypersonic air-breathing vehicle, for example, the combustion stabilization boundary, the combustor wall temperature limitation, the inlet channel unstarting boundary, and so on. Disastrous accidents may happen once the system works outside the safety boundaries or even approaches them. On the other hand, due to the requirement of air and space travel, hypersonic vehicle must have a broader flight envelope than any other aircrafts, which brings large parametric uncertainties and quick dynamical changes. To pursue more excellent flight performances, the hypersonic air-breathing vehicles are usually required to flight near the safety boundaries as possible as it can. In order to balance the contradiction between the performance and safety, protection measures must be taken to the hypersonic air-breathing vehicle when it is approaching the safety boundaries. And one of the effective measures to finish this objective is to adopt the multi-objective switching control strategy [3].

When asked to tracking a large step command as quickly as possible, it is necessary to design regulating/protecting switching controllers for a hypersonic air-breathing vehicle since, in common sense, it is obviously that the rapidity usually conflicts with the security. A regulation/protection switching control system consists of a regulation loop and several protection loops, when the controlled plant is working within the allowed safety limits, the regulation loop works to achieve perfect performances, when the controlled plant trends towards the neighborhood of some safety boundary, the corresponding protection loop will be switched on automatically according to certain previous design to ensure safety. The protection loops are expected to be constructed according to the boundaries so that the designers could relax the safety condition when they are dealing with the regulation loop. It is with more efficiency and is less conservative.

It is notable that the safety protection control problem is similar to the state constrained control problem which is studied via the barrier Lyapunov function in [4], but there are differences between them. First, for example, not only boundary but also transient performances, especially that of the states except the constrained ones, should be considered in the safety protection control problem. Second, the safety protection controller is usually expected to be designed separately from the regulation controller, and sufficient safety margin is also usually needed.

Consisting of the determination of both the controllers and the switching law, design of the switching controller is a complex synthesis procedure of nonlinear controller, even it is for a linear plant. A systematic designing method for the safety protection switching controller is anticipated to come out, by

This work was supported in part by the NSFC under Grants 61174073, 90816028, 61174058 and 61004020, and the FRFCU under Grant N100604005, and the Program for New Century Excellent Talents in University NCET-09-0257, and the Fundamental Research Funds for the Central Universities DUT11SX07.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

both the researchers working on the switched systems theory and the engineers working on the practical fields. Most of the currently existing theory results [5], however, focus only on the analysis of stability of switched systems, and few of the switching laws present in the existing theoretical results are so easy to implement. And few results can be referred to design regulation/protection switching controllers. The work [6] uses switching static feedback controller with the maximum control switching law to deal with the constant tracking problem for a linear system with state constraints, but the dynamical response performance is not discussed. In [3], the regulation/ protection switching controller based on the minimum law is studied through simulation, without stability analysis and designing steps mentioned. In our previous work [7] and [8], the switching control based on minimum control switching law and the command switching control scheme are investigated respectively.

Based on our previous work in [8], this paper continue to studies the output regulation/safety protection control problem for the simplified model of the aeroengine via command switching control based on the safety margin. Several modifications are performed to improve the response performance. The proposed asymptotic tracking condition could be used to design the control parameters through solving programming problems. The design of the regulation loop and the protection loop can be finished in steps, and the control parameters can be optimized in a numerical method to get optimal ITEA index. Conditions for the finite times switching and estimation of the safety margin are also mentioned. Finally, in the section of simulation researches, the effectiveness of the given results verified, and it is revealed that the proposed command switching strategy is with superiority both in safety performance and ITAE (Integral of Time-weighted Absolute Error) index comparing to the single controller scheme.

## II. Problem Depiction

As is shown in Figure 1, the linear system studied in this paper can be considered as the simplified model of a hypersonic air-breathing propulsion, where the actuator and the basic engine are simply modeled as inertia units $G_1(s)$ and $G_2(s)$ respectively, while the temperature sensor is considered as a leading unit $G_3(s)$ approximately, where

$$G_1(s) = \frac{1}{T_4 s + 1}, \quad G_2(s) = \frac{1}{T_3 s + 1}, \quad G_3(s) = \frac{T_1 s + 1}{T_2 s + 1}.$$

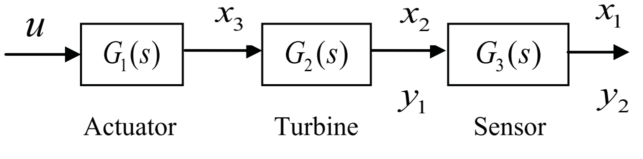


Figure 1. A simplified model of hypersonic air-breathing propulsion

As the regulated output, the rotation velocity of the turbine is expected to track the reference signal as quick as possible. Simultaneously, however, as the protected output, the temperature is not allowed to touch the constant safety boundary. To make the state available, the state variables $x_1$, $x_2$ and $x_3$ are chosen as the temperature, the rotation speed and the output of the actuator respectively, the state space model of the plant is

$$\dot{x} = Ax + bu, \tag{1}$$

where

$$A = -\begin{bmatrix} T_3T_4 & T_4(T_1 - T_3) & -T_1T_4 \\ 0 & T_2T_4 & -T_2T_4 \\ 0 & 0 & T_2T_3 \end{bmatrix} / T_2T_3T_4,$$

$$b = \begin{bmatrix} 0 & 0 & 1/T_4 \end{bmatrix}^T, x = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}^T \in R^3,$$

$u$ is the output of the controller, $T_1$, $T_2$, $T_3$ and $T_4$ are time constants determined according to identification, the regulated output and the protected output are

$$\begin{aligned} y_1 &= C_1 x, \\ y_2 &= fC_2 x, \end{aligned} \tag{2}$$

where $f$ is a constant feedback gain, and

$$C_1 = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}, C_2 = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}.$$

The control objective is to design control strategy to make

$$\lim_{t \to \infty} |y_{r1} - y_1(t)| = 0, \tag{3}$$

$$y_2(t) < y_{boundary}, \forall t, \tag{4}$$

where $y_{r1}$ is a constant reference signal, and $y_{boundary}$ is the constant safety boundary for the temperature $y_2$. In the paper, it is assumed that $y_{r1} < y_{r2} / f$, which stands for that the tracking task does not destroy the safety boundary and the desired equilibrium point of the closed loop is in the safety region.

Usually, it is more practical to think about the safety with some margin than only to consider the critical safety. In this paper, the safety margin $\gamma$ is defined as followed:

$$\gamma \overset{def}{=} \frac{y_{boundary} - y_2}{y_{boundary}}. \tag{5}$$

It is supposed in this paper that $y_{boundary} > 0$, and $0 < \gamma < 1$ implies that the system is working in the safe region, and a larger $\gamma$ stands for the safer status.

## III. Main Results

In this section, the command switching based strategy is adopted to solve the output regulation/safety protection control problem. The asymptotic tracking conditions are given, and the control parameters of the protecting controller could be optimized to minimize the ITAE performance index after the design of the regulation loop is finished.

### *A. Command Switching Control Based on Safety Margin*

In this section, we try to modify the command switching control scheme mentioned in our previous work [8] to obtain better control performances. The modified controllers are directly illustrated as followed:

First of all, a full state dynamical feedback controller is employed instead of a PI controller in the regulation loop to make the poles of the closed loop arbitrarily assignable under the controllable condition:

$$u_1(t) = K_x x(t) + K_q q(t), \tag{6}$$

where $q(t) = \int_0^t (y_{r1} - y_1(\tau))d\tau$ , and $\begin{bmatrix} K_x & K_q \end{bmatrix}$ is to be determined according to the desired poles.

The resulted closed loop equation of the regulation loop is

$$\dot{X} = A_{c1}(X - X_e), \tag{7}$$

where

$$X = [x^T, q]^T, A_{c1} = A_1 + B_1\begin{bmatrix} K_x & K_q \end{bmatrix},$$

$$X_e = -A_{c1}^{-1}\begin{bmatrix} 0_{3\times1} \\ y_{r1} \end{bmatrix}, A_1 = \begin{bmatrix} A & 0_{3\times1} \\ -C_1 & 0 \end{bmatrix}, B_1 = \begin{bmatrix} b \\ 0 \end{bmatrix}.$$

Second, a proportional (P) controller is adopted in the protection loop to help the protected output escape from danger as soon as possible, as the same time, the output of the integrator $q$ is set to vary more quickly during the protection interval than the regulation interval to improve the regulating performance:

$$\begin{aligned} &u_2(t) = K_p(Ly_{boundary} - y_2(t)), \\ &\dot{q}(t) = K(y_{r1} - y_1(t)), \sigma(t) = 2, \end{aligned} \tag{8}$$

where $L < 1, K > 1$ and $K_p$ are constants to be designed. It can be known that when $y_1(t) < y_{r1}$ , a quickly increasing $q$ results in a quickly increasing positive control input, which helps $y_1(t)$ to reach $y_{r1}$ quickly, and that when $y_1(t) > y_{r1}$ , a quickly decreasing $q$ brings a quickly decreasing negative control input, which helps $y_1(t)$ to reduce overshooting and to reduce times of switches. The corresponding equation of the protection loop is

$$\dot{X} = A_{c2}X + \xi_2, \tag{9}$$

where

$$A_{c2} = \begin{bmatrix} A - fK_p bC_2 & 0 \\ -KC_1 & 0 \end{bmatrix}, \xi_2 = \begin{bmatrix} Ly_b K_p b \\ Ky_{r1} \end{bmatrix}.$$

Third, the switching law is based on the safety margin to make the margin easy to estimate and is with hysteresis to better coin with practice, chattering also could be weakened due to the hysteresis:

$$u = u_{\sigma(t)}, \sigma(t) = \begin{cases} 1, \gamma > \gamma_{off}, \\ \sigma(t^-), \gamma_{on} < \gamma < \gamma_{off}, \\ 2, \gamma < \gamma_{on}, \end{cases} \tag{10}$$

where $0 < \gamma_{on} < \gamma_{off} < 1$ are constants to be designed. The switching law means that the protection loop will be switched on when the safety margin falls down to a given small number and will be switched off until the margin rises up to a sufficient large value. It is reasonable to assume that $\gamma(t_0) > \gamma_{off}$ , which means that it is safe enough at the initial moment, so $\sigma(t_0) = 1$ .The asymptotic tracking conditions are proposed as followed:

**Theorem1** Consider the system (1)-(2) controlled by the command switching controller (6), (8) and (10). Suppose that, there exists $t_1 > t_0$ such that for any $t > t_1$ , we have $y_1(t) > y_{r1}$ . If there exist a symmetric positive definite matrix $R$ and a positive constant $\beta$ satisfying

$$A_{c1}^T R + RA_{c1} \le -\beta I, \tag{11}$$

and the maximum value of the following quadratic programming problem (12) is non-positive, then the goal of asymptotic tracking (3) could be achieved:

$$\begin{aligned} \max \quad & (X - X_e)^T (A_{c2}^T R + RA_{c2})(X - X_e) \\ & + 2(A_{c2}X_e + \xi_2)R(X - X_e) \\ s.t. \quad & \begin{bmatrix} fC_2 & 0 \end{bmatrix} X \ge (1 - \gamma_{on}) y_{boundary}, \\ & \begin{bmatrix} C_1 & 0 \end{bmatrix} X \ge y_{r1.} \end{aligned} \tag{12}$$

**Proof**: If the candidate Lyapunov function is selected as

$$V(X) = (X - X_e)^T R(X - X_e), \tag{13}$$

when $\sigma(t) = 2$ , according to (9), we have

$$\begin{aligned} \dot{V} = &(X - X_e)^T (A_{c2}^T R + RA_{c2})(X - X_e) \\ &+ 2(A_{c2}X_e + \xi_2)R(X - X_e). \end{aligned} \tag{14}$$

It can be known from the condition (11) and (12) that $V(X)$ decreases in the whole time weather the regulation or the protection loop is switched on. The objective (3) can be thus realized. The proof is thus completed.

The quadratic programming problem (12) could be straightly solved using the optimal toolbox of Matlab. And the assumption of achievement of $y_1(t) > y_{r1}$ in finite time will be discussed combining with the simulations later.

**Remark1** In common sense, the peak value which a given output of a fixed system could achieve is determined by the initial state. So the safety margin in the proposed frame of command switching control based on the safety margin can be estimated by the following nonlinear programming [9]:

$$\begin{aligned}&\max \quad fC_2[x_e+e^{A_2t_m}(x_0-x_e)]\\ &x_0,t_m s.t.\begin{cases} fC_2x_0=(1-\gamma_{on})y_{boundary},\\ fC_2A_2x_0\geq 0,\\ fC_2A_2[x_e+e^{A_2t_m}(x_0-x_e)]=0.\end{cases}\end{aligned} \tag{15}$$

where $x_0\in R^3, t_m\in R^+$ are variables, and

$$A_2=A-fK_pC_2b,\ x_e=-Ly_bK_pA_2^{-1}b.$$

If the maximum value of (15) $y_m<y_{boundary}$, then we can easily get that $\gamma\geq\dfrac{y_{boundary}-y_m}{y_{boundary}}$.

**Remark2** Suppose that $\{t_{oni}\}$ and $\{t_{offi}\}$ $i=1,2,\cdots$ stand for the instants when the protection loop is switched on and switched off respectively, and $\{y_{2\max i}\}$ is the series formed by the peak values which $y_2(t)$ could achieve from the initial state $X(t_{offi})$ during the regulation intervals. One conclusion lies in that if $\{y_{2\max i}\}$ declines strictly when $y_1(t)>y_{r1}$, then switching must happen finite times and the finally acted loop must be the regulation loop, and the asymptotic tracking is consequently achieved.

### *B. Optimation for the Control Parameters*

The superiority of the multi-objective switching control schedule lies in that controllers corresponding to different goals could be designed separately. Take the command switching strategy proposed in this paper for an example, the control parameters of the protection loop, such as $K$, $K_p$ and $L$ are expected to be determined after the poles of the regulation loop are assigned. In other words, it is desirable that all the information of the regulation loop be available when the protected loop is under designing.

Under the asymptotic tracking condition, the three parameters mentioned above could be optimized by solving the following optimization problem:

$$\begin{aligned}&\min \quad J_{ITAE}\\ &K,K_p,L\ s.t. y_2(t)<y_{boundary},\forall t,\end{aligned} \tag{16}$$

where the performance index is chosen as the famous Integral of Time-weighted Absolute Error (ITAE) index:

$$J_{ITAE}=\int_0^\infty \tau|e(\tau)|d\tau=\int_0^\infty \tau|y_{r1}-y_1(\tau)|d\tau. \tag{17}$$

The parameters minimizing $J_{ITAE}$ of the closed-loop switched system (7)-(10) could be obtained by solving (16). With MATLAB, the powerful calculation tool, the optimization problem can be easily solved depending on pure numerical methods [10].

It is notable that if we can find a group of parameters corresponding to the minimum $J_{ITAE}$, then we have $J_{ITAE}<\infty$, and we have

$$\lim_{t\to\infty} t|y_{r1}-y_1(t)|=0, \tag{18}$$

which implies the conclusion of asymptotic tracking.

### *C. Simulation Researches*

According to the identification to the hypersonic air-breathing propulsion, the four time constants can be determined as followed:

$$T_1=0.4, T_2=0.15, T_3=0.3, T_4=0.15,$$

and we choose $y_{r1}=1, f=1$.

First of all, according to the rapidity requirement, we assign the poles of the regulation loop at $\lambda_{1,2,3,4}=-36$, resulted performance index is $J_{ITAE}=0.0101$, and as is shown in Figure 2, the corresponding trajectory of the protected output $y_2(t)$ goes through the given safety boundary $y_{boundary}=2$, which is not permissible for the safety demand.

The designing of the protection loop is thus necessary, and a group of permitted control parameters of the protection loop and that of the switching law are worked out according to the conditions in Theorem 1:

$$L=-19.5, K=15, K_p=20, \gamma_{on}=0.1, \gamma_{off}=0.11.$$

Figure 3 verifies the effectiveness of Theorem 1.

Moreover, the indices reflecting the performance can be obtained: controlled by the command switching controller, the maximum value of the protected output $y_2(t)$ is reduced to 1.9748, and the corresponding 2% error setting time of the regulated output $y_1(t)$ is $t_s=0.4168s$, and $J_{ITAE}=0.0122$.

It can also be seen from Figure 3 that when $y_1(t)<y_{r1}$, the output of the integrator $q$ is an increasing positive variable no matter which loop is switched on, which makes $y_1(t)$ rise during the regulation intervals, and the regulated output will go through its reference signal since $q$ increases even more quickly during the protection intervals according to (8). After the moment when $y_1(t)>y_{r1}$, during the regulation interval, the positive $q$ will lead $y_1(t)$ to rise, and the declining of the candidate Lyapunov function depicted as condition (12) limits the falling amplitude of $y_1(t)$ during the protection interval, which makes the assumption of achievement of $y_1(t)>y_{r1}$ in finite time used in Theorem1 possible under proper parameters. The curve of the regulated output in Figure 3 could illustrate this assumption.

What can also be illustrated by Figure 3 is the condition of the strict decline of $\{y_{2\max i}\}$ when $y_1(t) > y_{r1}$, as is mentioned in Remark2. After the instant when $y_1(t) > y_{r1}$, the decreasing $q$ leading to the decreasing control input may bring us the declining $\{y_{2\max i}\}$ as long as the positive feedback force during the protection interval is not quite large.

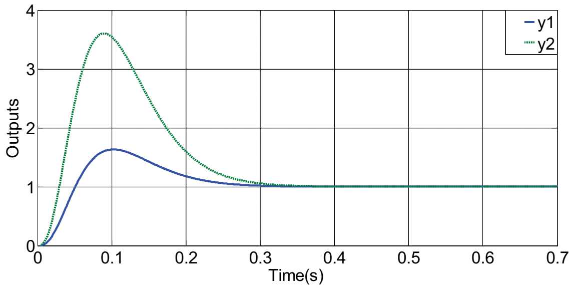


Figure 2. Curves of the regulated output and the protected output under single controller.

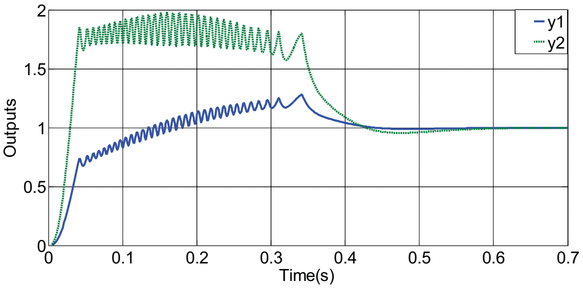


Figure 3. The improvement to the safety performance by the command switching control based on the safety margin.

Of course, replacing the poles of the closed loop nearer to the image axis could also reduce the peak value of the protected output $y_2(t)$, however, this will complicate the design procedure of the controller. And now we try to pull down the peak value of the protected output by replacing the poles in proportion, when the poles are assigned at $\lambda_{1,2,3,4} = -17.7$, as is shown in Figure 4, the peak value of $y_2(t)$ is 1.9911, and $t_s$ is 0.5018s, $J_{ITAE} = 0.0287$.

It could be deduced that pulling down the peak value of $y_2(t)$ to a lower level will make the setting time of the regulated output even longer, which indicates the superiority of the proposed command switching strategy.

To get better performance, we choose

$$L = -19.5, K = 15, K_p = 20$$

as the initial value of the control parameters and solve the optimization problem (16) to search parameters which can minimize the performance index $J_{ITAE}$ using the method introduced in [10], the obtained optimal control parameters are:

$$\begin{bmatrix} L & K & K_p \end{bmatrix} = \begin{bmatrix} -89.2672 & 61.7672 & 15.2793 \end{bmatrix},$$

the corresponding performance index is $J_{ITAE} = 0.0092$, and the curves are shown in Figure 5.

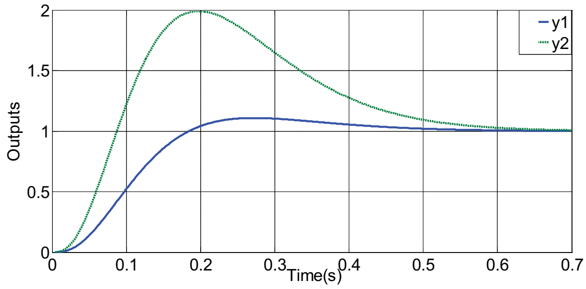


Figure 4. The improvement to the safety performance by replacing the poles of the closed loop nearer to the image axis.

It can be seen that under proper control parameters, the ITAE performance of a system using the command switching controller might be better than that of a system controlled by the single controller with the similar safety performance, and that under optimized control parameters, the ITAE performance of a system using the command switching controller could even be better than that of the single controller controlled system with the same regulation loop parameters (but without the permitted safety performances).

## IV. CONCLUSIONS

This paper investigated the output regulation/ safety protection switching control problem for the simplified aeroengine model. Focusing on the safety boundaries existing during the working progressing, the command switching strategy based on the safety margin are adopted and the scheme mentioned in our previous work is modified in several aspects to get better performances. A dynamical state feedback controller and a proportional protection controller switch in a hysteresis way to balance the regulation and the safety demands. The principles guaranteeing the asymptotic tracking in a safe way asymptotic tracking are proposed. The control

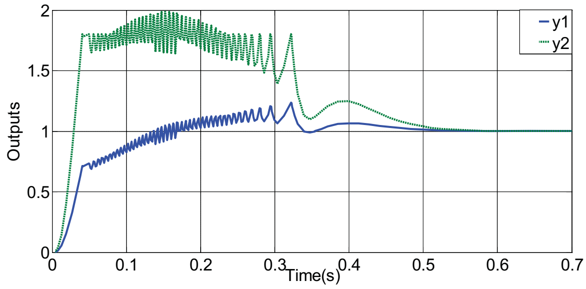


Figure 5. The improvement to the safety performance by the command switching control under the optimized control parameters.

parameters could be calculated by solving optimal problems and could be optimized to obtain some best performance indices. Finally, simulation researches are performed to verify the effectiveness of the given results. It also can be indicated through simulation researches that the commands switching controller bring more improvement in both safety margin and the ITAE index than the single controller.

## REFERENCES

[1] Rodriguez A., Dickeson J., Cifdaloz O., Kelkar A., Vogel J., Soloway D., Mc Cullen R., Benavides J., and Sridharan S., “Modeling and control of scramjet-powered hypersonic vehicles: challenges, trends, and tradeoffs,” 2008, AIAA paper 2008-6793

[2] Z. H. Yao, W. Bao and J. T. Chang, “Modelling for couplings of an airframe - propulsion integrated hypersonic vehicle with engine safety boundaries,” Proc. IMechE, Part G: J. Aerospace Engineering, 2010, 224(1): pp.43-55.

[3] Yu Daren, Liu Xiaofeng, “Multi-objective Robust Regulating and Protecting Control for Aero Engines. Journal of Engineering for Gas Turbines and Power,” July 2008, Vol. 130.

[4] K. P. Tee, S. S. Ge, and E. H. Tay, “Barrier Lyapunov Functions for the control of output-constrained nonlinear systems,” Automatica, 2009, vol. 45, no. 4, pp. 918-927.

[5] D. Liberzon, Switching in Systems and Control, Birkhauser, Boston, 2003.

[6] M. S. Branicky, Studies in hybrid systems: Modeling, analysis, and control, Ph.D. dissertation, Dept. of Electrical and Computer Engineering, Massachusetts Institute of Technology, 1995.

[7] Chen Chao, Yu Daren, Bao Wen, Zhao Jun, “Safety protection Switching Control for Aeroengines Based on Multiple Lyapunov Functions Method,” Proceedings of the 29th Chinese Control Conference, 2010, pp. 5953-5957.

[8] Chao Chen, Daren Yu, Wen Bao, Jun Zhao, “Safety Protection Control for a Class of Linear Systems via Switching Strategy,” Proceedings of the 30th Chinese Control Conference, 2011, pp. 1739-1744.

[9] Bertsekas D. P., Nonlinear Programming: Second Edition, Belmont, Massachusetts, Athena Scientific, 1999.

[10] Xue Dingyu, Computer Aided Control Systems Design Using Matlab Language (Second Edition), Tsinghua University Press, Beijing, 2006.

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# Inverse Optimal Robust Control of Singularly Impulsive Dynamical Systems

Dr. Nataša A. Kablar and Vlada Kvrgić

*Abstract*— **In this paper for the class of nonlinear uncertain singularly impulsive dynamical systems we present optimal robust control and inverse robust optimal control results. We consider a control problem for nonlinear uncertain singularly impulsive dynamical systems involving a notion of optimality with respect to an *auxiliary cost* which guarantees a bound on the worst-case value of a nonlinear-nonquadratic hybrid cost criterion over a prescribed uncertainty set. Further we specialize result to affine uncertain systems to obtain controllers predicated on an *inverse optimal hybrid control problem*. In particular, to avoid the complexity in solving the steady-state hybrid Hamilton-Jacobi-Bellman equation we parameterize a family of stabilizing hybrid controllers that minimize some *derived* hybrid cost functional that provides flexibility in specifying the control law. The performance integrand is shown to explicitly depend on the nonlinear singularly impulsive system dynamics, the Lyapunov function of the closed-loop system, and the stabilizing hybrid feedback control law wherein the coupling is introduced via the hybrid Hamilton-Jacobi-Bellman equation. By varying the parameters in the Lyapunov function and the performance integrand, the proposed framework can be used to characterize a class of globally stabilizing hybrid controllers that can meet the closed-loop system response constraints. Obtained results for nonlinear case are further specialized to linear singularly impulsive dynamical systems with polynomial and multilinear performance functional.**

*Index Terms*— **mathematical model, singularly impulsive dynamical systems, optimal robust control, inverse optimal robust control**

## I. Introduction

For the class of nonlinear uncertain singularly impulsive dynamical systems presented in [2], we have developed robust stability results in [7]. In this paper we give optimal robust control and inverse robust optimal control results. For that purpose, we generalize results developed in [3]. We consider a control problem for nonlinear uncertain singularly impulsive dynamical systems involving a notion of optimality with respect to an *auxiliary cost* which guarantees a bound on the worst-case value of a nonlinear-nonquadratic hybrid cost criterion over a prescribed uncertainty set. Further we specialize result to affine uncertain systems to obtain controllers predicated on an *inverse optimal hybrid control problem*. In particular, to avoid the complexity in solving the steady-state hybrid Hamilton-Jacobi-Bellman equation we parameterize a family of stabilizing hybrid controllers that minimize some *derived* hybrid cost functional that provides flexibility in specifying the control law. Obtained results for nonlinear case are further specialized to linear singularly impulsive dynamical systems with polynomial and multilinear performance functional.

Finally, in this paper we use the following standard notation. Let $\mathbb{R}$ denote the set of real numbers, let $\mathcal{N}$ denote the set of nonnegative integers, let $\mathbb{R}^n$ denote the set of $n\times 1$ real column vectors, let $\mathbb{R}^{n\times m}$ denote the set of $n\times m$ real matrices, let $\mathbb{S}^n$ denote the set of $n\times n$ symmetric matrices, and let $\mathbb{N}^n$ (resp., $\mathbb{P}^n$) denote the set of $n\times n$ nonnegative (resp., positive) definite matrices, and let $I_n$ or $I$ denote the $n\times n$ identity matrix. Furthermore, $A\geq 0$ (resp., $A>0$) denotes the fact that the Hermitian matrix is nonnegative (resp., positive) definite and $A\geq B$ (resp., $A>B$) denotes the fact that $A-B\geq 0$ (resp., $A-B>0$). In addition, we write $V'(x)$ for the Fréchet derivative of $V(\cdot)$ at x. Finally, let $\mathrm{C}^0$ denote the set of continuous functions and $\mathrm{C}^\mathrm{r}$ denote the set of functions with $r$ continuous derivatives.

## II. Optimal Robust Control for Nonlinear Uncertain Singularly Impulsive Dynamical Systems

In this section we consider a control problem for nonlinear uncertain singularly impulsive dynamical systems involving a notion of optimality with respect to an *auxiliary cost* which guarantees a bound on the worst-case value of a nonlinear-nonquadratic hybrid cost criterion over a prescribed uncertainty set. The optimal robust hybrid time-invariant feedback controllers are derived as a direct consequence of Theorem 2.1 given in [7] and provide a generalization of the Hamilton-Jacobi-Bellman conditions for state-dependent singularly impulsive dynamical systems with optimality notions over the infinite horizon with an infinite number of resetting times, for addressing robust feedback controllers of nonlinear uncertain singularly impulsive dynamical systems. To address robust optimal control problem let $\mathcal{D}\subset\mathbb{R}^n$ be an open set with $0\in\mathcal{D}$, and let $\mathcal{C}_\mathrm{c}\subset\mathbb{R}^{m_\mathrm{c}}$, $\mathcal{C}_\mathrm{d}\subset\mathbb{R}^{m_\mathrm{d}}$, where $0\in\mathcal{C}_\mathrm{c}$ and $0\in\mathcal{C}_\mathrm{d}$. Furthermore, let $\mathcal{F}_\mathrm{c}\subset\{F_\mathrm{c}:\mathcal{D}\times\mathcal{C}_\mathrm{c}\to\mathbb{R}^n : F_\mathrm{c}(0,0)=0\}$, and $\mathcal{F}_\mathrm{d}\subset\{F_\mathrm{d}:\mathcal{D}\times\mathcal{C}_\mathrm{d}\to\mathbb{R}^n : F_\mathrm{d}(0,0)=0\}$. For simplicity of exposition, we also define $(F_\mathrm{c}(\cdot,\cdot),F_\mathrm{d}(\cdot,\cdot))\in\mathcal{F}_\mathrm{c}\times\mathcal{F}_\mathrm{d}\triangleq\mathcal{F}$. Next, consider the nonlinear uncertain singularly impulsive controlled dynamical system

$$
\begin{aligned}
E_\mathrm{c}\dot{x}(t) &= F_\mathrm{c}(x(t),u_\mathrm{c}(t)),\quad x(0)=0,\quad x(t)\notin\mathcal{Z}_x,\\
&\qquad\qquad u_\mathrm{c}(t)\in\mathcal{U}_\mathrm{c}, \qquad \text{(II.1)}\\
E_\mathrm{d}\Delta x(t) &= F_\mathrm{d}(x(t),u_\mathrm{d}(t)),\quad x(t)\in\mathcal{Z}_x,\\
&\qquad\qquad u_\mathrm{d}(t)\in\mathcal{U}_\mathrm{d}, \qquad \text{(II.2)}
\end{aligned}
$$

where $t\geq 0$, $x(t)\in\mathcal{D}$ is the state vector, $(u_\mathrm{c}(t),u_\mathrm{d}(t_k))\in\mathcal{U}_\mathrm{c}\times\mathcal{U}_\mathrm{d}\subset\mathcal{C}_\mathrm{c}\times\mathcal{C}_\mathrm{d}$, $k\in\mathcal{N}$, is the hybrid control input,

Lola Institute, Kneza Viseslava 70a, Belgrade 11000, Serbia nkablar.ae01@gtalumni.org

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

where the control constraint sets $\mathcal{U}_{\rm c},\mathcal{U}_{\rm d}$ are given. We assume $(0,0)\in\mathcal{U}_{\rm c}\times\mathcal{U}_{\rm d}$, $F_{\rm c}:\mathcal{D}\times\mathcal{U}_{\rm c}\to\mathbb{R}^n$ is Lipschitz continuous and satisfies $F_{\rm c}(0,0)=0$, $F_{\rm d}:\mathcal{D}\times\mathcal{U}_{\rm d}\to\mathbb{R}^n$ is continuous and satisfies $F_{\rm d}(0,0)=0$, and $\mathcal{Z}_x\subset\mathbb{R}^n$. To address the robust optimal nonlinear hybrid feedback control problem let $\phi_{\rm c}:\mathcal{D}\to\mathcal{U}_{\rm c}$ be such that $\phi_{\rm c}(0)=0$ and let $\phi_{\rm d}:\mathcal{D}\to\mathcal{U}_{\rm d}$ be such that $\phi_{\rm d}(0)=0$. If $(u_{\rm c}(t),u_{\rm d}(t))=(\phi_{\rm c}(E_{\rm c}x(t)),\phi_{\rm d}(E_{\rm d}x(t)))$, where $x(t)$, $t\geq0$, satisfies (II.1), (II.2), then $(u_{\rm c}(\cdot),u_{\rm d}(\cdot))$ is a *hybrid feedback control*. Given the hybrid feedback control $(u_{\rm c}(t),u_{\rm d}(t))=(\phi_{\rm c}(E_{\rm c}x(t)),\phi_{\rm d}(E_{\rm d}x(t)))$, the closed-loop state-dependent singularly impulsive dynamical system has the form

$$\begin{aligned}E_{\rm c}\dot{x}(t)&=F_{\rm c}(x(t),\phi_{\rm c}(E_{\rm c}x(t))),\qquad x(0)=x_0,\\&\qquad t\geq0,\quad x(t)\notin\mathcal{Z}_x,\qquad\text{(II.3)}\\E_{\rm d}\Delta x(t)&=F_{\rm d}(x(t),\phi_{\rm d}(E_{\rm d}x(t))),\\&\qquad x(t)\in\mathcal{Z}_x,\qquad\text{(II.4)}\end{aligned}$$

for all $(F_{\rm c}(\cdot,\cdot),F_{\rm d}(\cdot,\cdot))\in\mathcal{F}$.

Next we present sufficient conditions for characterizing robust nonlinear hybrid feedback controllers that guarantee robust stability over a class of nonlinear uncertain singularly impulsive dynamical systems and minimize an auxiliary hybrid performance functional. For the statement of this result let $L_{\rm c}:\mathcal{D}\times\mathcal{U}_{\rm c}\to\mathbb{R}$, $L_{\rm d}:\mathcal{D}\times\mathcal{U}_{\rm d}\to\mathbb{R}$ and define the set of asymptotically stabilizing controllers for the nominal nonlinear singularly impulsive dynamical system $(F_{\rm c0}(\cdot,\cdot),F_{\rm d0}(\cdot,\cdot))$ by

$$\begin{aligned}\mathcal{C}(x_0)\triangleq\{(u_{\rm c}(\cdot),u_{\rm d}(\cdot))\ :\ &(u_{\rm c}(\cdot),u_{\rm d}(\cdot))\,\text{is admissible}\\&\text{and the zero solution}\,x(t)\equiv0\\&\text{to (II.1), (II.2)}\\&\text{is asymptotically stable with}\\&(F_{\rm c}(\cdot,\cdot),F_{\rm d}(\cdot,\cdot))\\&=(F_{\rm c0}(\cdot,\cdot),F_{\rm d0}(\cdot,\cdot))\}.\end{aligned}$$

Consider the nonlinear uncertain singularly impulsive dynamical system (II.1), (II.2) with hybrid performance functional

$$\begin{aligned}J(E_{\rm c}x_0,u_{\rm c}(\cdot),u_{\rm d}(\cdot))\ =\ &\int_0^\infty L_{\rm c}(E_{\rm c}x(t),u(t)){\rm d}t\\&+\sum_{k\in\mathcal{N}_{[0,\infty)}}L_{\rm d}(E_{\rm d}x(t_k),u_{\rm d}(t_k)),\qquad\text{(II.5)}\end{aligned}$$

where $(\mathcal{F}_{\rm c}(\cdot,\cdot),\mathcal{F}_{\rm d}(\cdot,\cdot))\in\mathcal{F}$ and $(u_{\rm c}(\cdot),u_{\rm d}(\cdot))$ is an admissible control. Assume there exist functions $V:\mathcal{D}\to\mathbb{R}$, $\Gamma_{\rm c}:\mathcal{D}\times\mathcal{U}_{\rm c}\to\mathbb{R}$, $\Gamma_{\rm d}:\mathcal{D}\times\mathcal{U}_{\rm d}\to\mathbb{R}$, and a hybrid control law $\phi_{\rm c}:\mathcal{D}\to\mathcal{U}_{\rm c}$ and $\phi_{\rm d}:\mathcal{D}\to\mathcal{U}_{\rm d}$, where $V(\cdot)$ is a C$^1$ function, such that

$$\begin{aligned}V(0)&=0,&&\text{(II.6)}\\V(E_{\rm c}x)&\geq0,\quad x\in\mathcal{D},\,x\neq0,&&\text{(II.7)}\\\phi_{\rm c}(0)&=0,&&\text{(II.8)}\\\phi_{\rm d}(0)&=0,&&\text{(II.9)}\\V'(E_{\rm c}x)F_{\rm c}(x,\phi_{\rm c}(x))&\leq V'(E_{\rm c}x)F_{\rm c0}(x,\phi_{\rm c}(x))\\&\quad+\Gamma_{\rm c}(x,\phi_{\rm c}(x)),\quad x\notin\mathcal{Z}_x,\quad F_{\rm c}(\cdot,\cdot)\in\mathcal{F}_{\rm c},&&\text{(II.10)}\end{aligned}$$

$$V'(E_{\rm c}x)F_{\rm c0}(x,\phi_{\rm c}(x))\ +\ \Gamma_{\rm c}(x,\phi_{\rm c}(x))<0,\ x\notin\mathcal{Z}_x,\ x\neq0,\qquad\text{(II.11)}$$

$$\begin{aligned}V(E_{\rm d}x+F_{\rm d}(x,\phi_{\rm d}(x)))\ &-\ V(E_{\rm d}x)\leq\\&V(E_{\rm d}x+F_{\rm d0}(x,\phi_{\rm d}(x)))-V(E_{\rm d}x)\\&+\Gamma_{\rm d}(x,\phi_{\rm d}(x)),\quad x\in\mathcal{Z}_x,\ F_{\rm d}(\cdot,\cdot)\in\mathcal{F}_{\rm d},\qquad\text{(II.12)}\\V(E_{\rm d}x+F_{\rm d0}(x,\phi_{\rm d}(x)))\ &-\ V(E_{\rm d}x)+\Gamma_{\rm d}(x,\phi_{\rm d}(x))\\&\leq0,\qquad x\in\mathcal{Z}_x,\qquad\text{(II.13)}\end{aligned}$$

$$\begin{aligned}H_{\rm c}(E_{\rm c}x,\phi_{\rm c}(x))&=0,\quad x\notin\mathcal{Z}_x,&&\text{(II.14)}\\H_{\rm c}(E_{\rm c}x,u_{\rm c}(x))&\geq0,\quad x\notin\mathcal{Z}_x,\ u_{\rm c}\in\mathcal{U}_{\rm c},&&\text{(II.15)}\\H_{\rm d}(E_{\rm d}x,\phi_{\rm d}(E_{\rm c}x))&=0,\quad x\in\mathcal{Z}_x,&&\text{(II.16)}\\H_{\rm d}(E_{\rm d}x,u_{\rm d}(x))&\geq0,\quad x\in\mathcal{Z}_x,\ u_{\rm d}\in\mathcal{U}_{\rm d},&&\text{(II.17)}\end{aligned}$$

where $(F_{\rm c0}(\cdot,\cdot),F_{\rm d0}(\cdot,\cdot))\in\mathcal{F}$ defines the nominal singularly impulsive dynamical system and

$$H_{\rm c}(E_{\rm c}x,u_{\rm c})\ \triangleq\ L_{\rm c}(E_{\rm c}x,u_{\rm c})+V'(E_{\rm c}x)F_{\rm c0}(x,u_{\rm c})+\Gamma_{\rm c}(x,u_{\rm c}),\qquad\text{(II.18)}$$

$$\begin{aligned}H_{\rm d}(E_{\rm d}x,u_{\rm d})\ \triangleq\ &L_{\rm d}(E_{\rm d}x,u_{\rm d})+V(E_{\rm d}x+F_{\rm d0}(x,u_{\rm d}))\\&-V(xE_{\rm d})+\Gamma_{\rm d}(x,u_{\rm d}).\qquad\text{(II.19)}\end{aligned}$$

Then, with the hybrid feedback control $(u_{\rm c}(\cdot),u_{\rm d}(\cdot))=(\phi_{\rm c}(E_{\rm c}x(\cdot)),\phi_{\rm d}(E_{\rm d}x(\cdot)))$, there exists a neighborhood of the origin $\mathcal{D}_0\subset\mathcal{D}$ such that if $x_0\in\mathcal{D}_0$, the zero solution $x(t)\equiv0$ of the closed-loop system (II.3), (II.4) is locally asymptotically stable for all $(F_{\rm c}(\cdot,\cdot),F_{\rm d}(\cdot,\cdot))\in\mathcal{F}$. Furthermore,

$$\begin{aligned}\sup_{(F_{\rm c}(\cdot,\cdot),F_{\rm d}(\cdot,\cdot))\in\mathcal{F}}\ &J(E_{\rm c}x_0,\phi_{\rm c}(E_{\rm c}x(\cdot)),\phi_{\rm d}(E_{\rm d}x(\cdot)))\\&\leq\mathcal{J}(E_{\rm c}x_0,\phi_{\rm c}(\cdot),\phi_{\rm d}(\cdot))\\&=V(E_{\rm c}x_0),\quad x_0\in\mathcal{D}_0,\qquad\text{(II.20)}\end{aligned}$$

where

$$\begin{aligned}\mathcal{J}(E_{\rm c}x_0,&u_{\rm c}(\cdot),u_{\rm d}(\cdot))\ \triangleq\\&\int_0^\infty[L_{\rm c}(E_{\rm c}x(t),u_{\rm c}(t))+\Gamma_{\rm c}(x(t),u_{\rm c}(t))]{\rm d}t\\&+\sum_{k\in\mathcal{N}_{[0,\infty)}}[L_{\rm d}(E_{\rm d}x(t_k),u_{\rm d}(t_k))+\Gamma_{\rm d}(x(t_k),u_{\rm d}(t_k))],\end{aligned}\qquad\text{(II.21)}$$

and where $(u_{\rm c}(\cdot),u_{\rm d}(\cdot))$ is an admissible control and $x(t)$, $t\geq0$, is a solution of (II.1), (II.2) with $(F_{\rm c}(x(t),u_{\rm c}(t)),F_{\rm d}(x(t),u_{\rm d}(t)))=(F_{\rm c0}(x(t),u_{\rm c}(t)),F_{\rm d0}(x(t),u_{\rm d}(t)))$. In addition, if $x_0\in\mathcal{D}_0$ then the hybrid feedback control $(u_{\rm c}(\cdot),u_{\rm d}(\cdot))=(\phi_{\rm c}(E_{\rm c}x(\cdot)),\phi_{\rm d}(E_{\rm d}x(\cdot)))$ minimizes $J(E_{\rm c}x_0,u_{\rm c}(\cdot),u_{\rm d}(\cdot))$ in the sense that

$$\begin{aligned}&J(E_{\rm c}x_0,\phi_{\rm c}(E_{\rm c}x(\cdot)),\phi_{\rm d}(E_{\rm d}x(\cdot)))=\\&\min\nolimits_{(u_{\rm c}(\cdot),u_{\rm d}(\cdot))\in\mathcal{C}(x_0)}J(E_{\rm c}x_0,u_{\rm c}(\cdot),u_{\rm d}(\cdot)).\qquad\text{(II.22)}\end{aligned}$$

Finally, if $\mathcal{D}=\mathbb{R}^n$, and

$$V(E_{\rm c/d}x)\to\infty\quad\text{as}\quad\|x\|\to\infty,\qquad\text{(II.23)}$$

then the zero solution $x(t)\equiv0$ of the closed-loop system (II.3), (II.4) is globally asymptotically stable for all $(F_{\rm c}(\cdot,\cdot),F_{\rm d}(\cdot,\cdot))\in\mathcal{F}$, [3] and [7].

*Proof:* Local and global asymptotic stability is a direct consequence of (II.6)–(II.13) by applying Theorem 2.1 of [7] to the closed-loop system (II.3), (II.4). Next, let

$(u_{\mathrm{c}}(\cdot), u_{\mathrm{d}}(\cdot)) \in \mathcal{C}(x_0)$ and let $x(\cdot)$ be the solution of (II.1), (II.2) with $(F_{\mathrm{c}}(\cdot,\cdot), F_{\mathrm{d}}(\cdot,\cdot)) = (F_{\mathrm{c}0}(\cdot,\cdot), F_{\mathrm{d}0}(\cdot,\cdot))$.

Then it follows that

$$\begin{aligned} 0 &= -\dot{V}(E_{\mathrm{c}}x(t)) + V'(E_{\mathrm{c}}x(t))F_{\mathrm{c}}(x(t), u_{\mathrm{c}}(t)),\ x(t) \notin \mathcal{Z}_x, && \text{(II.24)} \\ 0 &= -\Delta V(E_{\mathrm{d}}x(t)) + V(E_{\mathrm{d}}x + F_{\mathrm{d}}(x(t), u_{\mathrm{d}}(t))) \\ &\quad -V(E_{\mathrm{d}}x(t)), \qquad x(t) \in \mathcal{Z}_x. && \text{(II.25)} \end{aligned}$$

Hence,

$$\begin{aligned} L_{\mathrm{c}}(E_{\mathrm{c}}x(t), u_{\mathrm{c}}(t)) + \Gamma_{\mathrm{c}}(E_{\mathrm{c}}\tilde{x}(t), u_{\mathrm{c}}(t)) \ &= \\ -\dot{V}(E_{\mathrm{c}}x(t)) + L_{\mathrm{c}}(E_{\mathrm{c}}x(t), u_{\mathrm{c}}(t)) & \\ +V'(E_{\mathrm{c}}x(t))F_{\mathrm{c}0}(x(t), u_{\mathrm{c}}(t)) + \Gamma_{\mathrm{c}}(E_{\mathrm{c}}\tilde{x}(t), u_{\mathrm{c}}(t)) & \\ = -\dot{V}(E_{\mathrm{c}}x(t)) + H_{\mathrm{c}}(E_{\mathrm{c}}x(t), u_{\mathrm{c}}(t)), & \\ x(t) \notin \mathcal{Z}_x. & \qquad \text{(II.26)} \end{aligned}$$

Similarly,

$$\begin{aligned} L_{\mathrm{d}}(E_{\mathrm{d}}x(t), u_{\mathrm{d}}(t)) + \Gamma_{\mathrm{d}}(x(t), u_{\mathrm{d}}(t)) \ &= \\ -\Delta V(E_{\mathrm{d}}x(t)) + L_{\mathrm{d}}(E_{\mathrm{d}}x(t), u_{\mathrm{d}}(t)) & \\ +\Delta V(E_{\mathrm{d}}x(t)) + \Gamma_{\mathrm{d}}(x(t), u_{\mathrm{d}}(t)) & \\ = -\Delta V(E_{\mathrm{d}}x(t)) + H_{\mathrm{d}}(E_{\mathrm{d}}x(t), u_{\mathrm{d}}(t)), & \\ x(t) \in \mathcal{Z}_x. & \qquad \text{(II.27)} \end{aligned}$$

Now, over the interval $[0, t)$ yields

$$\begin{aligned} &\int_0^t [L_{\mathrm{c}}(E_{\mathrm{c}}x(t), u_{\mathrm{c}}(t)) \quad + \quad \Gamma_{\mathrm{c}}(\tilde{x}(t), u_{\mathrm{c}}(t))]\mathrm{d}t \\ &+ \sum_{k \in \mathcal{N}_{[0,t)}} [L_{\mathrm{d}}(E_{\mathrm{d}}x(t_k), u_{\mathrm{d}}(t_k)) + \Gamma_{\mathrm{d}}(x(t_k), u_{\mathrm{d}}(t_k))] \\ &= \int_0^t [-\dot{V}(E_{\mathrm{c}}x(t)) + H_{\mathrm{c}}(x(t), u_{\mathrm{c}}(t))]\mathrm{d}t \\ &+ \sum_{k \in \mathcal{N}_{[0,t)}} [-\Delta V(E_{\mathrm{d}}x(t_k)) + H_{\mathrm{d}}(E_{\mathrm{d}}x(t_k), u_{\mathrm{d}}(t_k))] \\ &= -V(E_{\mathrm{c}}x(t)) + V(E_{\mathrm{c}}x_0) + \int_0^t H_{\mathrm{c}}(E_{\mathrm{c}}x(t), u_{\mathrm{c}}(t))\mathrm{d}t \\ &+ \sum_{k \in \mathcal{N}_{[0,t)}} H_{\mathrm{d}}(E_{\mathrm{d}}x(t_k), u_{\mathrm{d}}(t_k)) \\ &\geq V(E_{\mathrm{c}}x_0) \\ &= \mathcal{J}(E_{\mathrm{c}}x_0, \phi_{\mathrm{c}}(x(\cdot)), \phi_{\mathrm{d}}(x(\cdot))). \qquad \text{(II.28)} \end{aligned}$$

Letting $t \to \infty$ and noting that $V(E_{\mathrm{c/d}}x(t)) \to 0$ for all $x_0 \in \mathcal{D}_0$ yields (II.22). □

Next, we specialize Theorem II to linear uncertain singularly impulsive dynamical systems. Specifically, in this case we consider $\mathcal{F} \triangleq \mathcal{F}_{\mathrm{c}} \times \mathcal{F}_{\mathrm{d}}$ to be the set of uncertain linear singularly impulsive dynamical systems, where

$$\begin{aligned} \mathcal{F}_{\mathrm{c}} &= \{(A_{\mathrm{c}} + \Delta A_{\mathrm{c}})x + B_{\mathrm{c}}u_{\mathrm{c}} : x \in \mathbb{R}^n, A_{\mathrm{c}} \in \mathbb{R}^{n\times n}, \\ &\quad B_{\mathrm{c}} \in \mathbb{R}^{n\times m_{\mathrm{c}}}, \Delta A_{\mathrm{c}} \in \boldsymbol{\Delta}_{A_{\mathrm{c}}}\}, \\ \mathcal{F}_{\mathrm{d}} &= \{(A_{\mathrm{d}} + \Delta A_{\mathrm{d}} - E_{\mathrm{d}})x + B_{\mathrm{d}}u_{\mathrm{d}} : x \in \mathbb{R}^n, A_{\mathrm{d}} \in \mathbb{R}^{n\times n}, \\ &\quad B_{\mathrm{d}} \in \mathbb{R}^{n\times m_{\mathrm{d}}}, \Delta A_{\mathrm{d}} \in \boldsymbol{\Delta}_{A_{\mathrm{d}}}\}, \end{aligned}$$

where $\boldsymbol{\Delta}_{A_{\mathrm{c}}}, \boldsymbol{\Delta}_{A_{\mathrm{d}}} \subset \mathbb{R}^{n\times n}$, are given bounded uncertainty sets of uncertain perturbations $\Delta A_{\mathrm{c}}, \Delta A_{\mathrm{d}}$ of the nominal system matrices $A_{\mathrm{c}}, A_{\mathrm{d}}$, such that $0 \in \boldsymbol{\Delta}_{A_{\mathrm{c}}}$ and $0 \in \boldsymbol{\Delta}_{A_{\mathrm{d}}}$. For simplicity of exposition, we also define $(\Delta A_{\mathrm{c}}, \Delta A_{\mathrm{d}}) \in \boldsymbol{\Delta}_{A\mathrm{c}} \times \boldsymbol{\Delta}_{A\mathrm{d}} \triangleq \boldsymbol{\Delta}$. For the following result let $R_{\mathrm{c}1} \in \mathbb{P}^n$, $R_{\mathrm{c}2} \in \mathbb{P}^{m_{\mathrm{c}}}$, $R_{\mathrm{d}1} \in \mathbb{N}^n$, $R_{\mathrm{d}2} \in \mathbb{N}^{m_{\mathrm{d}}}$ be given.

Consider the linear state-dependent uncertain singularly impulsive controlled dynamical system

$$\begin{aligned} E_{\mathrm{c}}\dot{x}(t) &= (A_{\mathrm{c}} + \Delta A_{\mathrm{c}})x(t) + B_{\mathrm{c}}u_{\mathrm{c}}(t), \quad x(0) = x_0, \\ &\qquad t \geq 0, \quad x(t) \notin \mathcal{Z}, \qquad \text{(II.29)} \\ E_{\mathrm{d}}\Delta x(t) &= (A_{\mathrm{d}} + \Delta A_{\mathrm{d}} - E_{\mathrm{d}})x(t) + B_{\mathrm{d}}u_{\mathrm{d}}(t),\ x(t) \in \mathcal{Z}, \\ &\qquad \text{(II.30)} \end{aligned}$$

with performance functional

$$\begin{aligned} &J_{\Delta A_{\mathrm{c}}, \Delta A_{\mathrm{d}}}(E_{\mathrm{c}}x_0, u_{\mathrm{c}}(\cdot), u_{\mathrm{d}}(\cdot)) \triangleq \\ &\quad \int_0^\infty [x^{\mathrm{T}}(t)E_{\mathrm{c}}^{\mathrm{T}}R_{\mathrm{c}1}E_{\mathrm{c}}x(t) + u_{\mathrm{c}}^{\mathrm{T}}(t)R_{\mathrm{c}2}u_{\mathrm{c}}(t)]\mathrm{d}t \\ &\quad + \sum_{k \in \mathcal{N}_{[0,\infty)}} [x^{\mathrm{T}}(t_k)E_{\mathrm{d}}^{\mathrm{T}}R_{\mathrm{d}1}E_{\mathrm{d}}x(t_k) + u_{\mathrm{d}}^{\mathrm{T}}(t_k)R_{\mathrm{d}2}u_{\mathrm{d}}(t_k)], \\ &\qquad \text{(II.31)} \end{aligned}$$

where $(u_{\mathrm{c}}(\cdot), u_{\mathrm{d}}(\cdot))$ is admissible, $(\Delta A_{\mathrm{c}}, \Delta A_{\mathrm{d}}) \in \boldsymbol{\Delta}$. Furthermore, assume there exist $P \in \mathbb{P}^n$, $\Omega_{\mathrm{c}} : \mathbb{P}^n \to \mathbb{N}^n$, $\Omega_{\mathrm{d}xx} : \mathbb{P}^n \to \mathbb{N}^n$, $\Omega_{\mathrm{d}xu_{\mathrm{d}}} : \mathbb{N}^n \to \mathbb{R}^{n\times m_{\mathrm{d}}}$, and $\Omega_{\mathrm{d}u_{\mathrm{d}}u_{\mathrm{d}}} : \mathbb{N}^n \to \mathbb{N}^{m_{\mathrm{d}}}$, such that

$$\begin{aligned} x^{\mathrm{T}}(\Delta A_{\mathrm{c}}^{\mathrm{T}}E_{\mathrm{c}}^{\mathrm{T}}P + P\Delta A_{\mathrm{c}}E_{\mathrm{c}})x &\leq x^{\mathrm{T}}E_{\mathrm{c}}^{\mathrm{T}}\Omega_{\mathrm{c}}(P)E_{\mathrm{c}}x, \\ x \notin \mathcal{Z}, \quad \Delta A_{\mathrm{c}} &\in \boldsymbol{\Delta}_{A\mathrm{c}}, \qquad \text{(II.32)} \end{aligned}$$

$$\begin{aligned} &x^{\mathrm{T}}(\Delta A_{\mathrm{d}}^{\mathrm{T}}PA_{\mathrm{d}} + A_{\mathrm{d}}^{\mathrm{T}}P\Delta A_{\mathrm{d}} - \Delta A_{\mathrm{d}}PB_{\mathrm{d}}(R_{\mathrm{d}2} \\ &\quad +B_{\mathrm{d}}^{\mathrm{T}}PB_{\mathrm{d}} + \Omega_{\mathrm{d}u_{\mathrm{d}}u_{\mathrm{d}}}(P))^{-1} \\ &\quad (B_{\mathrm{d}}^{\mathrm{T}}PA_{\mathrm{d}} + E_{\mathrm{d}}^{\mathrm{T}}\Omega_{\mathrm{d}xu_{\mathrm{d}}}^{\mathrm{T}}(P)E_{\mathrm{d}}) - (B_{\mathrm{d}}^{\mathrm{T}}PA_{\mathrm{d}} \\ &\quad +\Omega_{\mathrm{d}xu_{\mathrm{d}}}^{\mathrm{T}}(P))^{\mathrm{T}}(R_{\mathrm{d}2} + B_{\mathrm{d}}^{\mathrm{T}}PB_{\mathrm{d}} + \Omega_{\mathrm{d}u_{\mathrm{d}}u_{\mathrm{d}}}(P))^{-1} \\ &\quad B_{\mathrm{d}}^{\mathrm{T}}P\Delta A_{\mathrm{d}} + \Delta A_{\mathrm{d}}^{\mathrm{T}}P\Delta A_{\mathrm{d}})x \leq x^{\mathrm{T}}(E_{\mathrm{d}}^{\mathrm{T}}\Omega_{\mathrm{d}xx}(P)E_{\mathrm{d}}^{\mathrm{T}} \\ &\quad -\Omega_{\mathrm{d}xu_{\mathrm{d}}}(P) \\ &\quad (R_{\mathrm{d}2}+B_{\mathrm{d}}^{\mathrm{T}}PB_{\mathrm{d}} + \Omega_{\mathrm{d}u_{\mathrm{d}}u_{\mathrm{d}}}(P))^{-1} \\ &\quad (B_{\mathrm{d}}^{\mathrm{T}}PA_{\mathrm{d}}+\Omega_{\mathrm{d}xu_{\mathrm{d}}}^{\mathrm{T}}(P)) \\ &\quad -(B_{\mathrm{d}}^{\mathrm{T}}P+\Omega_{\mathrm{d}xu_{\mathrm{d}}}^{\mathrm{T}}(P))^{\mathrm{T}}(R_{\mathrm{d}2} + B_{\mathrm{d}}^{\mathrm{T}}PB_{\mathrm{d}} + \Omega_{\mathrm{d}u_{\mathrm{d}}u_{\mathrm{d}}}(P))^{-1} \\ &\quad \Omega_{\mathrm{d}xu_{\mathrm{d}}}^{\mathrm{T}}(P) \\ &\quad +(B_{\mathrm{d}}^{\mathrm{T}}PA_{\mathrm{d}} + \Omega_{\mathrm{d}xu_{\mathrm{d}}}^{\mathrm{T}}(P))^{\mathrm{T}}(R_{\mathrm{d}2} + B_{\mathrm{d}}^{\mathrm{T}}PB_{\mathrm{d}}+ \\ &\quad \Omega_{\mathrm{d}u_{\mathrm{d}}u_{\mathrm{d}}}(P))^{-1} \\ &\quad \cdot\Omega_{\mathrm{d}u_{\mathrm{d}}u_{\mathrm{d}}}(P)(R_{\mathrm{d}2} + B_{\mathrm{d}}^{\mathrm{T}}PB_{\mathrm{d}} \\ &\quad +\Omega_{\mathrm{d}u_{\mathrm{d}}u_{\mathrm{d}}}(P))^{-1}(B_{\mathrm{d}}^{\mathrm{T}}PA_{\mathrm{d}} + \Omega_{\mathrm{d}xu_{\mathrm{d}}}^{\mathrm{T}}(P)))x, \\ &\quad x \in \mathcal{Z}, \\ &\quad \Delta A_{\mathrm{d}} \in \boldsymbol{\Delta}_{A\mathrm{d}}. \qquad \text{(II.33)} \end{aligned}$$

Furthermore, suppose there exists $P \in \mathbb{P}^n$ satisfying

$$\begin{aligned} 0 &= x^{\mathrm{T}}(A_{\mathrm{c}}^{\mathrm{T}}PE_{\mathrm{c}} + E_{\mathrm{c}}^{\mathrm{T}}PA_{\mathrm{c}} + E_{\mathrm{c}}^{\mathrm{T}}R_{\mathrm{c}1}E_{\mathrm{c}} + \Omega_{\mathrm{c}}(P) - \\ &\quad PB_{\mathrm{c}}R_{\mathrm{c}2}^{-1}B_{\mathrm{c}}^{\mathrm{T}}P)E_{\mathrm{c}}x, \qquad x \notin \mathcal{Z}, && \text{(II.34)} \\ 0 &< R_{\mathrm{d}2} + B_{\mathrm{d}}^{\mathrm{T}}PB_{\mathrm{d}} + \Omega_{\mathrm{d}u_{\mathrm{d}}u_{\mathrm{d}}}(P), && \text{(II.35)} \\ 0 &= x^{\mathrm{T}}(A_{\mathrm{d}}^{\mathrm{T}}PA - E_{\mathrm{d}}^{\mathrm{T}}PE_{\mathrm{d}} + E_{\mathrm{d}}^{\mathrm{T}}R_{\mathrm{d}1}E_{\mathrm{d}} \\ &\quad +\Omega_{\mathrm{d}xx}(P) - (B_{\mathrm{d}}^{\mathrm{T}}PA_{\mathrm{d}} + \Omega_{\mathrm{d}xu_{\mathrm{d}}}^{\mathrm{T}}(P))^{\mathrm{T}} \\ &\quad (R_{\mathrm{d}2} + B_{\mathrm{d}}^{\mathrm{T}}PB_{\mathrm{d}} + \Omega_{\mathrm{d}u_{\mathrm{d}}u_{\mathrm{d}}}(P))^{-1} \\ &\quad \cdot(B_{\mathrm{d}}^{\mathrm{T}}PA_{\mathrm{d}} + \Omega_{\mathrm{d}xu_{\mathrm{d}}}^{\mathrm{T}}(P)))x, \\ &\qquad x \in \mathcal{Z}. && \text{(II.36)} \end{aligned}$$

Then, with hybrid feedback control $(u_{\rm c}, u_{\rm d}) = (\phi_{\rm c}(x), \phi_{\rm d}(x)) = (-R_{{\rm c}2}^{-1} B_{\rm c}^{\rm T} P E_{\rm c} x,\ -(R_{{\rm d}2} + B_{\rm d}^{\rm T} P B_{\rm d} + \Omega_{{\rm d}u_{\rm d}u_{\rm d}}(P))^{-1}(B_{\rm d}^{\rm T} P A_{\rm d} + \Omega_{{\rm d}xu_{\rm d}}^{\rm T}(P))x)$ the zero solution $x(t) \equiv 0$ to (II.29), (II.30) is globally asymptotically stable for all $x_0 \in \mathbb{R}^n$, $(\Delta A_{\rm c}, \Delta A_{\rm d}) \in \boldsymbol{\Delta}_{A{\rm c}} \times \boldsymbol{\Delta}_{A{\rm d}}$ and

$$\sup_{(\Delta_{\rm c},\Delta_{\rm d})\in\boldsymbol{\Delta}} J_{(\Delta A_{\rm c},\Delta A_{\rm d})}(E_{\rm c}x_0) \leq \mathcal{J}(E_{\rm c}x_0, \phi_{\rm c}(\cdot), \phi_{\rm d}(\cdot)) = x_0^{\rm T} E_{\rm c}^{\rm T} P E_{\rm c} x_0, \quad x_0 \in \mathbb{R}^n, \tag{II.37}$$

where

$$\begin{aligned}\mathcal{J}(E_{\rm c}x_0, u_{\rm c}(\cdot), u_{\rm d}(\cdot)) &\triangleq \int_0^\infty [x^{\rm T}(t)E_{\rm c}^{\rm T}R_{{\rm c}1}E_{\rm c}x(t) \\ &+u_{\rm c}^{\rm T}(t)R_{{\rm c}2}u_{\rm c}(t) + x^{\rm T}(t)\Omega_{\rm c}(P)x(t)]{\rm d}t \\ &+ \sum_{k\in\mathcal{N}_{[0,\infty)}} [x^{\rm T}(t_k)E_{\rm d}^{\rm T}R_{{\rm d}1}E_{\rm d}x(t_k) \\ &+u_{\rm d}^{\rm T}(t_k)R_{{\rm d}2}u_{\rm d}(t_k) + x^{\rm T}(t_k)\Omega_{{\rm d}xx}(P)x(t_k) \\ &+2x^{\rm T}(t_k)\Omega_{{\rm d}xu_{\rm d}}(P)u_{\rm d}(t_k) + u_{\rm d}^{\rm T}(t_k) \\ &\cdot\Omega_{{\rm d}u_{\rm d}u_{\rm d}}(P)u_{\rm d}(t_k)],\end{aligned} \tag{II.38}$$

(II.39)

and where $(u_{\rm c}, u_{\rm d})$ is admissible and $x(t)$, $t \geq 0$, is a solution to (II.29), (II.30) with $(\Delta A_{\rm c}, \Delta A_{\rm d}) = (0, 0)$. Furthermore,

$$\mathcal{J}(E_{\rm c}x_0, \phi_{\rm c}(x(\cdot)), \phi_{\rm d}(x(\cdot))) = min_{(u_{\rm c}(\cdot),u_{\rm d}(\cdot))\in\mathcal{C}(x_0)}\mathcal{J}(E_{\rm c}x_0, u_{\rm c}(\cdot), u_{\rm d}(\cdot)), \tag{II.40}$$

where $\mathcal{C}(x_0)$ is the set of asymptotically stabilizing hybrid controllers for the nominal singularly impulsive dynamical system and $x_0 \in \mathbb{R}^n$, [3] and [7].

*Proof:* The detailed proof is given in [7]. □

## III. Inverse Optimal Robust Control for Nonlinear Affine Uncertain Singularly Impulsive Dynamical Systems

In this section we specialize Theorem II to affine uncertain systems. The controllers obtained are predicated on an *inverse optimal hybrid control problem*. In particular, to avoid the complexity in solving the steady-state hybrid Hamilton-Jacobi-Bellman equation we do not attempt to minimize a *given* hybrid cost functional, but rather, we parametrize a family of stabilizing hybrid controllers that minimize some *derived* hybrid cost functional that provides flexibility in specifying the control law. The performance integrand is shown to explicitly depend on the nonlinear singularly impulsive system dynamics, the Lyapunov function of the closed-loop system, and the stabilizing hybrid feedback control law wherein the coupling is introduced via the hybrid Hamilton-Jacobi-Bellman equation. Hence, by varying the parameters in the Lyapunov function and the performance integrand, the proposed framework can be used to characterize a class of globally stabilizing hybrid controllers that can meet the closed-loop system response constraints.

Consider the state-dependent affine (in the control) uncertain singularly impulsive dynamical system

$$\begin{aligned}E_{\rm c}\dot{x}(t) &= f_{\rm c}(x(t)) + \Delta f_{\rm c}(x(t)) + G_{\rm c}(x(t))u_{\rm c}(t), \\ &\qquad x(0) = x_0,\ x(t) \notin \mathcal{Z}_x,\end{aligned} \tag{III.41}$$

$$\begin{aligned}E_{\rm d}\Delta x(t) &= f_{\rm d}(x(t)) + \Delta f_{\rm d}(x(t)) + G_{\rm d}(x(t))u_{\rm d}(t), \\ &\qquad x(t) \in \mathcal{Z}_x,\end{aligned} \tag{III.42}$$

where $t \geq 0$, $f_{{\rm c}0}, f_{{\rm d}0} : \mathcal{D} \to \mathbb{R}^n$ and satisfies $f_{{\rm c}0}(0) = 0, f_{{\rm d}0}(0) = 0$, $\mathcal{D} = \mathbb{R}^n$, $\mathcal{U}_{\rm c} = \mathcal{C}_{\rm c} = \mathbb{R}^{m_{\rm c}}$, $\mathcal{U}_{\rm d} = \mathcal{C}_{\rm d} = \mathbb{R}^{m_{\rm d}}$, and $(\Delta f_{\rm c}, \Delta f_{\rm d}) \in \mathcal{F}_{\rm c} \times \mathcal{F}_{\rm d} \triangleq \mathcal{F}$, where

$$\Delta f_{\rm c}(\cdot) \in \mathcal{F}_{\rm c} \subset \{\Delta f_{\rm c} : \mathbb{R}^n \to \mathbb{R}^n : \Delta f_{\rm c}(0) = 0\},$$
$$\Delta f_{\rm d}(\cdot) \in \mathcal{F}_{\rm d} \subset \{\Delta f_{\rm d} : \mathbb{R}^n \to \mathbb{R}^n : \Delta f_{\rm d}(0) = 0\}.$$

In this section no explicit structure is assumed for the elements of $\mathcal{F}$. Furthermore, we consider performance integrands $L_{\rm c}(E_{\rm c}x, u_{\rm c})$ and $L_{\rm d}(E_{\rm d}x, u_{\rm d})$ of the form

$$L_{\rm c}(E_{\rm c}x, u_{\rm c}) = L_{{\rm c}1}(E_{\rm c}x) + u_{\rm c}^{\rm T}R_{{\rm c}2}(x)u_{\rm c},\ x \notin \mathcal{Z}, \tag{III.43}$$
$$L_{\rm d}(E_{\rm d}x, u_{\rm d}) = L_{{\rm d}1}(E_{\rm d}x) + u_{\rm d}^{\rm T}R_{{\rm d}2}(x)u_{\rm d}, \quad x \in \mathcal{Z}, \tag{III.44}$$

where $L_{{\rm c}1} : \mathbb{R}^n \to \mathbb{R}$ and satisfies $L_{{\rm c}1}(E_{\rm c}x) \geq 0$, $x \in \mathbb{R}^n$, $R_{{\rm c}2} : \mathbb{R}^n \to \mathbb{P}^{m_{\rm c}}$, $L_{{\rm d}1} : \mathbb{R}^n \to \mathbb{R}$ and satisfies $L_{{\rm d}1}(E_{\rm d}x) \geq 0$, $x \in \mathbb{R}^n$, and $R_{{\rm d}2} : \mathbb{R}^n \to \mathbb{P}^{m_{\rm d}}$ so that (II.5) becomes

$$\begin{aligned}J(E_{\rm c}x_0, u_{\rm c}(\cdot), u_{\rm d}(\cdot)) &= \int_0^\infty [L_{{\rm c}1}(E_{\rm c}x(t)) + \\ &u_{\rm c}^{\rm T}(t)R_{{\rm c}2}(x(t))u_{\rm c}(t)]{\rm d}t + \sum_{k\in\mathcal{N}_{[0,\infty)}} [L_{{\rm d}1}(E_{\rm d}x(t_k)) \\ &+u_{\rm d}^{\rm T}(t_k)R_{{\rm d}2}(x(t_k))u_{\rm d}(t_k)].\end{aligned} \tag{III.45}$$

Consider the nonlinear uncertain controlled affine singularly impulsive system (III.41), (III.42) with performance functional (III.45). Assume there exists a $C^1$ function $V : \mathbb{R}^n \to \mathbb{R}$, and functions $P_{12} : \mathbb{R}^n \to \mathbb{R}^{1\times m_{\rm d}}$, $P_2 : \mathbb{R}^n \to \mathbb{N}^{m_{\rm d}}$, $P_{1f_{\rm d}} : \mathbb{R}^n \to \mathbb{R}^{1\times n}$, $P_{2f_{\rm d}} : \mathbb{R}^n \to \mathbb{N}^{n\times n}$, $P_{u_{\rm d}f_{\rm d}} : \mathbb{R}^n \to \mathbb{R}^{m_{\rm d}\times n}$, $\Gamma_{\rm c} : \mathbb{R}^n \to \mathbb{R}$, $\Gamma_{{\rm d}xx} : \mathbb{R}^n \to \mathbb{R}^n$, $\Gamma_{{\rm d}xu_{\rm d}} : \mathbb{R}^n \to \mathbb{R}^{1\times m_{\rm d}}$, and $\Gamma_{{\rm d}u_{\rm d}u_{\rm d}} : \mathbb{R}^n \to \mathbb{N}^{m_{\rm d}}$ such that

$$P_{12}(0) = 0, \tag{III.46}$$
$$P_{1f_{\rm d}}(0) = 0, \tag{III.47}$$
$$\Gamma_{{\rm d}xu_{\rm d}}(0) = 0, \tag{III.48}$$
$$V(0) = 0, \tag{III.49}$$
$$V(E_{\rm c}x) \geq 0, \quad x \in \mathbb{R}^n, \quad x \neq 0, \tag{III.50}$$

$$\begin{aligned}V'(E_{\rm c}x)\Delta f_{\rm c}(x) \leq \Gamma_{\rm c}(x),\ x \notin \mathcal{Z}_x, \\ \Delta f_{\rm c} \in \mathcal{F}_{\rm c},\end{aligned} \tag{III.51}$$

$$\begin{aligned}V'(E_{\rm c}x)[f_{{\rm c}0}(x) - \tfrac{1}{2}G_{\rm c}(x)R_{{\rm c}2}^{-1}(x)G_{\rm c}^{\rm T}(x)V'^{\rm T}(E_{\rm c}x)] \\ +\Gamma_{\rm c}(x) < 0, \\ x \notin \mathcal{Z}_x, \quad x \neq 0,\end{aligned} \tag{III.52}$$

$$\begin{aligned}P_{1f_{\rm d}}(x)\Delta f_{\rm d}(x) + \Delta f_{\rm d}^{\rm T}(x)P_{1f_{\rm d}}^{\rm T}(x) \\ +\Delta f_{\rm d}^{\rm T}(x)P_{2f_{\rm d}}(x)\Delta f_{\rm d}(x) + \phi_{\rm d}^{\rm T}(E_{\rm d}x)P_{u_{\rm d}f_{\rm d}}(x)\Delta f_{\rm d}(x) \\ +\Delta f_{\rm d}^{\rm T}(x)P_{u_{\rm d}f_{\rm d}}^{\rm T}(x)\phi_{\rm d}(x) \\ \leq \Gamma_{{\rm d}xx}(x) + \Gamma_{{\rm d}xu_{\rm d}}(x)\phi_{\rm d}(x) + \phi_{\rm d}^{\rm T}(x)\Gamma_{{\rm d}u_{\rm d}u_{\rm d}}(x)\phi_{\rm d}(x), \\ x \in \mathcal{Z}_x, \Delta f_{\rm d}(\cdot) \in \mathcal{F}_{\rm d},\end{aligned} \tag{III.53}$$

$$V(E_{\rm d}x+f_{\rm d0}(x))-V(E_{\rm d}x)+P_{12}(x)\phi_{\rm d}(x)+$$
$$\phi_{\rm d}(x)^{\rm T}P_2(x)\phi_{\rm d}(x)$$
$$+\Gamma_{{\rm d}xx}(x)+\Gamma_{{\rm d}xu_{\rm d}}(x)\phi_{\rm d}(x)+$$
$$\phi_{\rm d}^{\rm T}(x)\Gamma_{{\rm d}u_{\rm d}u_{\rm d}}(x)\phi_{\rm d}(x)\leq 0,\quad x\in\mathcal{Z}_x,$$

$$V(E_{\rm d}x+f_{\rm d0}(x)+G_{\rm d}(x)u_{\rm d})=V(E_{\rm d}x+f_{\rm d0}(x))$$
$$+P_{12}(x)u_{\rm d}+u_{\rm d}^{\rm T}P_2(x)u_{\rm d},$$

$$V(E_{\rm d}x+f_{\rm d0}(x)+\Delta f_{\rm d}(x)+G_{\rm d}(x)u_{\rm d})-V(E_{\rm d}x)=$$
$$V(E_{\rm d}x+f_{\rm d0}(x)+G_{\rm d}(x)u_{\rm d})-V(E_{\rm d}x)+P_{1f_{\rm d}}(x)\Delta f_{\rm d}(x)+$$
$$\Delta f_{\rm d}^{\rm T}(x)P_{1f_{\rm d}}^{\rm T}(x)+\Delta f_{\rm d}^{\rm T}(x)$$
$$\cdot P_{2f_{\rm d}}(x)\Delta f_{\rm d}(x)+u_{\rm d}^{\rm T}P_{u_{\rm d}f_{\rm d}}(x)\Delta f_{\rm d}(x)$$
$$+\Delta f_{\rm d}^{\rm T}(x)P_{u_{\rm d}f_{\rm d}}^{\rm T}(x)u_{\rm d},$$
$$x\in\mathcal{Z}_x,\quad u_{\rm d}\in\mathbb{R}^{m_{\rm d}},\Delta f_{\rm d}(\cdot)\in\mathcal{F}_{\rm d},$$

and

$$V(E_{\rm c/d}x)\to\infty \text{ as } \|x\|\to\infty. \tag{III.61}$$

Then the zero solution $x(t)\equiv 0$ to the closed-loop system

$$\begin{aligned}E_{\rm c}\dot{x}(t) &= f_{\rm c}(x(t))+\Delta f_{\rm c}(x(t))+G_{\rm c}(x(t))\phi_{\rm c}(x(t)),\\ &\qquad x(0)=x_0,\ x(t)\notin\mathcal{Z}_x, \qquad\qquad (\text{III.62})\\ E_{\rm d}\Delta x(t) &= f_{\rm d}(x(t))+\Delta f_{\rm d}(x(t))+G_{\rm d}(x(t))\phi_{\rm d}(x(t)),\\ &\qquad x(t)\in\mathcal{Z}_x, \qquad\qquad (\text{III.63})\end{aligned}$$

is globally asymptotically stable for all $(\Delta f_{\rm c},\Delta f_{\rm d})\in\mathcal{F}$ with the hybrid feedback control law

$$\begin{aligned}\phi_{\rm c}(x) &= -\tfrac{1}{2}R_{\rm c2}^{-1}(x)G_{\rm c}^{\rm T}(x)V'^{\rm T}(E_{\rm c}x),\quad x\notin\mathcal{Z}_x, \qquad (\text{III.64})\\ \phi_{\rm d}(x) &= -\tfrac{1}{2}(R_{\rm d2}(x)+P_2(x)+\Gamma_{{\rm d}u_{\rm d}u_{\rm d}}(x))^{-1}\\ &\quad\cdot(P_{12}+\Gamma_{{\rm d}xu_{\rm d}(x)})^{\rm T}(x),\ x\in\mathcal{Z}_x, \qquad (\text{III.65})\end{aligned}$$

and performance functional (III.45), satisfies

$$J(E_{\rm c}x_0,\phi_{\rm c}(x(\cdot)),\phi_{\rm d}(x(\cdot)))\ =\ \min_{(u_{\rm c}(\cdot),u_{\rm d}(\cdot))\in\mathcal{C}(x_0)} J(E_{\rm c}x_0,u_{\rm c}(\cdot),u_{\rm d}(\cdot)),\ x_0\in\mathbb{R}^n, \tag{III.66}$$

where

$$\begin{aligned}\mathcal{J}(E_{\rm c}x_0,u_{\rm c}(\cdot),u_{\rm d}(\cdot)) &\triangleq \int_0^\infty [L_{\rm c}(E_{\rm c}x(t),u_{\rm c}(t))\\ &+\Gamma_{\rm c}(\tilde{x}(t),u_{\rm c}(t))]{\rm d}t+\sum_{k\in\mathcal{N}_{[0,\infty)}}[L_{\rm d}(x(t_k),u_{\rm d}(t_k))+\\ &\Gamma_{\rm d}(x(t_k),u_{\rm d}(t_k))],\end{aligned} \tag{III.67}$$

and

$$\begin{aligned}\Gamma_{\rm c}(x,u_{\rm c}) &= \Gamma_{{\rm c}xx}(x),\quad x\notin\mathcal{Z}_x, \qquad (\text{III.68})\\ \Gamma_{\rm d}(x,u_{\rm d}) &= \Gamma_{{\rm d}xx}(x)+\Gamma_{{\rm d}xu_{\rm d}}(x)u_{\rm d}+u_{\rm d}^{\rm T}\Gamma_{{\rm d}u_{\rm d}u_{\rm d}}(x)u_{\rm d},\\ &\qquad x\in\mathcal{Z}_x, \qquad (\text{III.69})\end{aligned}$$

and where $(u_{\rm c}(\cdot),u_{\rm d}(\cdot))$ is an admissible control and $x(t)$, $t\geq 0$, is a solution of (III.41), (III.42) with $(\Delta f_{\rm c},\Delta f_{\rm d})=(0,0)$.

(III.54)In addition, the hybrid performance functional (III.67), with

$$\begin{aligned}L_{\rm c1}(E_{\rm c}x) &= \phi_{\rm c}^{\rm T}(x)R_{\rm c2}(x)\phi_{\rm c}(x)-V'(E_{\rm c}x)f_{\rm c0}(x)\\ &\quad-\Gamma_{{\rm c}xx}(x), \qquad (\text{III.70})\\ L_{\rm d1}(E_{\rm d}x) &= \phi_{\rm d}^{\rm T}(x)(R_{\rm d2}(x)+P_2(x)+\Gamma_{{\rm d}u_{\rm d}u_{\rm d}}(x))\phi_{\rm d}(x)\\ &\quad-V(E_{\rm d}x+f_{\rm d0}(x))+V(E_{\rm d}x)-\Gamma_{{\rm d}xx}(x), \qquad (\text{III.71})\end{aligned}$$

(III.55)

(III.56)

(III.57)is minimized in the sense that

(III.58)

$$\mathcal{J}(E_{\rm c}x_0,\phi_{\rm c}(E_{\rm c}x(\cdot)),\phi_{\rm d}(E_{\rm d}x(\cdot)))\ =\ \min_{(u_{\rm c}(\cdot),u_{\rm d}(\cdot))\in\mathcal{C}(x_0)}\mathcal{J}(E_{\rm c}x_0,u_{\rm c}(\cdot),u_{\rm d}(\cdot)). \tag{III.72}$$

[3] and [7].

*Proof:* The result is a direct consequence of Theorem II with $\mathcal{D}=\mathbb{R}^n$, $\mathcal{U}_{\rm c}=\mathbb{R}^{m_{\rm c}}$, $\mathcal{U}_{\rm d}=\mathbb{R}^{m_{\rm d}}$, $F_{\rm c}(x,u_{\rm c})=f_{\rm c0}(x)+\Delta f_{\rm c}(x)+G_{\rm c}(x)u_{\rm c}$, $F_{\rm c0}(x,u_{\rm c})=f_{\rm c0}(x)+G_{\rm c}(x)u_{\rm c}$, $L_{\rm c}(E_{\rm c}x,u_{\rm c})$ (III.59) given by (III.43), $\Gamma_{\rm c}(x,u_{\rm c})$ given by (III.68), for (III.60) $x\notin\mathcal{Z}_x$, $F_{\rm d}(x,u_{\rm d})=f_{\rm d0}(x)+\Delta f_{\rm d}(x)+G_{\rm d}(x)u_{\rm d}$, $F_{\rm d}(x,u_{\rm d})=f_{\rm d0}(x)+G_{\rm d}(x)u_{\rm d}$, $L_{\rm d}(E_{\rm d}x,u_{\rm d})$ given by (III.44), $\Gamma_{\rm d}(x,u_{\rm d})$ given by (III.69), for $x\in\mathcal{Z}$. Specifically, with (III.41)–(III.44), (III.68), and (III.69), the Hamiltonian have the form

$$\begin{aligned}H_{\rm c}(E_{\rm c}x,u_{\rm c}) &= L_{\rm c1}(E_{\rm c}x)+u_{\rm c}^{\rm T}R_{\rm c2}(x)u_{\rm c}\\ &\quad+V'(E_{\rm c}x)(f_{\rm c0}(x)+G_{\rm c}(x)u_{\rm c})+\Gamma_{{\rm c}xx}(x),\\ &\quad x\notin\mathcal{Z}_x,\ u_{\rm c}\in\mathcal{U}_{\rm c}, \qquad (\text{III.73})\\ H_{\rm d}(E_{\rm d}x,u_{\rm d}) &= L_{\rm d1}(E_{\rm d}x)+u_{\rm d}^{\rm T}R_{\rm d2}(x)u_{\rm d}\\ &\quad+V(E_{\rm d}x+f_{\rm d0}(x)+G_{\rm d}(x)u_{\rm d})-V(E_{\rm d}x)\\ &\quad+\Gamma_{{\rm d}xx}(x)+\Gamma_{{\rm d}xu_{\rm d}}(x)u_{\rm d}+u_{\rm d}^{\rm T}\Gamma_{{\rm d}u_{\rm d}u_{\rm d}}(x)u_{\rm d},\\ &\qquad x\in\mathcal{Z}_x,\quad u_{\rm d}\in\mathcal{U}_{\rm d}. \qquad (\text{III.74})\end{aligned}$$

Now, the hybrid feedback control law (III.64), (III.65) is obtained by setting $\frac{\partial H_{\rm c}}{\partial u_{\rm c}}=0$ and $\frac{\partial H_{\rm d}}{\partial u_{\rm d}}=0$. With (III.64) and (III.65) it follows that (III.51)–(III.60) imply (II.10)–(II.13). Next, since $V(\cdot)$ is C$^1$ and $x=0$ is a local minimum of $V(\cdot)$, it follows that $V'(0)=0$, and hence, since by assumption $P_{12}(0)=0$ and $\Gamma_{{\rm d}xu_{\rm d}}(0)=0$, it follows that $\phi_{\rm c}(0)=0$ and $\phi_{\rm d}(0)=0$ which proves (II.8), (II.9). Next, with $L_{\rm c1}(E_{\rm c}x)$ and $L_{\rm d1}(E_{\rm d}x)$ given by (III.70) and (III.71), respectively, and $\phi_{\rm c}(x)$, $\phi_{\rm d}(x)$ given by (III.64) and (III.65), (II.14) and (II.16) hold. Finally, since

$$\begin{aligned}H_{\rm c}(E_{\rm c}x,u_{\rm c}) &= H_{\rm c}(E_{\rm c}x,u_{\rm c})-H_{\rm c}(E_{\rm c}x,\phi_{\rm c}(x))\\ &= [u_{\rm c}-\phi_{\rm c}(x)]^{\rm T}R_{\rm c2}(x)[u_{\rm c}-\phi_{\rm c}(x)],\quad x\notin\mathcal{Z}_x, \qquad (\text{III.75})\\ H_{\rm d}(E_{\rm d}x,u_{\rm d}) &= H_{\rm d}(E_{\rm d}x,u_{\rm d})-H_{\rm d}(E_{\rm d}x,\phi_{\rm d}(x))\\ &= [u_{\rm d}-\phi_{\rm d}(x)]^{\rm T}(R_{\rm d2}(x)\\ &\quad+P_2(x+\Gamma_{{\rm d}u_{\rm d}u_{\rm d}}(x)))[u_{\rm d}-\phi_{\rm d}(x)],\\ &\qquad x\in\mathcal{Z}_x, \qquad (\text{III.76})\end{aligned}$$

where $R_{\rm c2}(x)>0$, $x\notin\mathcal{Z}_x$, and $R_{\rm d2}(x)+P_2(x)+\Gamma_{{\rm d}u_{\rm d}u_{\rm d}}(x)>0$, $x\in\mathcal{Z}_x$, conditions (II.15) and (II.17) hold. The result now follows as a direct consequence of Theorem II. □

## IV. Robust Nonlinear Hybrid Control with Polynomial Performance Functional

In this section we specialize the results of Section IV to linear singularly impulsive systems controlled by inverse optimal nonlinear hybrid controllers that minimize a derived polynomial cost functional. Specifically, assume $\mathcal{F} \triangleq \mathcal{F}_{\mathrm{c}} \times \mathcal{F}_{\mathrm{d}}$ to be the set of uncertain systems, where

$$\begin{aligned}\mathcal{F}_{\mathrm{c}} = \{&(A_{\mathrm{c}} + \Delta A_{\mathrm{c}})x + B_{\mathrm{c}}u_{\mathrm{c}} : x \in \mathbb{R}^n,\\ &A_{\mathrm{c}} \in \mathbb{R}^{n\times n}, B_{\mathrm{c}} \in \mathbb{R}^{n\times m_{\mathrm{c}}}, \Delta A_{\mathrm{c}} \in \mathbf{\Delta}_{A_{\mathrm{c}}}\},\end{aligned} \tag{IV.77}$$

$$\begin{aligned}\mathcal{F}_{\mathrm{d}} = \{&(A_{\mathrm{d}} + \Delta A_{\mathrm{d}})x : x \in \mathbb{R}^n,\\ &A_{\mathrm{d}} \in \mathbb{R}^{n\times n}, \Delta A_{\mathrm{d}} \in \mathbf{\Delta}_{A_{\mathrm{d}}}\},\end{aligned} \tag{IV.78}$$

where $\mathbf{\Delta}_{A_{\mathrm{c}}}, \mathbf{\Delta}_{A_{\mathrm{d}}} \subset \mathbb{R}^{n\times n}$ are given bounded uncertainty sets of uncertain perturbations $\Delta A_{\mathrm{c}}, \Delta A_{\mathrm{d}}$ of the nominal asymptotically stable system matrices $A_{\mathrm{c}}, A_{\mathrm{d}}$ such that $0 \in \mathbf{\Delta}_{A_{\mathrm{c}}}$ and $0 \in \mathbf{\Delta}_{A_{\mathrm{d}}}$. For simplicity of exposition, we also define $(\Delta A_{\mathrm{c}}, \Delta A_{\mathrm{d}}) \in \mathbf{\Delta}_{A\mathrm{c}} \times \mathbf{\Delta}_{A\mathrm{d}} \triangleq \mathbf{\Delta}$. For the results in this section we assume $u_{\mathrm{d}}(t) \equiv 0$. Furthermore, let $R_{1\mathrm{c}} \in \mathbb{P}^n$, $R_{1\mathrm{d}} \in \mathbb{N}^n$, $R_{2\mathrm{c}} \in \mathbb{P}^{m_{\mathrm{c}}}$, $\hat{R}_q, \hat{\hat{R}}_q \in \mathbb{N}^n$, $q = 2, \ldots, r$, be given, where $r$ is a positive integer, and define $S_{\mathrm{c}} \triangleq B_{\mathrm{c}}R_{2\mathrm{c}}^{-1}B_{\mathrm{c}}^{\mathrm{T}}$.

Consider the linear uncertain controlled singularly impulsive system

$$\begin{aligned}E_{\mathrm{c}}\dot{x}(t) &= (A_{\mathrm{c}} + \Delta A_{\mathrm{c}})x(t) + B_{\mathrm{c}}u_{\mathrm{c}}(t), \quad x(0) = x_0,\\ &\qquad x(t) \notin \mathcal{Z}_x,\end{aligned} \tag{IV.79}$$

$$E_{\mathrm{d}}\Delta x(t) = (A_{\mathrm{d}} + \Delta A_{\mathrm{d}} - E_{\mathrm{d}})x(t), \quad x(t) \in \mathcal{Z}_x, \tag{IV.80}$$

where $u_{\mathrm{c}}$ is admissible and $(\Delta A_{\mathrm{c}}, \Delta A_{\mathrm{d}}) \in \mathbf{\Delta}$. Let $\Omega_{\mathrm{c}}, \Omega_{\mathrm{d}} : \mathcal{N}_{\mathrm{P}} \subseteq \mathbb{S}^n \to \mathbb{N}^n$, $P \in \mathcal{N}_{\mathrm{P}}$, be such that

$$\begin{aligned}&x^{\mathrm{T}}(\Delta A_{\mathrm{c}}^{\mathrm{T}}PE_{\mathrm{c}} + E_{\mathrm{c}}^{\mathrm{T}}P\Delta A_{\mathrm{c}})x \le x^{\mathrm{T}}\Omega_{\mathrm{c}}(P)x,\\ &\qquad x \notin \mathcal{Z}, \quad \Delta A_{\mathrm{c}} \in \mathbf{\Delta}_{A\mathrm{c}},\end{aligned} \tag{IV.81}$$

$$\begin{aligned}&x^{\mathrm{T}}(\Delta A_{\mathrm{d}}^{\mathrm{T}}P\Delta A_{\mathrm{d}} + \Delta A_{\mathrm{d}}^{\mathrm{T}}PA_{\mathrm{d}} + \Delta A_{\mathrm{d}}^{\mathrm{T}}P\Delta A_{\mathrm{d}})x\\ &\le x^{\mathrm{T}}\Omega_{\mathrm{d}}(P)x, \qquad x \in \mathcal{Z}, \quad \Delta A_{\mathrm{d}} \in \mathbf{\Delta}_{A\mathrm{d}}.\end{aligned} \tag{IV.82}$$

Assume there exist $P \in \mathbb{P}^n$ and $M_q \in \mathbb{N}^n$, $q = 2, \ldots, r$, such that

$$\begin{aligned}0 &= x^{\mathrm{T}}(A_{\mathrm{c}}^{\mathrm{T}}PE_{\mathrm{c}} + E_{\mathrm{c}}^{\mathrm{T}}PA_{\mathrm{c}} + E_{\mathrm{c}}^{\mathrm{T}}R_{1\mathrm{c}}E_{\mathrm{c}} + \Omega_{\mathrm{c}}(P) - PS_{\mathrm{c}}P)x,\\ &\qquad x \notin \mathcal{Z}_x,\end{aligned} \tag{IV.83}$$

$$\begin{aligned}0 &= x^{\mathrm{T}}[(A_{\mathrm{c}} - S_{\mathrm{c}}P)^{\mathrm{T}}M_qE_{\mathrm{c}} + E_{\mathrm{c}}^{\mathrm{T}}M_q(A_{\mathrm{c}} - S_{\mathrm{c}}P) + \hat{R}_q]x,\\ &\qquad x \notin \mathcal{Z}_x, q = 2, \ldots, r,\end{aligned} \tag{IV.84}$$

$$\begin{aligned}0 &= x^{\mathrm{T}}(A_{\mathrm{d}}^{\mathrm{T}}PA_{\mathrm{d}} - E_{\mathrm{d}}^{\mathrm{T}}PE_{\mathrm{d}} + E_{\mathrm{d}}^{\mathrm{T}}R_{1\mathrm{d}}E_{\mathrm{d}} + \Omega_{\mathrm{d}}(P))x,\\ &\qquad x \in \mathcal{Z}_x,\end{aligned} \tag{IV.85}$$

$$\begin{aligned}0 &= x^{\mathrm{T}}(A_{\mathrm{d}}^{\mathrm{T}}M_qA_{\mathrm{d}} - E_{\mathrm{d}}^{\mathrm{T}}M_qE_{\mathrm{d}} + \hat{\hat{R}}_q)x,\\ &\qquad x \in \mathcal{Z}_x, q = 2, \ldots, r.\end{aligned} \tag{IV.86}$$

Then the zero solution $x(t) \equiv 0$ of the uncertain closed-loop system

$$\begin{aligned}E_{\mathrm{c}}\dot{x}(t) &= (A_{\mathrm{c}} + \Delta A_{\mathrm{c}})x(t) + B_{\mathrm{c}}\phi_{\mathrm{c}}(x(t)), \; x(0) = x_0,\\ &\qquad x(t) \notin \mathcal{Z}_x,\end{aligned} \tag{IV.87}$$

$$E_{\mathrm{d}}\Delta x(t) = (A_{\mathrm{d}} + \Delta A_{\mathrm{d}} - E_{\mathrm{d}})x(t), \qquad x(t) \in \mathcal{Z}_x, \tag{IV.88}$$

is globally asymptotically stable with the feedback control law

$$\begin{aligned}\phi_{\mathrm{c}}(x) &= -R_{2\mathrm{c}}^{-1}B_{\mathrm{c}}^{\mathrm{T}}(P + \sum_{q=2}^{r}(x^{\mathrm{T}}E_{\mathrm{c}}^{\mathrm{T}}M_qE_{\mathrm{c}}x)^{q-1}M_q)E_{\mathrm{c}}x,\\ &\qquad x \notin \mathcal{Z}_x,\end{aligned} \tag{IV.89}$$

and the performance functional (III.45) satisfies

$$\begin{aligned}&\sup_{(\Delta A_{\mathrm{c}},\Delta A_{\mathrm{d}})\in\mathbf{\Delta}} J_{\Delta A_{\mathrm{c}},\Delta A_{\mathrm{d}}}(E_{\mathrm{c}}x_0, \phi_{\mathrm{c}}(x_0))\\ &\qquad \le \mathcal{J}(E_{\mathrm{c}}x_0, \phi_{\mathrm{c}}(x_0))\\ &= x_0^{\mathrm{T}}E_{\mathrm{c}}^{\mathrm{T}}PE_{\mathrm{c}}x_0 + \sum_{q=2}^{r}\frac{1}{q}(x_0^{\mathrm{T}}E_{\mathrm{c}}^{\mathrm{T}}M_qE_{\mathrm{c}}x_0)^q, \; x_0 \in \mathbb{R}^n,\end{aligned} \tag{IV.90}$$

where

$$\begin{aligned}\mathcal{J}(E_{\mathrm{c}}x_0, u_{\mathrm{c}}(\cdot)) \triangleq &\int_0^{\infty}[L_{\mathrm{c}}(E_{\mathrm{c}}x(t), u_{\mathrm{c}}(t)) + \Gamma_{\mathrm{c}}(\tilde{x}(t), u_{\mathrm{c}}(t))]\mathrm{d}t\\ &+ \sum_{k\in\mathcal{N}_{[0,\infty)}}[L_{\mathrm{d}}(E_{\mathrm{d}}x(t_k)) + \Gamma_{\mathrm{d}}(x(t_k))],\end{aligned} \tag{IV.91}$$

and where $u_{\mathrm{c}}$ is admissible, and $x(t)$, $t \ge 0$, is a solution to (IV.79), (IV.80) with $(\Delta A_{\mathrm{c}}, \Delta A_{\mathrm{d}}) = (0, 0)$, and

$$\begin{aligned}\Gamma_{\mathrm{c}}(x, u_{\mathrm{c}}) &= x^{\mathrm{T}}(\Omega_{\mathrm{c}}(P) + \sum_{q=2}^{r}(x^{\mathrm{T}}E_{\mathrm{c}}^{\mathrm{T}}M_qE_{\mathrm{c}}x)^{q-1}\Omega_{\mathrm{c}}(M_q))E_{\mathrm{c}}x,\\ &\qquad x \notin \mathcal{Z}_x\end{aligned} \tag{IV.92}$$

$$\begin{aligned}\Gamma_{\mathrm{d}}(x) &= x^{\mathrm{T}}\Omega_{\mathrm{d}}(P)x + \sum_{q=2}^{r}\frac{1}{q}[(x^{\mathrm{T}}\hat{\hat{R}}_qx)\sum_{j=1}^{q}(x^{\mathrm{T}}E_{\mathrm{d}}^{\mathrm{T}}M_qE_{\mathrm{d}}x)^{j-1}\\ &\quad \cdot(x^{\mathrm{T}}(A_{\mathrm{d}}^{\mathrm{T}}M_qA_{\mathrm{d}} + \Omega_{\mathrm{d}}(x))x)^{q-j}\\ &\quad -(x^{\mathrm{T}}A_{\mathrm{d}}^{\mathrm{T}}M_qA_{\mathrm{d}}x)^{q-j}], \quad x \in \mathcal{Z}_x,\end{aligned} \tag{IV.93}$$

where $u_{\mathrm{c}}$ is admissible and $(\Delta A_{\mathrm{c}}, \Delta A_{\mathrm{d}}) \in \mathbf{\Delta}$. In addition, the performance functional (III.45), with $R_{2\mathrm{c}}(x) = R_{2\mathrm{c}}$ and

$$\begin{aligned}L_{1\mathrm{c}}(E_{\mathrm{c}}x) &= x^{\mathrm{T}}(E_{\mathrm{c}}^{\mathrm{T}}R_{1\mathrm{c}}E_{\mathrm{c}} + \sum_{q=2}^{r}(x^{\mathrm{T}}E_{\mathrm{c}}^{\mathrm{T}}M_qE_{\mathrm{c}}x)^{q-1}\hat{R}_q\\ &+[\sum_{q=2}^{r}(x^{\mathrm{T}}E_{\mathrm{c}}^{\mathrm{T}}M_qE_{\mathrm{c}}x)^{q-1}M_q]^{\mathrm{T}}S_{\mathrm{c}}\\ &\qquad [\sum_{q=2}^{r}(x^{\mathrm{T}}E_{\mathrm{c}}^{\mathrm{T}}M_qE_{\mathrm{c}}x)^{q-1}M_q])x,\end{aligned} \tag{IV.94}$$

$$\begin{aligned}L_{1\mathrm{d}}(E_{\mathrm{d}}x) &= x^{\mathrm{T}}E_{\mathrm{d}}^{\mathrm{T}}R_{1\mathrm{d}}E_{\mathrm{d}}x + \sum_{q=2}^{r}\frac{1}{q}[(x^{\mathrm{T}}\hat{\hat{R}}_qx)\\ &\qquad \sum_{j=1}^{q}(x^{\mathrm{T}}E_{\mathrm{d}}^{\mathrm{T}}M_qE_{\mathrm{d}}x)^{j-1}\\ &\cdot(x^{\mathrm{T}}A_{\mathrm{d}}^{\mathrm{T}}M_qA_{\mathrm{d}}x)^{q-j}],\end{aligned} \tag{IV.95}$$

is minimized in the sense that J(E$_{\mathrm{c}}x_0, \phi_{\mathrm{c}}(x(\cdot))) = \min_{u_{\mathrm{c}}(\cdot)\in\mathcal{C}(x_0)} J(E_{\mathrm{c}}x_0, u_{\mathrm{c}}(\cdot)), \quad x_0 \in \mathbb{R}^n, (IV.95)$where $\mathcal{C}(x_0)$ is the set of asymptotically stabilizing controllers for the nominal system and $x_0 \in \mathbb{R}^n$, [3] and [7].

*Proof:* The result is a direct consequence of Corollary III. □

## V. Robust Nonlinear Hybrid Control with Multilinear Performance Functional

Finally, we specialize the results of Section VI to linear singularly impulsive systems controlled by inverse optimal hybrid controllers that minimize a derived multilinear functional. First, however, we give several definitions involving multilinear forms. A scalar function $\psi : \mathbb{R}^n \to \mathbb{R}$ is $q$-*multilinear* if $q$ is a positive integer and $\psi(x)$ is a linear combination of terms of the form $x_1^{i_1}x_2^{i_2}\ldots x_n^{i_n}$, where $i_j$ is a nonnegative integer for $j = 1, \ldots, n$, and $i_1 + i_2 + \ldots + i_n = q$. Furthermore, a $q$-multilinear function $\psi(\cdot)$ is nonnegative definite (resp., positive definite) if $\psi(x) \geq 0$ for all $x \in \mathbb{R}^n$ (resp., $\psi(x) > 0$ for all nonzero $x \in \mathbb{R}^n$). Note that if $q$ is odd then $\psi(x)$ cannot be positive definite. If $\psi(\cdot)$ is a $q$-multilinear function then $\psi(\cdot)$ can be represented by means of Kronecker products, that is, $\psi(x)$ is given by $\psi(x) = \Psi x^{[q]}$, where $\Psi \in \mathbb{R}^{1\times n^q}$ and $x^{[q]} \triangleq x \otimes x \otimes \cdots \times x$ ($q$ times), where $\otimes$ denotes Kronecker product. For the next result recall the definition of $S_{\mathrm{c}}$, let $R_{1\mathrm{c}} \in \mathbb{P}^n$, $R_{1\mathrm{d}} \in \mathbb{P}^n$, $R_{2\mathrm{c}} \in \mathbb{P}^{m_{\mathrm{c}}}$, $\hat{R}_{2q}, \hat{\hat{R}}_{2q} \in \mathcal{N}^{(2q,n)}$, $q = 2, \ldots, r$, be given, where $\mathcal{N}^{(2q,n)} \triangleq \{\Psi \in \mathbb{R}^{1\times n^{2q}} : \Psi x^{[2q]} \geq 0,\ x \in \mathbb{R}^n\}$, and define the repeated ($q$ times) Kronecker sum as $\overset{q}{\oplus} A \triangleq A \oplus A \oplus \cdots \oplus A$.

Consider the linear controlled singularly impulsive system (IV.79), (IV.80). Assume there exist $P \in \mathbb{P}^n$ and $\hat{P}_q \in \mathcal{N}^{(2q,n)}$, $q = 2, \ldots, r$, such that

$$\begin{aligned} 0 &= x^{\mathrm{T}}(A_{\mathrm{c}}^{\mathrm{T}}PE_{\mathrm{c}} + E_{\mathrm{c}}^{\mathrm{T}}PA_{\mathrm{c}} + E_{\mathrm{c}}^{\mathrm{T}}R_{1\mathrm{c}}E_{\mathrm{c}} - PB_{\mathrm{c}}R_{2\mathrm{c}}^{-1}B_{\mathrm{c}}^{\mathrm{T}}P)x, \\ &\quad x \notin \mathcal{Z}_x, \qquad \text{(V.96)} \\ 0 &= x^{\mathrm{T}}(\hat{P}_q[\overset{2q}{\oplus}(E_{\mathrm{c}}^{\mathrm{T}}A_{\mathrm{c}} - S_{\mathrm{c}}P)] + \hat{R}_{2q})x, \\ &\quad x \notin \mathcal{Z}_x, \quad q = 2, \ldots, r, \qquad \text{(V.97)} \\ 0 &= x^{\mathrm{T}}(A_{\mathrm{d}}^{\mathrm{T}}PA_{\mathrm{d}} - E_{\mathrm{d}}^{\mathrm{T}}PE_{\mathrm{d}} + E_{\mathrm{d}}^{\mathrm{T}}R_{1\mathrm{d}}E_{\mathrm{d}})x, \\ &\quad x \in \mathcal{Z}_x, \qquad \text{(V.98)} \\ 0 &= x^{\mathrm{T}}(\hat{P}_q[A_{\mathrm{d}}^{[2q]} - E_{\mathrm{d}}^{[2q]}] + \hat{\hat{R}}_{2q})x, \\ &\quad x \in \mathcal{Z}_x, \qquad q = 2, \ldots, r. \qquad \text{(V.99)} \end{aligned}$$

Then the zero solution $x(t) \equiv 0$ of the closed-loop system (IV.79), (IV.80) is globally asymptotically stable with the feedback control law

$$\begin{aligned} \phi_{\mathrm{c}}(x) &= -R_{2\mathrm{c}}^{-1}B_{\mathrm{c}}^{\mathrm{T}}(PE_{\mathrm{c}}x + \tfrac{1}{2}g'^{\mathrm{T}}(E_{\mathrm{c}}x)), \\ &\quad x \notin \mathcal{Z}_x, \qquad \text{(V.100)} \end{aligned}$$

where $g(x) \triangleq \sum_{q=2}^{r} \hat{P}_q E_{\mathrm{c}} x^{[2q]}$, and the performance functional (III.45), with $R_{2\mathrm{c}}(x) = R_{2\mathrm{c}}$ and

$$\begin{aligned} L_{1\mathrm{c}}(E_{\mathrm{c}}x) &= x^{\mathrm{T}}E_{\mathrm{c}}R_{1\mathrm{c}}xE_{\mathrm{c}} + \sum_{q=2}^{r}\hat{R}_{2q}E_{\mathrm{c}}x^{[2q]} \\ &\quad + \tfrac{1}{4}g'(E_{\mathrm{c}}x)S_{\mathrm{c}}g'^{\mathrm{T}}(E_{\mathrm{c}}x), \qquad \text{(V.101)} \\ L_{1\mathrm{d}}(x) &= x^{\mathrm{T}}E_{\mathrm{d}}^{\mathrm{T}}R_{1\mathrm{d}}E_{\mathrm{d}}x + \sum_{q=2}^{r}\hat{\hat{R}}_{2q}E_{\mathrm{d}}x^{[2q]}, \text{(V.102)} \end{aligned}$$

is minimized in the sense that

$$\begin{aligned} J(E_{\mathrm{c}}x_0, \phi_{\mathrm{c}}(x(\cdot))) &= \\ &\min_{u_{\mathrm{c}}(\cdot)\in\mathcal{C}(x_0)} J(E_{\mathrm{c}}x_0, u_{\mathrm{c}}(\cdot)), \quad x_0 \in \mathbb{R}^n. \text{(V.103)} \end{aligned}$$

Finally,

$$\begin{aligned} J(E_{\mathrm{c}}x_0, \phi_{\mathrm{c}}(x(\cdot))) &= x_0^{\mathrm{T}}E_{\mathrm{c}}^{\mathrm{T}}PE_{\mathrm{c}}x_0 + \sum_{q=2}^{r}\hat{P}_qE_{\mathrm{c}}x_0^{[2q]}, \\ &\quad x_0 \in \mathbb{R}^n. \qquad \text{(V.104)} \end{aligned}$$

[3] and [7].

*Proof:* The result is a direct consequence of Theorem II with $f_{\mathrm{c}}(x) = A_{\mathrm{c}}x$, $f_{\mathrm{d}}(x) = (A_{\mathrm{d}} - E_{\mathrm{d}})x$, $G_{\mathrm{c}}(x) = B_{\mathrm{c}}$, $G_{\mathrm{d}}(x) = 0$, $u_{\mathrm{d}} = 0$, $R_{2\mathrm{c}}(x) = R_{2\mathrm{c}}$, $R_{2\mathrm{d}}(x) = I_{m_{\mathrm{d}}}$, and $V(E_{\mathrm{c/d}}x) = x^{\mathrm{T}}E_{\mathrm{c/d}}^{\mathrm{T}}PE_{\mathrm{c/d}}x + \sum_{q=2}^{r}\hat{P}_qE_{\mathrm{c}}x^{[2q]}$. Specifically, for $x \notin \mathcal{Z}_x$ it follows from (IV.81), (V.97), and (V.100) that

$$\begin{aligned} V'(E_{\mathrm{c}}x)[f_{\mathrm{c}}(x) &- \tfrac{1}{2}G_{\mathrm{c}}(x)R_{2\mathrm{c}}^{-1}(x)G_{\mathrm{c}}^{\mathrm{T}}(x)V'^{\mathrm{T}}(E_{\mathrm{c}}x)] = \\ &-x^{\mathrm{T}}E_{\mathrm{c}}^{\mathrm{T}}R_{1\mathrm{c}}E_{\mathrm{c}}x - \sum_{q=2}^{r}\hat{R}_{2q}E_{\mathrm{c}}x^{[2q]} \\ &-\phi_{\mathrm{c}}^{\mathrm{T}}(x)R_{2\mathrm{c}}\phi_{\mathrm{c}}(x) - \tfrac{1}{4}g'(E_{\mathrm{c}}x)S_{\mathrm{c}}g'^{\mathrm{T}}(x), \end{aligned}$$

which implies (2.2.13). For $x \in \mathcal{Z}_x$ it follows from (V.98) and (V.99) that

$$\begin{aligned} \Delta V(E_{\mathrm{d}}x) &= V(E_{\mathrm{d}}x + f_{\mathrm{d}}(x)) - V(E_{\mathrm{d}}x) = \\ &-x^{\mathrm{T}}E_{\mathrm{d}}^{\mathrm{T}}R_{1\mathrm{d}}E_{\mathrm{d}}x - \sum_{q=2}^{r}\hat{\hat{R}}_{2q}x^{[2q]}, \end{aligned}$$

which implies (2.2.14) with $G_{\mathrm{d}}(x) = 0$. Finally, with $u_{\mathrm{d}} = 0$, condition is automatically satisfied so that all the conditions of Corollary V are satisfied. □

## VI. Conclusion

In this paper we have developed optimal robust control and inverse optimal robust control results for the class of nonlinear uncertain singularly impulsive dynamical systems [5]. Results are based on Lyapunov and asymptotic stability theorems developed in [6], and results presented in [7].

## VII. Future Work

Further work will specialize results of this paper to time-delay systems.

## Acknowledgments

The author thanks to all of that were supportive and creative in achieving this results.

## REFERENCES

[1] HADDAD W. M., N. A. KABLAR, AND V. CHELLABOINA, ”Nonlinear Robust Control for Nonlinear Uncertain Impulsive Dynamical Systems,” *Proc. IEEE Conf. Dec. and Contr.*, pp. 2959–2964, Sidney, Australia, December (2000).

[2] HADDAD W. M., V. CHELLABONIA, AND N. A. KABLAR, ”Nonlinear Impulsive Dynamical Systems Part I: Stability and Dissipativity,” *Int. J. Contr.*, Vol. 74, pp. 1631–1658, (2001a).

[3] HADDAD W. M., V. CHELLABONIA, AND N. A. KABLAR, ”Nonlinear Impulsive Dynamical Systems Part II: Feedback Interconnections and Optimality,” *Int. J. Contr.*, Vol. 74, pp. 1659–1677, (2001b).

[4] HADDAD W. M., V. CHELLABOINA, J. FAUSZ, AND A. LEONESSA, ”Optimal Nonlinear Robust Control for Nonlinear Uncertain Systems,” *Int. J. Contr.*, Vol. 73, pp. 329–342, (2000).

[5] KABLAR N. A., ”Singularly Impulsive or Generalized Impulsive Dynamical Systems,” *Proc. Amer. Contr Conf.*, Vol. 6, pp. 5292- 5293, Denver, USA, June (2003a).

[6] KABLAR N. A., ”Singularly Impulsive or Generalized Impulsive Dynamical Systems: Lyapunov and Asymptotic Stability,” *Proc. IEEE Conf. Dec. Contr.*, Vol. 1, pp. 173-175, Maui, Hawaii, December (2003b).

[7] KABLAR N. A., ”Robust Stability Analysis of Nonlinear Uncertain Singularly Impulsive Dynamical Systems,” *Proc. Amer. Contr. Conf.*, pp. 7-11, Portland, OR, June (2005).

[8] LAKSHMIKANTHAM V. , D. D. BAINOV, AND P. S. SIMEONOV, *Theory of Impulsive Differential Equations.* Singapore, World Scientific, (1989).

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Survey on Impact of QMS ISO 9001:2000 in an Organization, Increases the Effectiveness of its Operations

Salman Mahmood
Computer Engineering Department
Sir Syed University of Engineering & Technology
Karachi, Pakistan
salmanm@ssuet.edu.pk

Raza Hasan
Computer Engineering Department
Sir Syed University of Engineering & Technology
Karachi, Pakistan
raza_6@hotmail.com

***Abstract*— this research paper will set out to explore the degree to which the practice of ISO 9001:2000 is used in Searle Pakistan limited in developing quality products that meets the customer requirement. The paper focuses on developing the understanding the operation and effectiveness of quality management systems and why companies are also committed to the needs of ISO9001:2000. A survey was conducted on Searle Pakistan limited registered with ISO 9001:2000 to evaluate the head office unit and factory units of the company to know how effective ISO9001:2000 standard to meet customer requirement which enhance to customer satisfaction. Also, to investigate the role of ISO9001:2000 standard that how its implementation help to improve business operation and its impact regarding to quality of the Searle Pakistan limited within the head office unit and factory unit.**

***Keywords-component; Quality; Operation; Effectiveness; ISO; Searle; Customer Satisfaction; Business;***

## I. INTRODUCTION

At any time in history, the quest for qualities is probably more widespread and intense globally. Organizations have realized that the key to increased productivity and profitability is improving quality and in order to survive competition from local market and international market, they are forced to return to the basics of better quality management and cost competitiveness measures for their products and services. The factors that relate to business performance can be divided into two categories. First, those that improve the product or service quality differential against competitors and, second, those factors that reduces the cost of quality. An effective quality management system determines the product and service quality conformance as its primary goal. As Jacobson and Aaker found product quality had a positive influence on return on investment, market share and price [1].

Such quality efforts at an international level can be address with the implementation of the ISO 9001:2000 international standard. An effective quality management system based on ISO9001:2000 determine process control as an essential activity. Better process control would be consistently associated with less rework and hence lower costs. These lower costs will lead to better comparative business performance. Therefore, it is methodically line up regarding to importance of quality in line, with Deming who reasons that, as quality improves, waste is eliminated, costs are reduced, and performance improves [2].

An ISO9001:2000 is a set of processes that describe the organisational attitude and clearly define the areas of responsibility. It creates a common language, a common picture for the entire organisation. An ISO9001:2000 in this definition has two primary purposes.

Firstly, if one needs to perform some work that is not the usual day-to-day work task, the documents serve as support and guideline for carrying out the work.

Secondly, the documents are used if there is a need to review some processes and working procedures due to problems or improvement work.

Nowadays, pharmaceutical companies come across to many new challenges to ensure efficient business operations. There are external challenges from competitors, generic drug manufacturers, health-care organisations, in addition to internal challenges to decrease the costs of sales, R&D, and manufacturing. Government guidelines in the form of regulatory requirements need to be received, interpreted and disseminated in a timely manner to ensure compliance [8].

Therefore, the most fundamental focus of the pharmaceutical companies is an effective quality management system for efficient business operations and to improve the quality to meet the customer requirements which enhance to 'customer satisfaction'. Hence, ISO9001:2000 is an essential part of any organization regarding to its standardization. However, in order to ensure that Pakistani companies do not fall behind whilst competing in world markets all organizations implement quality management system practices for their customer satisfaction.

Searle is one of the leading Pakistani companies in pharmaceutical sector. Searle have a very smart manufacturing plant follow all modern rules. They are consistently investing heavily in new technologies and state-of-the-art equipment; this has been instrumental in improving plant efficiencies and curtailing manufacturing costs. Searle is one of the largest national pharmaceutical companies of Pakistan. As per 4th quarter 2008 of Pakistan Pharmaceutical Index (IMS) which is regarded as a reference publication by Pharmaceutical Industry,

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

Searle corporate ranking is 11th in the Pakistan Pharmaceuticals market with a market share of 2.4% [3].

The paper is organized as follows: Section II gives the research strategy. Section III describes the research method. Section IV introduces data collecting evidence. Section V describes how the data was collected. Section VI describes the quality of research. Section VII gives the research map. Section VIII describes how the sampling is done. Section IX concludes the result presentation and Section X concludes the paper.

## II. Research Strategy

A central part of research strategy of the survey is to evaluate current practice of ISO9001:2000, awareness of principle of quality management system, flexibility of ISO9001:2000 and give clear concept about the need of ISO9001:2000 standard by research techniques. The research strategy covered the following stages:

- Utilization of existing information, including published and unpublished research and secondary data.
- Qualitative method and Quantitative method for e.g. questionnaire based survey.

## III. Research Method

There are many research methods available which comes from these approaches. No research method is one hundred percent qualitative or quantitative but each method can be considered to be on a continuum.

This survey based on both qualitative and quantitative approaches. Random samples of employees are analysed by both approaches, which allows drawing conclusion based on the results.

## IV. Data Collecting Evidence

The methods use in this survey is:

- Documents; in the form of reports to shareholders, books, journal and newspapers.
- Articles in form of brochures and product descriptions.
- Direct observation in the form of informal observations and notes taken during the period of the study.
- Participant observation; in the form of field-notes taken after facilitating questionnaire survey.

## V. Collecting The Data

Questionnaire was designed to evaluate the current practices of ISO9001:2000 within the two areas of Searle Pakistan limited. Two types of questionnaire are formulated for the two different areas of the organisation which are head office and factory. A total of Fifty-Seven employees including directors, general manager ,manager assistant manager, technical staff and sale persons in both areas of **Searle Pakistan limited** approaches to provide the primary input into the study. Although there are commonalities in the nature of questions used in them, each one was adapted to ensure relevance and suitability for each type of areas of the **Searle Pakistan limited**.

Each questionnaire divides into three main areas:

- **Part A:** General
- **Part B:** Quality management system's strategies
- **Part C:** Practice and experience of ISO9001:2000

The questionnaire contains mainly 'open' questions where respondents are invited to express their opinions. In addition there are a number of questions involving rating or scale as well as Yes/No choice. The layout of the questionnaire design so that it seems easy to complete and this involve using some widely accepted principles of good practice:

- Provide a short covering letter explaining the purpose of research.
- A brief description of the subject areas under consideration.
- A brief instruction about how to complete.
- Started with simpler factual questions (e.g. biographical details such as age and length of service), moving on to items of opinion or values.
- Types of question varied occasionally.

Enclosed envelops is distribute to appropriate employees in head office and factory areas. Each envelope contains a covering letter, questionnaire and reply envelope which delivered and collected from the reception of head office and factory.

It requested to reply within a week such that the study may proceed to next stage.

It was assured to the respondents that completed forms will be treated as strictly confidential. A summary of the final results will be made available upon requested.

## VI. Quality of the Research

The quality of research depends on:

### A. *Validity*

For a successful research attitude the quality of it must be high, to judge this the *validity* and *reliability* is assessed.

Validity concerns the issue whether or not the findings can be shown to be valid for the problem that is being investigated. Data collected must be relevant to the problem and the purpose of the research otherwise there will be low validity. Irrelevant data and unnecessary information leads to low validity [4]. There are six strategies that can be used to check validity

- Triangulation,
- Checks,
- Long-term observation,
- Peer examination,

- Participatory/ collaborative models of research,
- Researcher's biases [5].

In this survey all the irrelevant questions have been ignored and all the relevant questions are included and secondary data collected from the reliable and trusted sources. This data base is updated at regular intervals of time, just after conducting them so that all information stores correctly and these factors will make to feel that the contents is relevant to the purpose of the research.

### *B. Reliability*

Reliability concerns the issue of consistent results of the study if it was replicated. A good guideline is to make sure that if someone did the project again, the same results would be found [6]. Reliability is an important aspect of doing a survey and the goal of reliability is to minimise biases and errors in the research study. A prerequisite for reliability is that all the documentation is in proper order and can be easily found [4].

Trying to maintain a transparency in how sense is made from the raw data the designed questions is follow a specific or set agenda in order to generate relevant and reliable results; however answers from employees is highly subjective as they give response on perception. Questionnaire is designed in a clear understandable manner so that question doesn't mislead employees and helps them in giving reliable and consistent answers.

A written statement is given about the extent to which will disclose the information about the organisation in this report, and also the secrecy of the questionnaire, so that they give vital and the exact information which helped in doing this research. Each case recorded and referred with a code but not with any personal detail.

The findings of the survey can be applied to other organisations in Pakistan. The results identify important factors in the development process and it believes that the research will be both valid and reliable.

## VII. Research Map

To fulfill the research aim and objectives, it is essential the collection of data in pre-planned manner and to use the available data in an effective format. Table I shows how the collected information fulfils the required objectives.

**TABLE I. HOW TO FIND THE ESSENTIAL INFORMATION TO FULFIL RESEARCH OBJECTIVES**

| **Research Objective** | **Research Map** |
|---|---|
| ***What to find Out*** | ***How the essential information find*** |
| To understand the operation and effectiveness of quality management systems is and why companies are also committed to the needs of ISO9001:2000. | Literature Review, Comprehensive search on internet such as company's web site, Journals, articles and annual report.<br>Literature Review, study of market segments, questionnaire with the managers of the firm follow up with lots of informal chats.<br>Articles about the performance of the firm in the recent years, annual reports. |
| To determine the actual experiences and opinions of the operating managers and employees in relation to ISO (9001:2000) in their organisation. | Questionnaire to target managers and employees in order to gauge view and some informal interviews if necessary |
| The main focus would be identifying how companies can exploit their use of ISO(9001:2000) to gain competitive advantage and also to improve the flexibility and efficiency of the organisation | Comprehensive data analysis from which required results can be collected and manipulated to establish conclusion and recommendations. |

## VIII. Sampling

The sampling frame that was adopted for this research findings is probability sampling because; smaller number of cases for possible a higher overall data means that more time can be spend designing and piloting the means of collecting these data. Using sampling makes possible a higher overall accuracy. And the sampling technique will be 'Simple Random Sampling' technique because it depends on research questions and objectives, selecting the sample of employees at random from the sampling edge [7].

## IX. Result Presentation

This section provides a result of the response of the survey. This analysis based on the survey respondents of gender, age group, number of years with present employer, and the source of quality management knowledge. This is followed by an overview of the results from the questionnaires sent to head office and factory of Searle pharmaceutical company.

### *A. Rate of response to the survey*

Table II shows that the overall response rate to the questionnaire survey is 66.6%.A breakdown of the response rate by head office and the factory are shown below:

**TABLE II.** OVERALL RESPONSE RATE

| **Unit** | **Total Sent Out** | **Total Recieved** | **% Received** |
|---|---|---|---|
| Factory | 32 | 20 | 62.5 |
| Head Office | 25 | 18 | 72 |
| Total | 57 | 38 | 66.6 |

Whilst, the overall response rate of 66.6 % is not significant surprise, the 72% returned from the head office and .62.5% returned from the factory. Keeping in mind that there are some major concerns over Factory's sensitivity and security therefore, given the current response from the factory is understandable.

The response rate could have been higher with improved questionnaire design. On reflection, the length of the questionnaire probably over reflected the amount of time required for completion. Whilst all sections did not necessarily require completing this may not have been readily apparent.

However, even with the responses received a great deal of consistent and quantitative results have been established.

### B. *Profile of respondents*

The proportion of women responding to the questionnaire is 29% and this consisted six individuals from head office and five individuals from factory. The proportion of men responding to the questionnaire is 71% and this consisted twelve individuals from head office and fifteen individuals from factory .all respondents are aged in over eighteen years consisted 5.55% under twenty,44.44% twenty plus ,33.33%thirty plus and 16.66% of forty years old. Just 16.66% of the respondents have spent ten plus years with their present employer, 16.66% up to two years service, 44.44% of two plus year's service and 22.22% of five years service with their present employer in the head office.

In the factory 10% up to two years,40% two plus years,30 % five plus years and 20 % ten plus years of service with their present employer in the factory.

By far most common source of ISO knowledge is by way of training consisted 85% in factory and 86% in the head office. From the work experience 15 % in factory and 12% in head office. 4% respondent from head office claimed that their source of knowledge regarding to ISO knowledge is seminars.

### C. *By source of ISO knowledge*

An evaluation of the responses by sex and age group failed to generate any significant differences in the analysis of questionnaires.

This is because tapered distribution across these ranges. However, respondent which indicated work experience as the source of their knowledge of ISO provided a detail qualitative data and this tends to be positive impact by comparison. But, the wide range of knowledge sources claimed by the head office and factory unit is training. It is because of number of years with present employer consisted of 40% from factory and 44.44% head office is 2-4years. These employees have only source of knowledge of ISO is training, which reflected in a more holistic view in terms of the depth and detail of the responses provided.

## X. Conclusion

Results indicated that a vast majority of respondents were satisfied with most aspects of the ISO9001:2000. The ISO9001:2000 indicate that achieving conformance to specification with low levels of rework has a direct effect on competitive advantage and as well as in terms of 'Customer Satisfaction'. In general, satisfied customers are likely to engaged in repeat purchase and reflect strong loyalty toward the Searle. Seeing as ISO9001:2000 is an important component of the total package of value required by customers, and the Searle use accreditation of ISO9001:2000 to attract new customers from local market as well as international markets. Employees are a significant part of the service delivery process and play a significant role in company-customer interaction.

The results of the study show that the majority of Searle employees are well trained regarding to ISO9001:2000 standard. Outstanding figures of respondents were aware of quality management principles and its generic requirements due to training conducted periodically in both units of Searle. Majority of employees are young. They are currently working in Searle which impact on the company is optimistic as aspects of improving their skills due to training conducted periodically. As it seen that young people can adopt changes more quickly and effectively which is very beneficial for both of units of Searle which directly lead to quality improvement.

Also, employee's awareness and usage rates are fairly high for key quality management principles and general requirements of ISO9001:2000. As high-quality control and processes are related to competitive advantage.ISO9001:2000 is an effective quality management system which has process control as an essential activity. Better process control, be consistently associated with less rework and hence lower costs. These lower costs will lead to better comparative business performance. This is lined up with Deming [2] who reasons that, as quality improves, waste is eliminated, costs are reduced, and financial performance improves. The results also indicate that ISO9001:2000 company's management system provides a framework for controlling and improving business activities. Which Adds value to products, services and competitiveness and provides a marketing edge.

## Acknowledgment

The authors acknowledge the encouragement, support and guidance given by **Finance Director, Qamar Mashkoor, Searle Pakistan Limited**.

Special thanks to all other faculty members for helping during the research and giving their valuable suggestion and guidance.

## References

[1] Jacobson, R., Aaker, A.D., ***"The Strategic Role of Product Quality"*** Journal of Marketing, 51(4), pp. 31-44,. 1987. DOI: 10.2307/1251246.

[2] Deming, W. Edwards, ***"Out of the Crisis"***, Cambridge, MA: MIT Center for Advanced Engineering Study, 1986.

[3] **Searle Pakistan Limited (SPL)**, URL: http://www.searlepak.com/index.html, retrieved 23 December 2011.

[4] Wikman, A., ***"Reliability, Validity and True Values in Surveys"***, Social Indicators Research, Vol.78, Iss.1; pg: 85, 2006.

[5] Saunders, M. Lewis, P. and Thornhill, A., ***"Research Methods for Business Students"***,3rd Edition, Pearson Education Limited, London, U.K., 2003.

[6] C. Robson, ***"Real World Research"***, Second Edition: Blackwell Publishing, 2002.

[7] Henry, G. T., ***"Practical Sampling"***. Newbury Park, CA: Sage Publications, 1990.

[8] Poksinska, Bozena, Jörgen A E Eklund, and Jens Jörn Dahlgaard. ***"ISO 9001:2000 in small organisations: Lost opportunities, benefits and influencing factors."*** International Journal of Quality Reliability Management 23.5 (2006) : 490-512.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Fault Detection and Isolation for Engine under Closed-Loop Control

Adnan Hamad, Dingli Yu, J B Gomm
Control system research group, school of engineering
Liverpool John Moore University
Liverpool, UK
adnanbohliga@yahoo.com

Mahavir S Sangha
Test Technology & Emissions, Cummins Inc
Daventry, UK

*Abstract*—**Fault detection and isolation (FDI) have become one of the most important aspects of automobile design. Fault detection and isolation for engine open loop system was investigated in many research. In fact, the simulation results obtained from engine open loop system do not reflect the real situation for automotive engine. In the practice the engine works as closed-loop control system. In this paper, a new FDI scheme is developed for automotive engines under closed-loop control system. Test the method using closed-loop system has been done. The method uses an independent radial basis function (RBF) neural network model to model engine dynamics, and the modeling errors are used to form the basis for residual generation. Furthermore, another RBF network is used as a fault classifier to isolate occurred fault from other possible faults in the system. The performance of the developed scheme is assessed using an engine benchmark, the Mean Value Engine Model (MVEM) with Matlab/Simulink. Six faults have been simulated on the MVEM, including four sensor faults, one component fault and one actuator fault. The simulation results show that all the simulated faults can be clearly detected and isolated in dynamic conditions throughout the engine operating range.**

***Keywords: Automotive engines under closed-loop control, independent RBF model, RBF neural network, fault detection, fault isolation.***

## I. INTRODUCTION

A fault is any type of malfunction of components that may happen in a system and this fault will degrade the system performance. Fault detection is the program which informs us that something wrong in the system and needs to be repaired. Also, fault isolation is way to determine which fault occurs among the possible faults. To detect faults we usually compare the outputs of the real system which is in this paper the MVEM, and the outputs of a neural network model of the engine. Rolf Isermann has proposed Model-based fault-detection and diagnosis methods for some technical processes [1]. On-line sensor fault detection, isolation, and accommodation in automotive engines had been studied by Domenico Capriglione [2]. Fault detection and isolation for MVEM open loop control system were achieved in previous paper of the authors [3]. However, in the practice the engine does not work as an open loop control system. Automotive engine works under close loop control system with feedback control. Fault detection and isolation using close loop system is quite different comparing with open loop system. The main points are: fault detection and isolation is much more difficult than using open loop system. The reason is the process outputs in close loop system will be fed back and this will affect the sensor faults. Secondly, in MVEM open loop system, Air fuel ratio (AFR) was not controlled at the 14.7, because the feature of feed-back did not use in the open loop control system. In this paper, a new fault detection and isolation method will be implemented by using MVEM when this model is under close loop control system. An independent RBFNN model is used to model a dynamic system using RAS throttle angle as an input. Feed-back (FB) and feed-forward (FF) control methods will be applied to the MVEM. The K-means clustering algorithm is used to choose the centres of RBFFNN. Recursive least squares (RLS) algorithm is used to update for each new sample the parameter matrix W.

## II. CONTROL STRUCTURE FOR MVEM

Fig. 1 shows the Simulink model of the automatic control loop for the MVEM including feed-forward and feed-back controllers. Where the MVEM control input u is the injected fuel mass mfi and the disturbance input Ø is the throttle angle position. The feed-forward controller that correlates the steady state value between the MVEM control input mfi and the disturbance Ø will be used in the feed-forward path. In order to achieve better transient response, feed-forward and feed-back controllers will be designed as following.

### *A. FF controller design*

The feed-forward controller will be implemented by look-up table configuration. The data of this table were determined from the MVEM. Firstly, throttle angle position value have been given to the MVEM starting from 20 to 60 degree by step 5 degree in order to cover 9 cases, secondly, the gain k has been changed for each case to adjust the air fuel ratio equal to 14.7. Finally, the suitable corresponding injected fuel mass can be determined for each throttle angle value by using (1).

$$mfi = k \times \phi \tag{1}$$

Omar AL-Mukhtar University, Libya

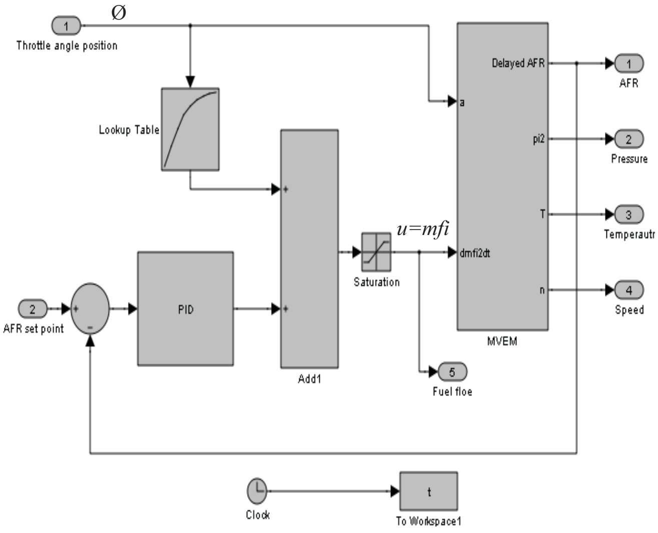


Figure 1. Simulink model for the automatic control loop for the MVEM including feed-forward and feedback controllers

### B. PID controller design

In general, the transfer function of PID is illustrated in (2), where, $K_p$, $K_i$ and $K_d$ are proportional, integral and differential gains respectively.

$$G_c(s) = K_p + \frac{K_i}{s} + K_d s \quad (2)$$

Where:

$$K_i = \frac{K_p}{T_i} \quad (3)$$

$$K_d = K_p T_d \quad (4)$$

Where:

$T_i$: Integral time.

$T_d$: Derivative time.

In order to find out the PID controller parameters ($K_p$, $Ki$ and $K_d$), many numerous tuning rules for PID controller can be found in the literature. The process of selecting the controller parameters to meet given performance specification is known as controller tuning. The rules for determining values of the $K_p$, $T_i$ and $T_d$ based on the transient response characteristics of given plant have been proposed by Zigler and Nichols [4]. All the PID parameters were determined by using the Matlab software R2009a. Equation 5 shows the transfer function of the PID controller after calculate all its parameters.

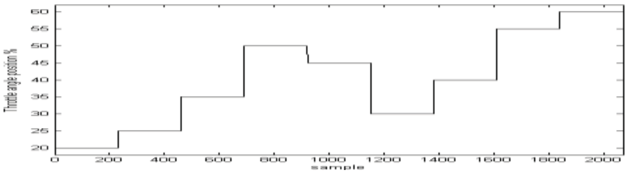
$$G_c(s) = 0.00001(1 + \frac{1}{100s} + 0.01s) \quad (5)$$

## III. Evaluation of closed loop control system

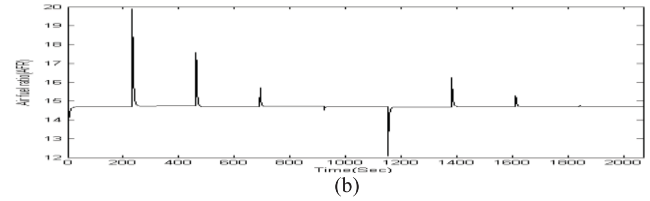
In order to evaluate the MVEM under close loop control system, a set of signal was used for the throttle angle position to obtain a representative set of input data.The range of this excitation signals was bounded between 20 and 60 degrees. This almost covers the whole throttle angle position in normal operation condition. The outputs of the PID and feed forward controllers will be used as a second input of the MVEM (see Fig. 1). Fig. 2(a ,b) illustrate the input signal of throttle angle position and the AFR output response of the MVEM close loop control obtained from the block diagram shown in Fig. 1. The AFR is to be controlled at 14.7. From the Fig. 2 it can be seen that the PID controller has good performance and the obtained results were very accurate, therefore the AFR has been controlled at 14.7.

## IV. Engine modeling

The first step in the engine modeling by using RBFNN is the generation of a suitable training RAS data set of throttle angle position and setpoint of AFR. As the training data will influence the accuracy of the neural network modeling performance, the objective of experiment design on training data is to make the measured data become maximally informative, subject to constraints that may be at hand. As mentioned above, a set of random amplitude signals (RAS) were designed for the throttle angle position and AFR setpoint to obtain a representative set of input data of MVEM close loop. The sample time of 0.02 sec was used. The second step is to determine the input variables of the RBF model. The SI engine to be modeled has two input variables: throttle angle and the outputs of the PID and feed forward controller which is fuel flow rate, and four outputs: air manifold temperature, air manifold pressure, crank shaft rotary speed and air fuel ratio. The network input that generated the smallest modeling errors was selected, and has first-order for the two process inputs and third order for process output. As selected above, the RBF model has 16 inputs and 4 outputs. The hidden layer nodes have been selected as 15. Before the training, 15 centres were chosen using the K-means clustering algorithm, and the width σ was chosen using the p-nearest-neighbours algorithm. All Gaussian functions in the 15 hidden layer nodes used the same width. For training the weights W the recursive least squares algorithm [5] was applied and the following initial values were used: $\mu$= 0.98, w (0) =$1.0\times10^{-6}\times U$ (nh×4), P(0)=$1.0\times10^{8}\times I$ (*nh*), where $\mu$ is the forgetting factor, $I$ is an identity matrix and $U$ is the matrix with all element unity, *nh* is the number of hidden layer nodes. Totally a data set with 7000 samples was collected from the MVEM. Before training and testing, the raw data is scaled linearly into the range of [0 1]. Fig. 3 show the model training results of the last 500 samples in the training data set and the first 500 samples in the test data set.

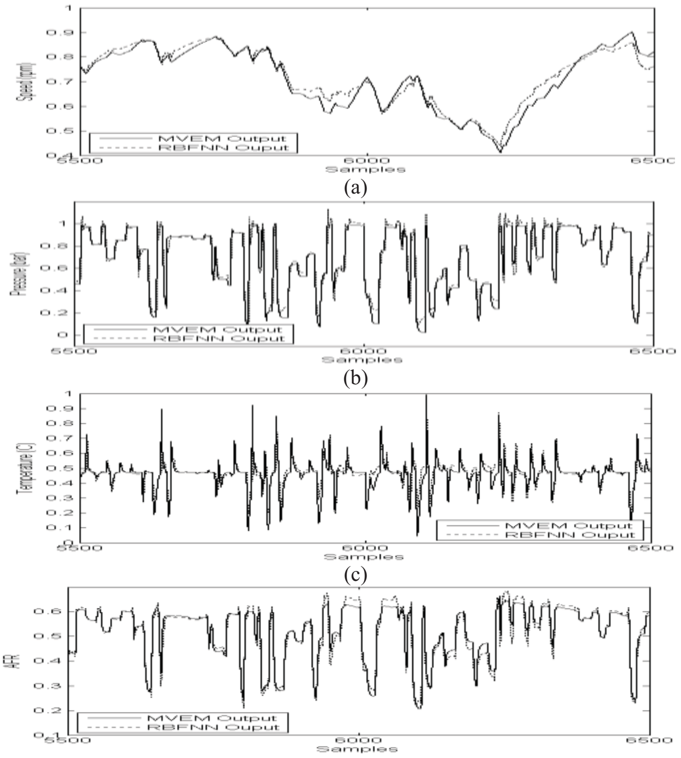


(a)



(b)

Figure 2. (a) the input signal of throttle angle position, (b) the AFR output response of the MVEM control loop



(a)

(b)

(c)

(d)

Figure 3. ,a,b,c and d. The simulation results of the speed, pressure, temperature and air fuel ratio engine model output and the RBFNN s output respectively

From Fig. 3, it can be seen that there is a good match between the two outputs with a very small error, in general. The modeling error of the training data set is smaller than the test data set. The mean absolute error (MAE) index is used to evaluate the modeling effects. For this model the MAE values of crankshaft speed, manifold pressure, manifold temperature and air fuel ratio are 0.0014, 0.0061, 0.0031 and 0.0031 respectively.

## V. SIMULATING FAULTS

Before the developed method is tested on a real engine with real faults, it was tested in this research on the nonlinear simulation of SI engines, the MVEM with different faults simulated on it. One component fault, one actuator fault and four sensor faults with different levels of intensity have been investigated as practical examples of spark ignition (SI) engine faults. The component fault is air leakage in the intake manifold. The actuator fault is a malfunction of the fuel injector. The four sensor faults are malfunction of the intake manifold pressure sensor, manifold temperature sensor, crank shaft speed sensor and air fuel ratio sensor. Details of the simulation of these faults are described as follows.

### *A. Component fault*

Equation (6) of the manifold pressure [6] is modified to (7) in order to collect the engine data subjected to the air leakage fault.

$$\dot{p}_i = \frac{T_i R}{V_i}(-\dot{m}_{ap} + \dot{m}_{at} + \dot{m}_{EGR}) \qquad (6)$$

$$\dot{p}_i = \frac{T_i R}{V_i}(-\dot{m}_{ap} + \dot{m}_{at} + \dot{m}_{EGR} - \Delta l) \qquad (7)$$

where $\dot{p}_i$ is the absolute manifold pressure (bar), $\dot{m}_{at}$ is the air mass flow rate past throttle plate (kg/sec), $\dot{m}_{ap}$ is the air mass flow rate into the intake port (kg/sec), $\dot{m}_{EGR}$ is the EGR mass flow rate (kg/sec). The added term is used to simulate the leakage from the air manifold, which is subtracted to increase the air outflow from the intake manifold. $\Delta l = 0$ represents no air leak in the intake manifold. The air leakage level is simulated as 20% of total air intake in the intake manifold. This fault occurs from the sample number 3750~ 3850 in the faulty data as shown in Fig.4, and was simulated by changing the Simulink model of the MVEM.

### *B. Actuator fault*

For SI engines, the target is to achieve an air–fuel mixture with a ratio of 14.7 kg air to 1 kg fuel. This means the normal value of air fuel ratio is 14.7. Because any mixture less than 14.7 to 1 is considered to be a rich mixture, any more than 14.7 to 1 is a lean mixture. Lean mixture causes the efficiency of the engine reduced, while rich mixture will cause emission increased. The fuel injector is controlled by the controller with correct amount of fuel. If the fuel injector has any fault the injected fuel amount will not be correct and affect the air/fuel ratio. Here, the malfunction of the fuel inject is simulated by reducing the injected fuel amount of 25% of the total fuel mass flow rate between the sample number 2550 and 2650 as shown in Fig.4. This fault is also simulated by changing the Simulink model of the MVEM.

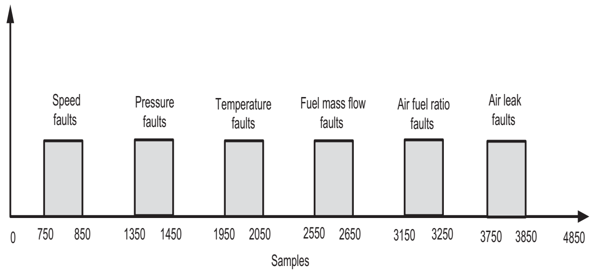


Figure 4. Distribution of the simulated faults

### C. *Sensor fault*

The four sensor faults considered are (10, 15, 10 and 20) % changes superimposed on the outputs of crankshaft speed, manifold pressure, and temperature and air fuel ratio sensors respectively. These faults are simulated from sample numbers 750 to 850, 1350 to 1450, 1950 to 2050 and 2550 to 2650 respectively as shown in Fig.4. The faulty data for the sensors is generated using multiplying factors (MFs) of 1.1, 1.1, 1.15 and 1.2 for the above over-reading faults respectively. Faulty data are generated by the Modified MVEM with throttle angle at different values between 20$^o$ and 60$^o$ for all the fault conditions. The 6 states with their multiplying factors (MFs) are given in table I. The sample time is chosen as 0.02 sec.

## VI. Fault detection

Fig.5 shows the information flow for the fault detection and isolation. Firstly, the 7000 samples data set of random amplitude sequence for throttle angle in the proper range and the outputs of the PID and feed forward controllers are fed into the MVEM under close loop control system. The collected four engine outputs together with the two inputs as well as their delayed values are used to train the RBF model. After training, all the six faults are simulated to the MVEM. Then, with another 4850 set of square signals of throttle angle position (see Fig. 6) and fuel flow which is output of PID and feed forward controller (see Fig. 7) fed into the MVEM under close loop control system, the model prediction error and the filtered residual are generated for fault detection. With the six faults simulated from samples 750 to 3850 as shown in Fig.4.After a low-pass filter is used the filtered prediction errors are shown in Fig.8 (b, c, d & e). The first 500 samples of data set which mean the beginning 10 second of engine operation has been ignored because contain noise signals. The first filtered model prediction error of air fuel ratio is shown in Fig.8, b. The second, third and fourth for air manifold temperature, pressure and engine speed are showed in Figs.8(c, d, e) respectively. In these Figs. the samples 0 to 750 are data without faults. Including them is to show the prediction error is under the selected threshold in "no fault case". Now it is evident that all simulated faults have a significant reflection on the model prediction errors. A threshold is chosen for each prediction error and is also displayed in these Figs. Moreover, Fig. 8, f shows the residual error (*re*) which is generated by (8).

$$re = \sqrt[2]{e_n^{\;2} + e_p^{\;2} + e_t^{\;2}} \qquad (8)$$

Where $e_n$, $e_p$ and $e_t$ are the error vectors of the speed, pressure and temperature respectively between the engine model and the RBF neural network.

TABLE I. THE 6 FAULTS STATES AND MULTIPLYING FACTORS

| No | Fault Name | multiplying factors (MFs) |
|---|---|---|
| 1 | Air Leak 20% | |
| 2 | Injected fuel mass flow 25% | |
| 3 | Speed sensor 10% over reading | 1.1 |
| 4 | Pressure sensor 15% over reading | 1.15 |
| 5 | Temp. sensor 10% over reading | 1.1 |
| 6 | Air fuel ratio 20% over reading | 1.25 |

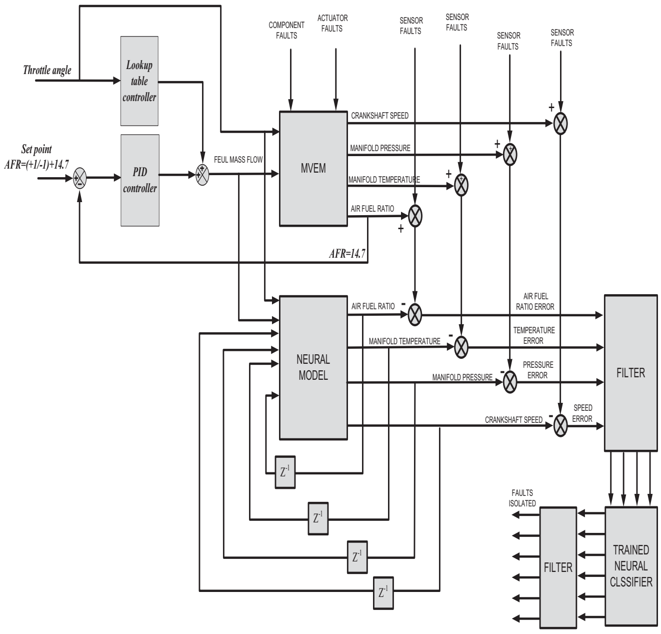


Figure 5. The information flow for the fault detection and isolation

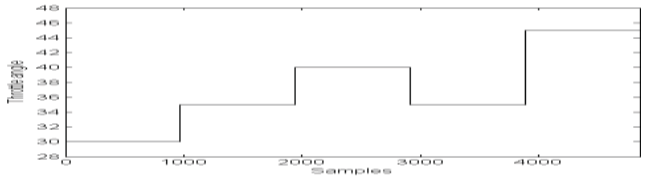


Figure 6. Square signals of throttle angle position

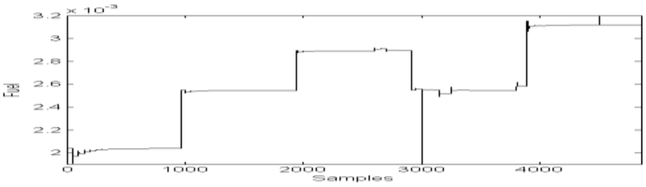
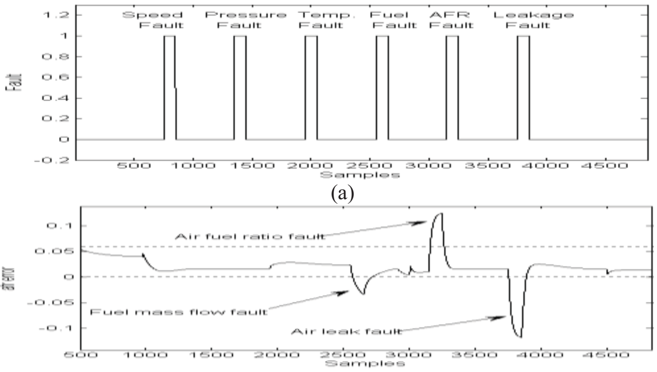
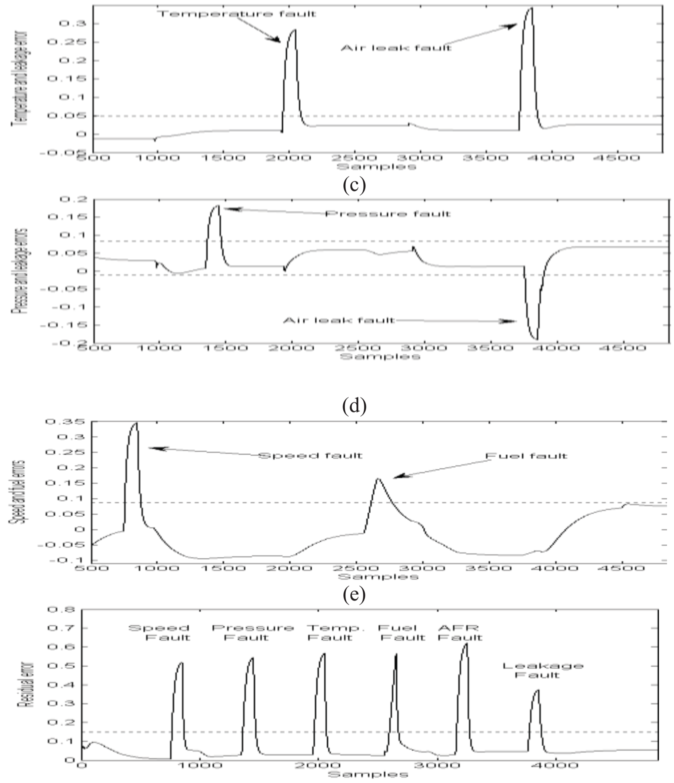


Figure 7. fuel flow rate (output of PID and feed forward controllers)



(a)

(b)



(c)

(d)

(e)

(f)

Figure 8. (a)simulated faults, (b), (c) , (d) and (e) Filtered model prediction error of air fuel ratio, air manifold temperature, air manifold pressure and engine speed , (f) Filtered residual

## VII. Fault isolation

As have been seen in Fig.8, f, though all simulated faults cause a significant deviation in the residual, this can be used to detect fault but cannot be used to isolate fault. When a fault occurs, only its associated output of the fault diagnosis system has a reflection, while all the other outputs should be insensitive to this fault, then the fault can be isolated from the other possible faults. In this research there are 6 possible faults but only four modeling errors. It can be concluded from Fig.8 that if only the first model output (air fuel ratio) goes over the threshold while the other three outputs remain under the thresholds, then it must be air fuel ratio sensor has fault. If second model output (temperature) goes over the threshold while the other three outputs remain under the thresholds, then it must be temperature sensor has fault. Similarly, if only the third model output (pressure) goes over the threshold while the other three outputs remain under the thresholds, then it must be the pressure sensor has fault. Moreover, if only the fourth model output (speed) goes over the threshold while the other three outputs remain under the thresholds, then it must be the speed sensor has fault. If both the first model output (air fuel ratio) and the forth model output (speed) go over the thresholds while the other two outputs remain under the thresholds, then it must be the fuel injector has fault. Finally, If the first, second and third model outputs go over the thresholds while the fourth output remain under the threshold, then it must be the air leak occurs. Therefore, to achieve a clear isolation among all possible faults, another RBF network is employed as a fault classifier. The classifier has four inputs each receiving one of the four modeling errors, and has seven outputs with one representing “no fault case” and the other six representing the six different faults. The classifier is trained in the following way. Collect 7 sets of data with first set without fault and the other six sets each with one fault only. For each data set of the seven, the target of the training for the output corresponding to the contained fault is set to “1”, while the targets for the other outputs set to “0”. Totally 4850 RAS data samples were collected with first 750 without fault and each set of 600 samples with one of the five faults, the last sixth fault occurred during the last set of 1100 samples. These data are fed into the new RBF model and the generated four modeling errors are fed into the RBF classifier to train it, with the targets given as described above. After training, the classifier is tested with the similar arrangement of data, totally 4850 data samples which are the modeling errors obtained from the RBF of fault detection and MVEM under close loop control system (see Fig. 8, b, c, d and e). The first 750 samples are fault-free, followed by six data sets. The first five data sets have 600 samples and having a single fault, the sixth data set has 1100 samples and has the last fault. Between any two of these six faulty data sets insert 500 fault-free samples and the final 1000 samples are fault-free, so that the residual rising time and disappearing time can be observed. The data samples with associated fault types are listed in Table II. Similar to the first RBF network, the centres and widths are also selected using the K-means clustering algorithm and the *P*-nearest centres method respectively. The network weights are trained using the recursive Least Squares algorithm with its parameters set as $\mu$= 1.0, w (0) =1.0×$10^{-6}$×U (nh×6), P (0) =1.0×105×I (*nh*). The number of the hidden layer nodes was tried several numbers and the one giving minimum training error was chosen and was 250. Figs. 9~14 show the test result after filtering. The isolation thresholds are chosen as shown in the Figs.

TABLE II. DATA SAMPLES AND FAULT TYPES

| Data samples | Fault types |
|---|---|
| 1 ~ 750 | No fault |
| 751 ~ 850 | Speed sensor fault |
| 801 ~ 1350 | No fault |
| 1351 ~ 1450 | Pressure sensor fault |
| 1451 ~ 1950 | No fault |
| 1951 ~ 2050 | Temperature sensor fault |
| 2051 ~ 2550 | No fault |
| 2551 ~ 2650 | Fuel injector fault |
| 2651 ~ 3150 | No fault |
| 3151 ~ 3250 | Air fuel ratio fault |
| 3251 ~ 3750 | No fault |
| 3751~3850 | Air leak fault |
| 3851~4850 | No fault |

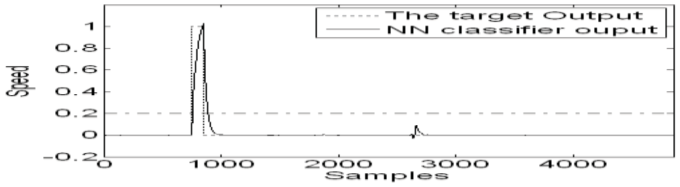


Figure 9. Filtered first output of fault classifier

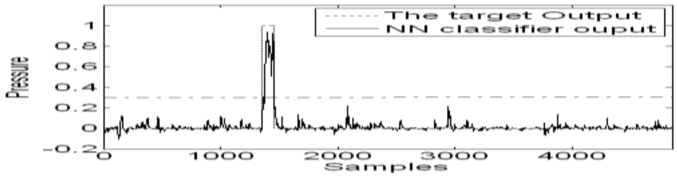


Figure 10. Filtered second output of fault classifier

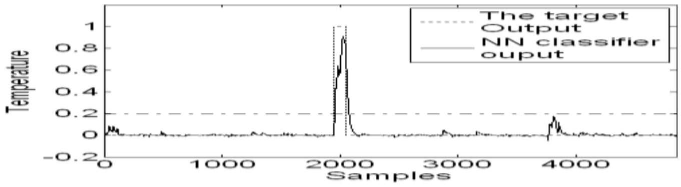


Figure 11. Filtered third output of fault classifier

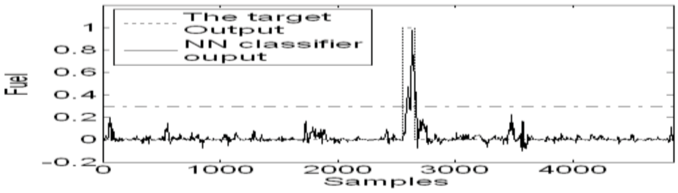


Figure 12. Filtered forth output of fault classifier

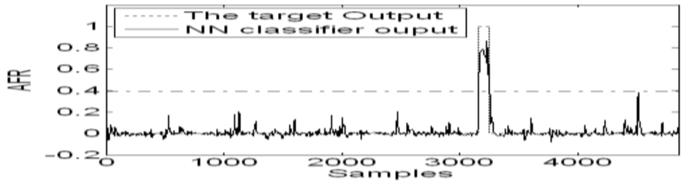


Figure 13. Filtered fifth output of fault classifier

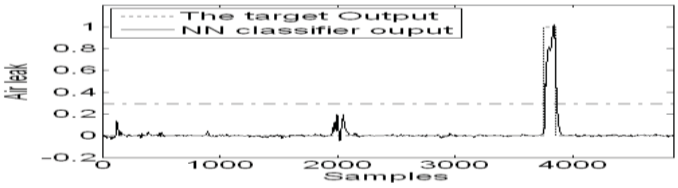


Figure 14. Filtered sixth output of fault classifier

## VIII. Discussion of Simulation Results

### A. Training the Neural Network model.

The simulation results of training and testing by using 15 hidden nodes were very good and in general, a good prediction between the engine model output and the RBF neural network output was achieved. From Fig. 3, it can be seen that the mean absolute error between the engine speed output and the RBFNN is very small.

### B. Detection of the Sensor, Component and Actuator Faults.

The Fig. 8 (b, c, d and e) show the test results of the fault detection of air fuel ratio/fuel flow/leakage, temperature/air leak, pressure/air leak, and speed/fuel injection respectively after filtering with 15 hidden nodes. It can be seen from Figs. after filtering operation that all kind of faults were detected clearly. The error values were between 0.05 and -0.1 except the samples in which faults occur. The detection thresholds were chosen as 0.09 for crankshaft speed / fuel injection, (-0.01/+0.08) for manifold pressure/air leak, 0.05 for manifold temperature/air leak and (0/0.06) for air fuel ratio.

### C. Isolation of the Sensor, Component and Actuator Faults.

In this section, another RBFNN called neural classifier were used in order to isolate all kinds of the faults. This neural was received error signals between the MVEM and the RBFNN model outputs. 50, 150 and 250 hidden nodes are used in order to try to obtain good simulation results. From the Figs. of the results, we found 250 hidden nodes after filtering operation were the best case and the simulated faults can be clearly isolated (see Figs. 9~14). The isolation thresholds are chosen as shown in the Figs. in case 250 hidden nodes.

## IX. Conclusions

In this research, the MVEM under close loop control including feed-forward and feed-back controllers was used. Look up table has been used as a feed-forward controller and PID used as feed-back controller. The data of the Look up table were determined by using the MVEM and the data of PID were calculated by Zigler and Nichols methods by using the Matlab software R2009a. The AFR output response of the MVEM close loop control was controlled at 14.7. Four sensor faults (intake manifold pressure, temperature, crankshaft speed and air fuel ratio), one component fault (leakage in the intake manifold) and one actuator fault (injected fuel mass flow) have been simulated. Two RBFNNs were used, the first one to model engine dynamics and the second one to isolate the sensor faults, component fault and actuator fault from the modeling errors. By using *p* – Nearest Neighbours method and K-means algorithm the width in hidden layer nodes of the RBF neural network σ and the centres c are calculated for both RBFNNs. The recursive least square algorithm was applied for training the weights *w* of the RBF neural networks. The proposed method can detect and isolate the faults and from the Figs. of simulation results, it can be seen that the methods were able to detect and isolate the faults.

## References

[1] R. Isermann, "Model-based fault-detection and diagnosis status and applications," Annual Reviews in Control, vol.29, pp.71–85 ,2005.

[2] D. Capriglione, C. Liguori, C. Pianese, and A. Pietrosanto, "Analytical redundancy for sensor fault isolation and accommodation in public transportation vehicles," IEEE Trans. on Instrumentation and Measurement, vol. 53, No.4, pp. 993-999, 2004.

[3] A. Hamad, D. Yu, J.B Gomm,. and S. Mahavir, "Automotive engine fault detection and isolation via radial basis function neural network," Proceeding of 16th international conference on automation & computing, editors Xun Chen, Jihong Wang ,Birmingham, UK.,pp.232-237 , ISBN978-0-9555293-6-8, 2010.

[4] K. Ogata, Modern control engineering. Prentice-Hall international, inc, 1997.

[5] Y. J. Zhai and D. Yu, "Radial Basis Function Based Feeddback Control for Air Fuel Ratio of Spark Ignition," Proc. IMACE Joural of Automobile Engineering, Vol. 222, pp. 415-428, 2007.

[6] E. Hendricks, D. Engler, and M. Fam, " A Generic Mean Value Engine Model for Spark Ignition Engines," Proceedings of 41st Simulation Conference, Denmark, DTU Lyngby, 2000.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# A Parameter-free Method for Sensor Fault Detection and Isolation in Bilinear Systems

Assia HAKEM, Komi Midzodzi PEKPE and Vincent COCQUEMPOT

*Abstract*— **This paper is concerned with Fault Detection and Isolation (FDI) and more specifically it focuses on a parameter-free residual generation method. The residual signals are obtained by projecting the measured signals onto the kernel of an extended input matrix, which depends on the structure of the system model. The method was not easily applicable in real-world applications due to a high computational complexity. In that paper, fault indicators are constructed differently, using kernels properties, to avoid this complexity problem. A simulated electromechanical actuator example is taken to illustrate the applicability of the method.**

*Index Terms*— **Fault detection and isolation, data driven methods, bilinear systems, electromechanical actuator.**

## I. Introduction

Real plants are subject to faults, that can affect the process parameters, actuators or sensors. Online fault detection and isolation (FDI) is an important task for human safety and system dependability. Many approaches have been reported in the literature to achieve this task. Two kinds of FDI methods are distinguished [16]. Model-based and model-free methods. Model-based methods consist in comparing the actual system behavior with the one given by an analytical model, i.e. a set of nonlinear differential equations. A signal called residual is used to evaluate this comparison. In the absence of fault and noise, if the process and the model are exactly matched, the residual is zero, otherwise it is different from zero which characterizes fault occurrence. The main common methods for model-based residual generation are:

- observer-based methods [5], [6] and [7]
- analytical redundancy relation (ARR) -based methods [1]
- parameters estimation methods [4], [8]

Unfortunately, the values of the model parameters are unknown in most practical applications. For such case, model-free FDI methods [2], [13] have been proposed. Some of these methods use signal processing techniques to extract special properties of measured signals, these methods are called signal-based methods (see [16] and [17]). Other data driven methods have been proposed recently for switching systems in [10] and [11], for bilinear systems in [11].

The residual generation method that is proposed in that paper is situated between model-based and model-free methods, since the only information we need is the knowledge of the input-output data and of the structure. The values of the model parameters are not needed.

The advantages of the proposed method are as follows:

- The only needed data are inputs and outputs.
- The generated residuals are structured which allows faults isolation.
- Multiple faults may be considered.

The bilinear system is a particular structure of nonlinear systems, this special class of systems has been widely studied in recent years [3]. Many real-world dynamical systems may be represented by a bilinear model and such model can approximate a large class of nonlinear systems. Consequently, bilinear models study is interesting from both theoretical and practical points of view.

The data-projection method for residual generation was extended for bilinear structure models in a previous publication by the authors [15].

The target of this paper is to enhance the parameter-free residual generation method proposed in [15]. The general principle of the method is kept. However, using kernel properties, fault indicators are computed differently, to avoid the computational complexity of method in [15].

The remainder of this paper is organized as follows. A general description of our residual generation method for bilinear models is provided in section 3.The input/output relation is derived in section 3 while the residual expression is derived in section 4. In section 5, simulation results on an electromechanical actuator are presented to show the effectiveness of our method. The final section gives a conclusion.

## II. Overview of the parameter-free residual generation method

Consider known inputs $u_k \in R^m$ and outputs $y_k \in R^\ell$ affected by colored white noise $w_k \in R^\ell$. These input/output signals are supposed to be collected on a physical plant that can be modeled as a discrete-time bilinear system given by:

$$\begin{cases} x_{k+1} = Ax_k + G(x_k \otimes u_k) + Bu_k \\ y_k = Cx_k + Du_k + f_k + w_k \end{cases} \tag{1}$$

where $\otimes$ represents the Kronecker product, and $f_k \in R^\ell$ is the sensor fault vector. It is supposed that the linear dynamic is stable i.e. $\|A\|_2^i \to 0$.

The aim is to detect and to isolate sensor faults when supposing that the only available information is the system structure (bilinear) and input/output data. The system parameters($A \in R^{n\times n}$, $B \in R^{n\times m}$, $C \in R^{\ell\times n}$, $D \in R^{\ell\times m}$, $G \in R^{n\times nm}$) are supposed to be unknown.

Assia HAKEM, Komi Midzodzi PEKPE and Vincent COCQUEMPOT, LAGIS UMR CNRS 8219, LILLE 1 University Villeneuve d'Ascq 59655, France,
assia.hakem@ed.univ-lille1.fr,
midzodzi.pekpe@univ-lille1.fr,
vincent.cocquempot@univ-lille1.fr

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

The data-based residual generation method for bilinear systems is detailed in the next section. Let us give the general principle of this method. Under the stability conditions, it is possible to express the vector of measured outputs on a given time-window as a function of the inputs. The following expression is thus obtained.

$$Y \cong HM(u, y \otimes u) \tag{2}$$

where $H$ depends on the system parameters, $Y$ is a matrix of outputs collected on a given time-window, $M(u, y \otimes u)$ is a matrix constructed using inputs and the Kronecker product between inputs and outputs. If the chosen time-window is sufficiently large, we can then project equation (2) on the right kernel $\Pi$ of $M(u, y \otimes u)$ $(UM(u, y \otimes u)\Pi = 0)$ and we can derive the relation:

$$Y\Pi = 0 \tag{3}$$

This relation must be verified in absence of disturbances and faults. Consider $Y_{online}$ and $\Pi_{online}$, the $Y$ and $\Pi$ matrices which are computed using online taken values of inputs and outputs signals. In the no fault situation, the signal $\epsilon = Y_{online}\Pi_{online}$, is not exactly null because of the measurement noise. However, it can be proved that $r = E[\epsilon]$ equals 0, with $E[.]$ the mathematical expectation. When a fault occurs, $r = E[Y_{online}\Pi_{online}]$ becomes different from zero. Thus, $r$ can be considered as a fault indicator (residual) to be used for FDI.

It is clear that no system parameter or state estimation is needed for residual computation since $\Pi$ depends only on inputs, which makes residual expression (3) independent on model parameters.

## III. Input-output expression of bilinear systems

The objective of this section is to show how to derive equation (2) from system (1). A general expression of the output $y_k$ is first obtained. Then it is shown that the influence of the state may be neglected under the stability conditions, which leads to equation (2).

A general output expression $y_k$ can be derived, which is given in the following proposition and proved by induction.

*Proposition 1:* The general expression of the output $y_k$ in function of the state $x_{k-i}$, the inputs and system parameters $A$, $B$, $C$, $D$, $G$ is given by:

$$\forall i \geq 0 : y_k = CA^i x_{k-i} + \widetilde{H}_i \overline{z}_{k,i} + \overline{H}_i \overline{u}_{k,i} + f_k + w_k \tag{4}$$

where $\widetilde{H}_i$, $\overline{z}_{k,i}$, $\overline{H}_i$, $\overline{u}_{k,i}$ are given as follows.
a) $\widetilde{H}_i$ and $\overline{H}_i$ depend only on system matrices.

- $\widetilde{H}_i$ depends on the system matrices $C$, $A$ and $G$:

$\widetilde{H}_i = \left[ CA^{i-1}G | \cdots | CAG | CG \right] \in R^{\ell \times nmi}$.

- $\overline{H}_i$ depends on the system parameters $C$, $A$ and $B$:

$\overline{H}_i = \left[ CA^{i-1}B | \cdots | CB | D \right] \in R^{\ell \times m(i+1)}$.
b) $\overline{z}_{k,i}$ and $\overline{u}_{k,i}$ depend on the system states and inputs on a time-window of size $i$.

$\overline{z}_{k,i} = \left[ z_{k-i}^T | \cdots | z_{k-2}^T | z_{k-1}^T \right]^T \in R^{nmi \times 1}$.
with $z_{k-i} = x_{k-i} \otimes u_{k_i}$.
and $\overline{u}_{k,i} = \left[ u_{k-i}^T | \cdots | u_{k-1}^T | u_k^T \right]^T \in R^{m(i+1) \times 1}$.

*Proof:*
Inductive method is used to prove correctness of the general output expression (4), which can be written differently without using matrix representation:

$$\begin{aligned} y_k &= CA^i x_{k-i} + C \sum_{j=0}^{i-1} A^j G.z_{k-j-1} \\ &+ C(Du_k + \sum_{j=0}^{i-1} A^j Bu_{k-j-1}) + f_k + w_k \end{aligned} \tag{5}$$

Expression (5) is verified for $i = 0$ and $i = 1$. Assuming that the proposal holds for $i$, let us prove that it holds for $i + 1$ also:

$$\begin{aligned} y_k &= CA^{i+1} x_{k-i-1} + C \sum_{j=0}^{i} A^j G.z_{k-j-1} \\ &+ C(Du_k + \sum_{j=0}^{i} A^j Bu_{k-j-1}) + f_k + w_k \end{aligned} \tag{6}$$

We replace $x_{k-i} = Ax_{k-i-1} + G.z_{k-i-1} + Bu_{k-i-1}$ into (5), and derive the following equation:

$$\begin{aligned} & y_k = CA^{i+1} x_{k-i-1} + \\ & C(A^i G.z_{k-i-1} + \sum_{j=0}^{i-1} A^j G.z_{k-j-1}) \\ & + C(A^i Bu_{k-i-1} + Du_k + \sum_{j=0}^{i-1} A^j Bu_{k-j-1}) \\ & + f_k + w_k \end{aligned} \tag{7}$$

By identifying expressions (6) and (7), it is straightforward to prove that the proposal for $i+1$ (equation (6)) holds. This ends the proof. ■
Because the linear dynamic is supposed to be stable, i.e. $A^i$ tends to zero for $i$ sufficiently large which results to

$$CA^i x_{k-i} \to 0 \tag{8}$$

As a consequence, for $i$ sufficiently large, the state influence may be neglected in expression (4). This leads to the following approximated expression of the output $y_k$:

$$\begin{aligned} y_k &\cong C \sum_{j=0}^{i-1} A^j G.z_{k-j-1} \\ &+ C(Du_k + \sum_{j=0}^{i-1} A^j Bu_{k-j-1}) \\ &+ f_k + w_k \end{aligned} \tag{9}$$

The matrix representation of expression (9) is given by:

$$\forall i \geq 0 : y_k = \widetilde{H}_i \overline{z}_{k,i} + \overline{H}_i \overline{u}_{k,i} + f_k + w_k \tag{10}$$

## IV. Data-projection residual generation

In this section, a data-based residual $\epsilon_k$ is generated for sensor fault detection and isolation.

By right-multiplying (Kronecker product) the measurement equation of system (1) by $u_k$, and using the following Kronecker product properties:

- $(Q_1 \otimes Q_2)(Q_3 \otimes Q_4)$=$(Q_1\ Q_3) \otimes (Q_2\ Q_4)$
- $(Q_1 \otimes Q_2)$=$(Q_1 \otimes I)Q_2$

where $Q_1$, $Q_2$, $Q_3$ and $Q_4$ are matrices with appropriate dimensions, we can derive the following expression:

$$\begin{aligned} p_k = (C \otimes I_m)z_k + (D \otimes I_m)q_k + \\ (f_k \otimes u_k) + (w_k \otimes u_k) \end{aligned} \tag{11}$$

where $I_m$ is the identity matrix of dimension $m \times m$ and

$$\begin{cases} z_k = x_k \otimes u_k \\ p_k = y_k \otimes u_k\ . \\ q_k = u_k \otimes u_k \end{cases}$$

Consider an integer $L$, which is chosen such that $L > mi + \ell$. The following subsequent vectors and matrices are introduced:

$$s_k = \begin{bmatrix} p_k \\ q_k \end{bmatrix} = \begin{bmatrix} y_k \otimes u_k \\ u_k \otimes u_k \end{bmatrix} \in R^{(\ell+m)m \times 1}$$

$$\overline{s}_{k,i} = \left[ s_{k-i}^T | \cdots | \, s_{k-2}^T | \, s_{k-1}^T \right]^T \in R^{(\ell+m)mi \times 1}.$$

$$S_k = \left[ \overline{s}_{k-L+1,i} \cdots \overline{s}_{k-1,i}\, \overline{s}_{k,i} \right] \in R^{(\ell+m)mi \times L}.$$

$$Z_k = \left[ \overline{z}_{k-L+1,i} \cdots \overline{z}_{k-1,i}\, \overline{z}_{k,i} \right] \in R^{nmi \times L}.$$

$$\begin{aligned} \overline{f}_{k,i} = \left[ (f_{k-i} \otimes u_{k-i})^T | \cdots | \, (f_{k-2} \otimes u_{k-2})^T | \right. \\ \left. (f_{k-1} \otimes u_{k-1})^T \right]^T \in R^{\ell mi \times 1}. \end{aligned}$$

$$\overline{F}_k = \left[ \overline{f}_{k-L+1,i} \cdots \overline{f}_{k-1,i}\, \overline{f}_{k,i} \right] \in R^{\ell mi \times L}.$$

$$\begin{aligned} \overline{w}_{k,i} = \left[ (w_{k-i} \otimes u_{k-i})^T | \cdots | \, (w_{k-2} \otimes u_{k-2})^T | \right. \\ \left. (w_{k-1} \otimes u_{k-1})^T \right]^T \overline{w}_{k,i} \in R^{\ell mi \times 1}. \end{aligned}$$

$$\overline{W}_k = \left[ \overline{w}_{k-L+1,i} \cdots \overline{w}_{k-1,i}\, \overline{w}_{k,i} \right] \in R^{\ell mi \times L}.$$

$$M_i = \begin{bmatrix} C \otimes I_m & 0_{\ell m \times nm} & \cdots & 0_{\ell m \times nm} \\ 0_{\ell m \times nm} & C \otimes I_m & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0_{\ell m \times nm} \\ 0_{\ell m \times nm} & \cdots & 0_{\ell m \times nm} & C \otimes I_m \end{bmatrix} \in R^{\ell mi \times nmi}.$$

$$K_i = \begin{bmatrix} \left[ I_{\ell m} | \, D \otimes I_m \right] & 0_{\ell m \times (\ell m + m^2)} & \cdots & 0_{\ell m \times (\ell m + m^2)} \\ 0_{\ell m \times (\ell m + m^2)} & \left[ I_{\ell m} | \, D \otimes I_m \right] & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0_{\ell m \times (\ell m + m^2)} \\ 0_{\ell m \times (\ell m + m^2)} & \cdots & 0_{\ell m \times (\ell m + m^2)} & \left[ I_{\ell m} | \, D \otimes I_m \right] \end{bmatrix}$$

where $K_i \in R^{\ell mi \times (\ell+m)mi}$.

By putting the terms dependent on the system input and output on the left side of equality (11), the rest of the terms are put on the right side, the following expression is derived at time $k$:

$$\begin{aligned} \left[ I_{\ell m} | \, D \otimes I_m \right] s_k = \\ (C \otimes I_m)z_k + (f_k \otimes u_k) + (w_k \otimes u_k) \end{aligned} \tag{12}$$

Expression (12) at $k-1$ is given similarly as follows:

$$\begin{aligned} \left[ I_{\ell m} | \, D \otimes I_m \right] s_{k-1} = \\ (C \otimes I_m)z_{k-1} + (f_{k-1} \otimes u_{k-1}) + (w_{k-1} \otimes u_{k-1}) \end{aligned} \tag{13}$$

Equation (13) can be rewritten differently as follows:

$$\begin{aligned} 0_{\ell m \times (\ell m + m^2)} s_{k-i} + \cdots + 0_{\ell m \times (\ell m + m^2)} s_{k-2} + \\ \left[ I_{\ell m} | \, D \otimes I_m \right] s_{k-1} = 0_{\ell m \times nm} z_{k-i} + \cdots + 0_{\ell m \times nm} z_{k-2} \\ + (C \otimes I_m)z_{k-1} + (f_{k-1} \otimes u_{k-1}) + (w_{k-1} \otimes u_{k-1}) \end{aligned} \tag{14}$$

By writing expression (12) at $k-2$

$$\begin{aligned} \left[ I_{\ell m} | \, D \otimes I_m \right] s_{k-2} = \\ (C \otimes I_m)z_{k-2} + (f_{k-2} \otimes u_{k-2}) + (w_{k-2} \otimes u_{k-2}) \end{aligned} \tag{15}$$

Equation (15) can be rewritten differently as follows:

$$\begin{aligned} 0_{\ell m \times (\ell m + m^2)} s_{k-i} + \cdots + 0_{\ell m \times (\ell m + m^2)} s_{k-3} + \\ \left[ I_{\ell m} | \, D \otimes I_m \right] s_{k-2} + 0_{\ell m \times (\ell m + m^2)} s_{k-1} = \\ 0_{\ell m \times nm} z_{k-i} + \cdots + 0_{\ell m \times nm} z_{k-3} + (C \otimes I_m)z_{k-2} + \\ 0_{\ell m \times nm} z_{k-1} + (f_{k-2} \otimes u_{k-2}) + (w_{k-2} \otimes u_{k-2}) \end{aligned} \tag{16}$$

By following the same procedure till $k-i$, the common matrix representation of all the obtained equations is given by:

$$K_i \overline{s}_k = M_i \overline{z}_k + \overline{f}_k + \overline{w}_k \tag{17}$$

By concatenating equation (17) over columns on a time-window of size $L$, we can derive the following equation:

$$K_i S_k = M_i Z_k + \overline{F}_k + \overline{W}_k \tag{18}$$

The derived equation (18) and following theorems will be useful in the sequel. We are now ready to express the main results of our paper as Theorem 4.2 and Proposition 2.

*Theorem 4.1:* If the number of independent rows of the matrix $C$ is equal or greater than the number of independent columns of $C$, then the matrix $M_i$ is left invertible. In other words, it exists a matrix $V_i$ such that the following relation holds:

$$V_i M_i = I_{i\alpha} \tag{19}$$

where $\alpha$ is the number of independent rows of the matrix $C$. If the number of independent rows of the matrix $C$ is equal to the number of independent columns of $C$, then
$V_i = M_i^{-1}$.

If the number of independent rows of the matrix $C$ is greater than the number of independent columns of $C$, then
$V_i = (M_i^T M_i)^{-1} M_i^T$.

*Theorem 4.2:* If $\Gamma \in R^{\cdot \times L}$ is a matrix, where the number of independent rows is equal or less than the number of independent columns, then the right kernel of $\Gamma$ is given by:

$$\Pi_{\Gamma} = I_L - \Gamma^T(\Gamma\Gamma^T)^{-1}\Gamma \in R^{L\times L} \qquad (20)$$

where $\Pi_{\Gamma}$ is the right projection matrix of $\Gamma$, and consequently we have:

$$\Gamma\Pi_{\Gamma} = 0 \qquad (21)$$

Using left invertibility property of $M_i$, and by left multiplying equation (18) by $V_i$, matrix $Z_k$ is given by

$$Z_k = V_i K_i S_k - V_i\overline{F}_k - V_i\overline{W}_k \qquad (22)$$

Right projecting matrix $Z_k$ on the right kernel matrix $\Pi_{\left[\begin{smallmatrix} S_k \\ U_k \end{smallmatrix}\right]}$ of $\begin{bmatrix} S_k \\ U_k \end{bmatrix}$, the equation (22) becomes:

$$Z_k\Pi_{\left[\begin{smallmatrix} S_k \\ U_k \end{smallmatrix}\right]} = -V_i\overline{F}_k\Pi_{\left[\begin{smallmatrix} S_k \\ U_k \end{smallmatrix}\right]} - V_i\overline{W}_k\Pi_{\left[\begin{smallmatrix} S_k \\ U_k \end{smallmatrix}\right]} \qquad (23)$$

where $U_k = \left[\overline{u}_{k-L+1,i}|\cdots|\,\overline{u}_{k-1,i}|\,\overline{u}_{k,i}\right] \in R^{m(i+1)\times L}$.

*Proposition 2:* The proposed parameter-free residual is defined as follows:

$$\epsilon_k = Y_k\Pi_{\left[\begin{smallmatrix} S_k \\ U_k \end{smallmatrix}\right]} \qquad (24)$$

where $Y_k = \left[y_{k-L+1,i}|\cdots|\,y_{k-1,i}|\,y_{k,i}\right] \in R^{\ell\times L}$.

The mathematical expectation of $\epsilon_k$,$r = E[\epsilon_k]$ is used for fault detection.

*Proof:*

By concatenating equation (10) over columns on a time-window of size $L$, we can derive the following equation:

$$\forall i \geq 0 : Y_k = \widetilde{H}_i Z_k + \overline{H}_i U_k + F_k + W_k \qquad (25)$$

where $F_k$ and $W_k$ are constructed similarly as $Y_k$.

1) If there is no sensor fault ($f_k = 0$): [we will prove in this case that $E[\epsilon_k] = 0$]

   By concatenating the output in equation (4) on a time-window of size $L$, the evaluation form of the proposed parameter-free residual is given by:

$$\epsilon_k = Y_k\Pi_{\left[\begin{smallmatrix} S_k \\ U_k \end{smallmatrix}\right]} \cong \widetilde{H}_i Z_k\Pi_{\left[\begin{smallmatrix} S_k \\ U_k \end{smallmatrix}\right]} + W_k\Pi_{\left[\begin{smallmatrix} S_k \\ U_k \end{smallmatrix}\right]} \qquad (26)$$

   where $W_k$ and $F_k$ are constructed similarly as $Y_k$. By replacing (23) into (26), and knowing that $\Pi_{\left[\begin{smallmatrix} S_k \\ U_k \end{smallmatrix}\right]}$ is a right kernel of $U_k$, we get:

$$\epsilon_k \cong -\widetilde{H}_i V_i\overline{W}_k\Pi_{\left[\begin{smallmatrix} S_k \\ U_k \end{smallmatrix}\right]} + W_k\Pi_{\left[\begin{smallmatrix} S_k \\ U_k \end{smallmatrix}\right]} \qquad (27)$$

   The evaluation form (27) of the proposed residual is a linear combination of a centered noise $w$, which implies that $E[\epsilon_k] = 0$ when there is no sensor fault.

2) If there is a sensor fault: [we will prove in this case that $E[\epsilon_k] \neq 0$]

   From equation (27), the evaluation form of the proposed parameter-free residual is given by:

$$\begin{aligned}\epsilon_k \cong{} & -\widetilde{H}_i V_i\overline{F}_k\Pi_{\left[\begin{smallmatrix} S_k \\ U_k \end{smallmatrix}\right]} - \widetilde{H}_i V_i\overline{W}_k\Pi_{\left[\begin{smallmatrix} S_k \\ U_k \end{smallmatrix}\right]} + \\ & W_k\Pi_{\left[\begin{smallmatrix} S_k \\ U_k \end{smallmatrix}\right]} + F_k\Pi_{\left[\begin{smallmatrix} S_k \\ U_k \end{smallmatrix}\right]}\end{aligned} \qquad (28)$$

   Following the same procedure as in the no fault case to get the equation (27), the evaluation form of the mathematical expectation of the proposed (28) parameter-free residual becomes:

$$\begin{aligned}E[\epsilon_k] \cong{} & -E[\widetilde{H}_i V_i\overline{F}_k\Pi_{\left[\begin{smallmatrix} S_k \\ U_k \end{smallmatrix}\right]}] + \\ & E[F_k\Pi_{\left[\begin{smallmatrix} S_k \\ U_k \end{smallmatrix}\right]}]\end{aligned} \neq 0 \qquad (29)$$

   which shows the sensitivity of the mathematical expectation of the proposed residual to sensor faults.

### A. Sensor fault isolability

As shown previously the sensor fault is detectable, in addition to that an important process is to isolate this fault which means the decision on which sensor is in faulty case. To achieve this process we distinguish two cases, when the sensor fault is a constant bias fault at least during the time window of $L + i + 1$ and when it is not.

When the sensor fault is a constant bias fault at least during the time window of $L+i+1$, we have the following expression:

$$-\widetilde{H}_i V_i\overline{F}_k\Pi_{\left[\begin{smallmatrix} S_k \\ U_k \end{smallmatrix}\right]} = 0 \qquad (30)$$

As a result the expression (28) becomes:

$$\epsilon_k \cong -\widetilde{H}_i V_i\overline{W}_k\Pi_{\left[\begin{smallmatrix} S_k \\ U_k \end{smallmatrix}\right]} + W_k\Pi_{\left[\begin{smallmatrix} S_k \\ U_k \end{smallmatrix}\right]} + F_k\Pi_{\left[\begin{smallmatrix} S_k \\ U_k \end{smallmatrix}\right]} \qquad (31)$$

Following the same procedure as in the no fault case to get the equation (27), the evaluation form of the mathematical expectation of the proposed (31) parameter-free residual becomes:

$$E[\epsilon_k] \cong F_k\Pi_{\left[\begin{smallmatrix} S_k \\ U_k \end{smallmatrix}\right]} \qquad (32)$$

which shows that the mathematical expectation of the proposed residual is structured, which means that the first row of $E[\epsilon_k]$ is dedicated for the first sensor, the second row of $E[\epsilon_k]$ is dedicated for the second sensor and so on, in this case only the corresponding mathematical expectation of the residual is not zero when the corresponding sensor is in a faulty case.

If the fault fluctuates instantaneously and $\widetilde{H}_i V_i = 0$, it is the same case as for constant bias sensor faults.

But if the fault fluctuates instantaneously and $\widetilde{H}_i V_i \neq 0$, a more convenient decision algorithm should be developed and this case is not treated in this paper for brevity reasons (for signature table the reader is referred to [20] and [21]).

■

## V. Example and simulation

An electromechanical actuator [19] is used to show the effectiveness of the proposed residual generation method for sensor fault detection and isolation.

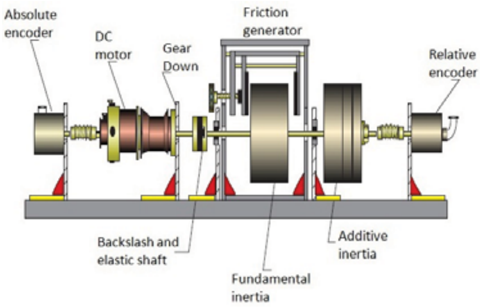


Fig. 1. Electromechanical actuator

This plant may be modeled by a bilinear state-space model:

$$\begin{cases} x_{k+1} = Ax_k + G(x_k \otimes u_k) + Bu_k \\ y_k = Cx_k + Du_k + f_k + w_k \end{cases} \tag{33}$$

with

$$A = \begin{pmatrix} -\frac{R_a}{L_a} & 0 & 0 & 0 \\ 0 & -\frac{F_m}{J_m} & -\frac{k_r}{NJ_m} & 0 \\ 0 & \frac{1}{N} & 0 & -1 \\ 0 & 0 & \frac{k_r}{J_c} & -\frac{F_c}{J_c} \end{pmatrix},$$

$$G = \begin{pmatrix} 0 & 0 & -\frac{k_a}{L_a} & 0\,0\,0\,0\,0 \\ \frac{k_a}{J_m} & 0 & 0 & 0\,0\,0\,0\,0 \\ 0 & 0 & 0 & 0\,0\,0\,0\,0 \\ 0 & 0 & 0 & 0\,0\,0\,0\,0 \end{pmatrix}, \; B = \begin{pmatrix} 0 & \frac{1}{L_a} \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix},$$

$$C = \begin{pmatrix} 1\,0\,0\,0 \\ 0\,0\,0\,1 \end{pmatrix}, \; D = \begin{pmatrix} 0\,0 \\ 0\,0 \\ 0\,0 \\ 0\,0 \end{pmatrix}$$

The model parameters are defined in table 1, these parameters are not used for residuals generation but they are used to simulate the model and to generate output data.

| Parameter | Description | Value | Unit |
|---|---|---|---|
| $T_e$ | sampling time | 0.03 | $[sec]$ |
| $J_m$ | motor shaft inertia | $2.4e-4$ | $[m^2kg]$ |
| $J_c$ | load shaft inertia | 0.0825 | $[m^2kg]$ |
| $F_m$ | motor viscous friction | 0.0032 | $[m^2kg/sec]$ |
| $F_c$ | load viscous friction | 0 | $[m^2kg/sec]$ |
| $k_a$ | motor torque constant | 0.156 | $[m^2kg/sec^2]$ |
| $k_r$ | coupling rigidity coefficient | 37.7 | $[m^2kg/sec^2]$ |
| $R_a$ | motor resistance | 1 | $[\Omega]$ |
| $L_a$ | motor inductance | 0.05 | $[H]$ |
| $N$ | gear ratio | 20 | |
| $i$ | time-window | 16 | |
| $L$ | time-window | 339 | |

Table 1

The 4 states are given in table 2:

| State | Description |
|---|---|
| $i_a$ | armature current |
| $w_m$ | motor shaft velocity |
| $\Delta$ | angular rotation |
| $w_c$ | load shaft angular velocity |

Table 2

The input vector is plotted in Fig.2:

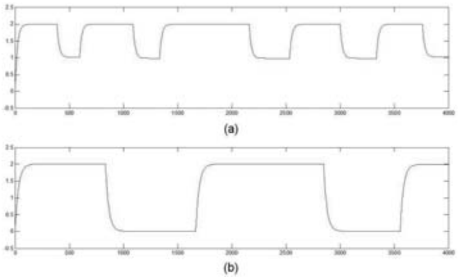


Fig. 2. (a): $u(1,:) = i_e$ is the stator current, (b): $u(2,:) = v_a$ is the armature voltage

The two outputs are plotted in Fig.3:

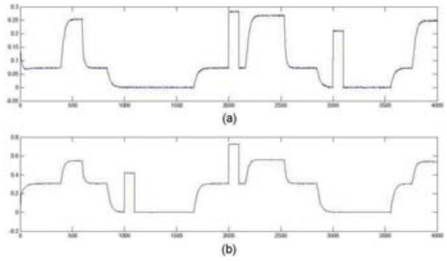


Fig. 3. (a): $y(1,:)$ is the armature current, (b): $y(2,:)$ is the angular velocity

There are two calculated residuals since the number of sensors is 2, the calculated residuals using the proposed method are presented in Fig.4.

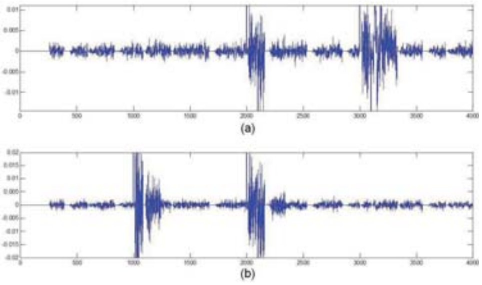


Fig. 4. (a): $\epsilon_k(1)$ is the first residual, (b): $\epsilon_k(2)$ is the second residual

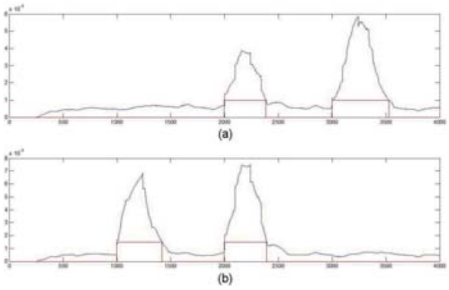


Fig. 5. Blue:(a): $E[\epsilon_k(1)]$, (b): $E[\epsilon_k(2)]$, Red: Thresholding the finite moving average of the proposed residual

In Fig. 5, the blue curve represents the mathematical expectation of the proposed residuals, this mathematical expectation is calculated in a moving time-window of size 339. The red curve represents the result of the decision procedure called Finite Moving Average (FMA) [22] (see the FMA algorithm in Fig. **??**), this is due to the fault sensitivity of the residual mean, which seems to be well dedicated to decide whether there is a fault or not, a good choice of the threshold is needed which can be achieved offline using a healthy database, where the threshold is chosen greater than the maximum value of $E[\epsilon_k]$, for this example the threshold is equal to $0.001$ for the first sensor and $0.0015$ for the second sensor. Moreover, the FDI can be realized if multiple faults occur simultaneously.

## CONCLUSION

A data-projection residual generation method is presented for bilinear systems, where an output matrix is projected on the input right kernel matrix. A new way for input-output matrices construction is proposed to avoid complexity problem of the method presented in paper [15]. The online diagnosis is then easily implemented. Simulation results of an electromechanical actuator show the effectiveness of the proposed method.

## REFERENCES

[1] A.Y. Chow and A. Willsky, "Analytical redundancy and the design of robust failure detection systems", *IEEE Transactions on Automatic Control*, volume 7(29), pp. 603-614, 1984.

[2] R. Isermann, "Process fault detection based on modelling and estimation methods: a survey", *Automatica*, volume 20(4), pp. 403-424, 1984.

[3] R.R. Mohler, "Nonlinear Systems: Applications to Bilinear Control. Prentice Hall", *Automatica*, 1991.

[4] R. Isermann, "Fault diagnosis of machines via parameter estimation and knowledge processing", *Automatica*, volume 29(4), pp. 815-836, 1993.

[5] J. Chen, R. Patton and H. Zhang, "Design of robust structured and directional residuals for fault isolation via unknown input observers", *European Control Conference*, (ECC95), Vol.1, 348-353, Rome, Sept 5-8, 1995.

[6] P.M. Frank and X. Ding, "Survey of robust residual generation and evaluation methods in observer-based fault detection systems", *Journal of Process Control*, volume 7(6), pp. 403-424, 1997.

[7] R. Patton and J. Chen, "Observer-based fault detection and isolation: Robustness and applications", *Control Engineering Practice*, volume 5(5), pp. 671682, 1997.

[8] S. Simani, C. Fantuzzi and R.J. Patton, "Model-based Fault Diagnosis in Dynamic Systems Using Identification Techniques", *Advances in industrial control*, Volume 15, Issue 11, pages 509512, 2005

[9] K.M. Pekpe, G. Mourot, J. Ragot, "Subspace method for sensor fault detection and isolation-application to grinding circuit monitoring", *11th IFAC Symposium on automation in Mining, Mineral and Metal processing*, 2004.

[10] A. Hakem, K.M. Pekpe and V. Cocquempot, "Parameter-free method for switching time estimation and current mode recognition", *Control and Fault-Tolerant Systems*, IEEE SysTol'10, Nice, France, October 6-8, 2010.

[11] A. Hakem, K.M. Pekpe, V. Cocquempot, "On Mode Discernibility and Switching Detectability for Linear Switching Systems using a Data-based Projection Method" , 23rd Chinese Control and Decision Conference, IEEE CCDC, Mianyang, China in May 23-25, 2011.

[12] A.G. Kyusung, Kim. Parlos, "Induction motor fault diagnosis based on neuropredictors and wavelet signal processing", *Mechatronics, IEEE/ASME Transactions on* , volume 7(2), pp. 201, 2002.

[13] M. Basseville, M. Abdelghani and A. Benveniste, "Subspace-based fault detection algorithms for vibration monitoring", *Automatica*, volume 1, pp. 1001-1009, 2000.

[14] M. Ekman, "Bilinear black-box identification and MPC of the activated sludge process", *Journal of Process Control*, volume 18(7-8), pp. 643-653, 2008.

[15] A. Hakem, K.M. Pekpe, V. Cocquempot, "Sensor fault diagnosis for bilinear systems using data-based residuals", *50th Conference on Decision and Control and European Control Conference*, IEEE CDC/ECC 2011, Orlando, Florida, USA, 12-15 December 2011.

[16] V. Venkatasubramanian, R. Rengaswamy, K. Yin and S.N. Kavuri, "A review of process fault detection and diagnosis. Part I: Quantitative model-based methods", *Computers and Chemical Engineering*, 27, 293-311, 2003.

[17] V. Venkatasubramanian, R. Rengaswamy, K. Yin and S.N. Kavuri, "A review of process fault detection and diagnosis. Part III: Process history based methods", *Computers and Chemical Engineering*, 27, 327-346, 2003.

[18] K.M . Pekpe, V. Cocquempot and C. Christophe, "Model-free residual generation for sensor fault detection and isolation in bilinear systems". *7th edition of the multi disciplinary international conference Qualita*, Tanger, Maroc 20-22 March 2007.

[19] M. Zasadzinski, H. Rafaralahy, C. Mechmeche and M. Darouach, "On Disturbance Decoupled Observers for a Class of Bilinear Systems", *Journal of Dynamic Systems, Measurement and Control*, Volume 120, Issue 3, 371, September 1998.

[20] S. CHENIKHER, J. P. CASSAR and K. M. PEKPE, "Fault detection and Isolation from an identified MIMO Takagi-Sugeno model of a bioreactor", Proceedings of the 7th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes Barcelona, Spain, June 30 - July 3, 2009.

[21] J. J. Gertler, "Fault detection and diagnosis in engineering systems", ISBN: 0-8247-9427-3, 1998.

[22] P. R. Bertrand and G. Fleury, "Detecting Small Shift on the Mean By Finite Moving Average", International Journal of Statistics and Management System, Vol. 3, No. 1-2, pp. 56-73, 2008.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# A new reconfigurable fault tolerant control design based on Laguerre series

R.Heidari
Isfahan University of Technology
Isfahan, Iran
Heidari.rahmat@gmail.com

M.Kamali
Isfahan University of Technology
Isfahan, Iran
m.kamaliandani@ec.iut.ac.ir

J. Askari
Isfahan University of Technology
Isfahan, Iran
j-askari@cc.iut.edu.ir

*Abstract*—**In this paper a new design method is presented for reconfigurable fault tolerant control of multivariable nonlinear time delay systems. The control reconfiguration algorithm comprises two main parts: fault detection and diagnosis and model reference adaptive control. First, a combination of generic residual generation method and its Laguerre approximation is proposed due to the problem of residual similarity in fault diagnostic process. Second, a novel approach for adaptive reconfigurable control synthesis is proposed using a nonlinear system identification approach based on the approximation of systems via Laguerre series. The Laguerre-based reconfiguration method is implemented and tested on a time delay form of COSY benchmark model as a preliminary study of reconfigurable control applied to a nonlinear time-delayed model. Actuator faults have been implemented in the COSY benchmark and used to evaluate the control reconfiguration schemes. Simulation results showed acceptable level of fault tolerance.**

***Keywords-Fault tolerant control systems, fault detection and diagnosis, residual generation, model reference adaptive control, Laguerre series.***

## I. Introduction

Fault Tolerant Control Systems (FTCS) design is becoming the key issue in technical systems due to rapidly increasing demands for higher system performance, product quality, productivity and cost efficiency. As it is well-known, FTCSs are control systems that possess the ability of accommodating system component failures automatically. In General, FTCSs can be classified into two types: Passive Fault Tolerant Control Systems (PFTCS) and Active Fault Tolerant Control Systems (AFTCS). The design objective of AFTCS is to 1) design a Fault Detection and Diagnosis (FDD) scheme providing as precise as possible information about a fault and 2) design a new reconfigurable controller in response to and compensating for the fault-induced changes in the system, such that the stability and acceptable close-loop system performance can be maintained.

A successful control system reconfiguration relies on a real-time FDD scheme to provide the most up-to-date information about the system. Model-based methods of fault-detection can be used in presence of input and output signals and applying dynamic process models. These methods are based, e.g., on parameter estimation, parity equations or state observers [1,2]. The goal of fault detection is to generate several symptoms indicating the difference between nominal and faulty status. Based on different symptoms, fault diagnosis procedure follows, determining the fault by applying classification or inference methods. In the FDD module, both the fault parameters and the system state variables need to be estimated online in real time. In the case of nonlinear systems one approach is to linearize nonlinear model using system identification methods. Several fault diagnosis methods have been proposed for systems linearized around an operating equilibrium point (see [3] and [4]). Some of these techniques have also been extended to nonlinear systems. In [5] nonlinear observers are synthesized to detect tank leakage.

The strategy of reconfigurable control is based on inherent information of nominal and faulty systems so as to make these systems consistent in some proper sense. This strategy fits the model following/matching scheme well if we regard the nominal systems as the reference models. Model Reference Adaptive Control (MRAC) is a mathematical approach which can be used to design a model-based reconfigurable control. Research on the application of adaptive methods to reconfigurable control has been extensively studied during past years. Poirot and co-authors in [6] have proposed a framework for constructing adaptive and reconfigurable control system based on the control loop paradigm. In [7] three adaptive algorithms for multivariable model reference control in flight control applications have been compared. A Failure Detection and Identification (FDI) and Adaptive Reconfigurable Control (ARC) procedure based on a decentralized FDI-ARC scheme integrating separate modules for aerospace applications has been carried out by Boskovic and co-authors in [8]. They also have done an adaptive fault tolerant control based on the estimation of damage-related parameter and switching among multiple controllers for flight control in [9]. Furthermore, Boskovic and co-authors proposed a new failure parameterization that models a large class of failures in terms of a single parameter in [10]. In order to accomplish the model following /matching scheme, MRAC is an appropriate approach. A number of investigators have proposed MRAC techniques for nonlinear systems [11], [12] and [13].

In practice, systems with time delays are frequently encountered (e.g., process control systems). Time-delayed linear systems have been intensively investigated in [14]. Lyapunov design has been proven to be an effective tool in controller design for nonlinear systems. Adaptive control has also proven its capability in controlling nonlinear time-delayed

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

systems [15], [16], [17]. The FDD problem in time delay systems has also been investigated recently in [18] and [19].

A common topic that has been gaining interest is nonlinear systems. Model linearization is the key issue in the case of nonlinear systems. The idea we used in this paper is approximation of systems via basis functions. This representation is much more parsimonious than the FIR or IIR ones, but contrary to the transfer function model does not require assumptions about the order and time delay of the process. Although it is implemented in a state-space form, it is straightforward to use in system identification. Furthermore, orthogonal polynomials can be used to make the polynomial coefficients uncorrelated, to minimize the error of approximation, and to minimize the sensitivity of calculations to round-off error. Laguerre models suggested here provide feasibility and stability for the closed-loop system. Laguerre lattice filter identifier counts for 1) the identification of the system's dynamics, that are mapped into the filter's structure through the utilization of estimated reflection coefficients and 2) the system's order identification. In [20] the usage of Laguerre parameterization of input sequences in model predictive control has been investigated. Also application of Volterra series to the modeling of static and dynamic nonlinear systems is investigated in [21]. A model predictive control strategy based on Laguerre functions has also been proposed in [22].

This paper looks at the problem of reconfigurable fault tolerant control systems in time-delayed nonlinear systems using Laguerre series in case that system is noise free and disturbances are dispensable. This work on the reconfigurable control synthesis is under assumption that the faulty system information is known by using Laguerre model. After achieving Laguerre model, MRAC is applied on the basis of this linearized model to have better control. At next step, in order to have an accurate FDD, we suggest a novel approach to the problem of residual similarity. A combination of residual generation methods and parameter estimation using Laguerre series would simplify decision making. We employed this concept to generate more detailed residuals. However, because increasing residual signals may leads to diagnosis complexity, classification methods will be helpful. The effectiveness of the Laguerre-based MRAC has been shown via simulation results on a time delay form of COSY benchmark.

## II. Problem Statement

Let us consider the following nonlinear multivariable system subjected to input delay

$$\begin{aligned}\dot{X}(t) &= f_1(X(t), U(t-\tau), F(t), d(t), w(t))\\ Y(t) &= f_2(X(t), U(t-\tau), F(t), d(t), w(t)).\end{aligned} \tag{1}$$

in which $X \in \mathcal{R}^N$ , $U \in \mathcal{R}^m$ and $Y \in \mathcal{R}^N$ are state vector, input and output signals respectively, $\tau \in \mathcal{R}^+$ is the known input delay , $F_{1\times nf}$ is the vector of occurred faults (nf is the number of faults), $d_{1\times nd}$ is the disturbance vector and $w_{1\times nw}$ represents uncertainty or noise vector.

We can approximate (1) using an Nth order Laguerre model

$$\begin{aligned}Y(t) = &\sum_{i=1}^{N} C_i(f)L_i(q,\alpha)U(t)\\ &+\sum_{i=1}^{N} d_i(f)L_i(q,\alpha)Y(t)\end{aligned} \tag{2}$$

where

$$L_i(q,\alpha) = \frac{\sqrt{1-\alpha^2}}{q-\alpha}\left(\frac{1-\alpha q}{q-\alpha}\right)^{i-1} \qquad i = 1,2,\dots, \tag{3}$$

$|\alpha| < 1$ is Laguerre parameter and Laguerre coefficients $C_i(f)$ and $d_i(f)$ could be calculated using recursive least square algorithm. For complicated systems, high orders of Laguerre polynomials give a simple realization of the system and would decrease the approximation error.

Assuming $d(t) = w(t) = 0$ - the system is noise free and disturbances are dispensable - we propose the MRAC scheme for multivariable plant model

$$Y(t) = C(f)L(q,\alpha)\, U(t) \tag{4}$$

where$Y(t) \in \mathcal{R}^N$ and $U(t) \in \mathcal{R}^N$ are output and control input vectors. Moreover

$$L(q,\alpha) = \begin{bmatrix} L_1(q,\alpha) & 0 & 0\\ \vdots & \vdots & \vdots\\ L_p(q,\alpha) & 0 & 0\\ 0 & L_1(q,\alpha) & 0\\ \vdots & \vdots & \vdots\\ 0 & L_p(q,\alpha) & 0\\ 0 & 0 & 0\\ \vdots & \vdots & \vdots\\ \vdots & \vdots & L_1(q,\alpha)\\ \vdots & \vdots & \vdots\\ 0 & 0 & L_p(q,\alpha)\end{bmatrix}_{np\times n} \tag{5}$$

is Laguerre matrix and

$$C(f)_{n\times np} = \begin{bmatrix} C_{111}(f) & \dots & C_{1p1}(f) & \cdots & C_{n11}(f) & \dots & C_{np1}(f)\\ & & \vdots & \ddots & & & \vdots\\ C_{11n}(f) & \dots & C_{1pn}(f) & \cdots & C_{n1n}(f) & \dots & C_{npn}(f)\end{bmatrix}_{n\times np} \tag{6}$$

is Laguerre coefficients matrix under the effect of fault .

The problem is to design an adaptive reconfigurable fault tolerant controller for this system such that all signals of the closed-loop system are bounded and the plant output $y(t)$, asymptotically exact, follows the output $Y_m(t)$ of the reference model

$$Y_m(t) = W_m(s)r = C_m L(q,\alpha)\, r \tag{7}$$

where $W_m(s) \in \mathcal{R}^{N\times N}$ is a stable rational transfer matrix, $r \in \mathcal{R}^N$ is a bounded reference input signal and $C_m$ is Laguerre coefficients matrix.

Let us write (4) as

$$Y_P = G(s)U_p \tag{8}$$

where $G(s) = C(f)L(q, \alpha)$ is the transfer matrix of the linear system. It should be noted that since the relative degree of the Laguerre model is $n^* = 1$

$$\mathrm{L_i}(\mathrm{s}, \alpha) = \frac{\sqrt{2\alpha}}{\mathrm{s} + \alpha}\left(\frac{\mathrm{s} - \alpha}{\mathrm{s} + \alpha}\right)^{\mathrm{i}-1} \tag{9}$$

the relative degree of $\mathrm{G(s)}$ is also $\mathrm{n}^* = 1$.

To meet the control objective the following assumptions are made about $\mathrm{G(s)}$[23]:

(A1): $\mathrm{G(s)}$ is strictly proper and has full rank.

(A2): The transmission of zeros of $\mathrm{G(s)}$ have negative real parts.

(A3): An upper bound $\bar{\nu}_0$ on the observability index $\nu_0$ of $\mathrm{G(s)}$ is known.

(A4): There exists a known matrix $\mathrm{S_p(f)}$ such that $\mathrm{K_p(f)\ S_p(f)} = (\mathrm{K_p(f)S_p(f)})^{\mathrm{T}} > 0$.

In (A4), the high frequency gain matrix $K_p(f)$ of the model is defined as

$$\mathrm{K_p(f)} = \mathrm{C_p(f)\ C_m}^{-1} \tag{10}$$

which is finite and nonsingular. It should be noted that the system dynamics may have some inputs which are not control inputs, but only play a leading role in modeling. In this situations, to achieve appropriate $\mathrm{K_p(f)}$ , matrixes $\mathrm{C_p(f)}$ and $\mathrm{C_m}^{-1}$ must be defined such that all of their elements refer to control inputs.

Without loss of generality, it is assumed that the transfer function of the reference model is diagonal SPR and of the form

$$\mathrm{W_m(s)} = \mathrm{diag}\left[\frac{1}{\mathrm{s} + \mathrm{a_{ri}}}\right], \mathrm{a_{ri}} > 0, i = 1, \dots, N \tag{11}$$

## III. FAULT DETECTION AND DIAGNOSIS METHOD

Early fault detection, which reduces the possibility of catastrophic damage, is possible by comparing the measured signals with a database that contains characteristic signals for machines operating with or without faulty conditions. To obtain the most up-to-date and comprehending database for fault diagnosis simplicity, we need to consider all possible conditions such as types and severities of faults.

In some cases, decision making is difficult because of similarly behaved residuals resulted from faults with different types or different severities. This is the motivation to search more residuals to distinguish these faults. Hence, a combination of residual generation approaches are suggested to overcome this weakness and to make the database better suited for the FDD. The new approach proposed in this paper, combines residual generation methods with parameter estimation using Laguerre series. In fact, residuals from basic FDD method which are weak to have a good decision are expanded via Laguerre series which results in Laguerre coefficients. These coefficients can be used as supplementary residuals to have a well-informed database and to facilitate the decision making easier. This approach is formulized as follows. Let

$$\mathrm{Residue(t)} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{r(i,q)\ Residue(t-i)} \tag{12}$$

be the expansion of resulting residuals from basic FDD method where q, $\mathrm{r(i,q)}$ and n denote respectively delay operator, coefficients and model order. Coefficients can be expanded in Laguerre model

$$\mathrm{r(i,q)} = \sum_{\mathrm{j}=1}^{\mathrm{p}} \mathrm{a_{ji}L_j}(\mathrm{q}, \alpha) \tag{13}$$

which leads to

$$\mathrm{Residue(t)} = \sum_{\mathrm{i}=1}^{\mathrm{n}}\sum_{\mathrm{j}=1}^{\mathrm{p}} \mathrm{a_{ji}L_j}(\mathrm{q}, \alpha)\ \mathrm{Residue(t-i)} \tag{14}$$

resulting in Laguerre coefficients $\mathrm{a_{ji}}'\mathrm{s}$ which can be used as supplementary residuals to make an easier fault detection.

Reconfigurable controller

In general, the existing reconfigurable controller design methods fall into one of the following approaches: Linear quadratic regulator, Eigen structure assignment, multiple-model, adaptive control, pseudo-inverse, perfect model following and feedback linearization [22]. Our objective is to design a reconfigurable control strategy to accommodate system component faults.

In this paper, a nonlinear MRAC strategy based on Laguerre model is presented. A Laguerre model is used to synthesize a linear controller that provides a satisfactory closed-loop performance near the nominal operating point.

The proposed control law is considered as

$$\mathrm{U} = \theta_1^{*\mathrm{T}}\mathrm{w_1} + \theta_2^{*\mathrm{T}}\mathrm{w_2} + \theta_3^*\mathrm{r} = \theta^{*\mathrm{T}}\mathrm{w} \tag{15}$$

where

$$\theta^{*\mathrm{T}} = \left[\theta_1^{*\mathrm{T}} \quad \theta_2^{*\mathrm{T}} \quad \theta_3^*\right], \tag{16}$$

$$\mathrm{w} = [\mathrm{w_1}^{\mathrm{T}} \quad \mathrm{w_2}^{\mathrm{T}} \quad \mathrm{r}^{\mathrm{T}}]^{\mathrm{T}}, \tag{17}$$

$$\mathrm{w_1} = \frac{\mathrm{A(s)}}{\Lambda(\mathrm{s})}\mathrm{U} \quad , \quad \mathrm{w_2} = \frac{\mathrm{A(s)}}{\Lambda(\mathrm{s})}\mathrm{Y}, \tag{18}$$

$$\mathrm{A(s)} = [\mathrm{Is}^{\bar{\nu}_0 - 1} \quad \mathrm{Is}^{\bar{\nu}_0 - 2} \quad \dots \quad \mathrm{Is} \quad \mathrm{I}]^{\mathrm{T}} \tag{19}$$

and $\Lambda(\mathrm{s})$ is a monic Hurwitz polynomial of degree $\bar{\nu}_0$ (A3).

**Proposition1**: Consider the system (1) with the assumptions (A1) – (A4) and the reference model (7). Then the adaptive control (15) with update laws

$$\dot{\theta}^{\mathrm{T}} = -\Gamma_{\mathrm{p}}(\mathrm{f})\epsilon\,\mathrm{w}^{\mathrm{T}} \tag{20}$$

assures that all closed-loop signals are bounded and the tracking error $\mathrm{e_1} = \mathrm{Y_p} - \mathrm{Y_m}$ converges to zero asymptotically. ■

In equation (20) $\epsilon = Y(t) - D_p(f) L(q, \alpha)\, \hat{\theta}^{*T} w$ is the output error, $Y(t) = D_p(f)\, L(q, \alpha)\, U(t)$ and $\hat{Y}_f(t) = D_p(f)\, L(q, \alpha)\, \hat{\theta}^{*T} w$ are output vector and its estimate, respectively.

## IV. Simulation Study

A time delay form of the COSY benchmark is considered for simulation studies. The benchmark is composed of three identical tanks with identical area of cross section. The measured variables are the levels of the first tank($h_1$), second tank ($h_2$) and third tank ($h_3$) respectively. Inflows of the first tank ($q_{in1}$) and third tank ($q_{in2}$) are chosen as known inputs. Inflows vary between 0 and 1 and the status of inter connecting pipes switches between 0 and 1, therefore the system can be considered as a hybrid system [4].

When the valve $V_{12}$ operates with delay $T = 10$ seconds, the objective is to control levels of the first and second tanks. The third tank is hardware redundant which may be used in the presence of faults. Control inputs are Inflow of first tank $q_{in1}$ and status of main inter connecting pipe $V_{12}$.

COSY benchmark without time delay has been investigated by many researchers [25]. But in order to apply Proposition1 to its time delay form, the nonlinear model is linearized via Laguerre series with sampling time $T_s = 0.1$ seconds.

### A. FDD method

In this paper the preliminary FDD approach considered is parity equations. A straightforward model-based method of fault detection is to take fix model $G_m$ and run it parallel to the process, thereby forming an output error [1]

$$r(t) = [G_P(s) - G_m(s)]u(t) \tag{21}$$

Since the modeling approach is system identification using Laguerre series, the parity equation approach would follow

$$r(t) = y(t) - \sum_{i=0}^{i=N} C_i(\alpha) L_i(z, \alpha)\, u(t). \tag{22}$$

Two main residuals used in our work are

$$residue1(t) = h_1(t) - \sum_{i=0}^{i=2} C_i(\alpha) L_i(z, \alpha)\, u_1(t) \tag{23}$$

and

$$residue2(t) = h_2(t) - \sum_{i=0}^{i=2} C_i(\alpha) L_i(z, \alpha)\, u_2(t). \tag{24}$$

When faults occur with similar residuals, Laguerre model of order 4 is used to derive Laguerre coefficients of residuals (output errors) to obtain secondary residuals. Preliminary residuals are expanded as

$$Residue1(t) = \sum_{i=1}^{4} a_{1i} L_i(q, \alpha)\, Residue1(t-1) \tag{25}$$

and

$$Residue2(t) = \sum_{i=1}^{4} a_{2i} L_i(q, \alpha)\, Residue2(t-1). \tag{26}$$

The difference between these coefficients ($a_{ji}$) and those of normal condition ($a_{jiN}$)

$$e_{ji} = a_{ji} - a_{jiN}. \tag{27}$$

are used as supplementary residuals.

For diagnosis, classification methods such as Artificial Neural Networks, Fuzzy clustering, Rule-based reasoning and Statistical methods are widely used [2]. The rule-based reasoning method is the one selected for this paper.

### B. Case one: Normal operation

We first start our study by considering the Laguerre-based MRAC strategy in the case with no faults. Hence we simulated the three tank model with $V_{23}(t) = 0$ , which means that the third tank is not under control. The resulting response is shown in Fig.1.

Leakage in first tank and blockage of main pipe between first and second tank are two common kinds of faults considered in three tank system. The severity of these faults can also be an important factor to take a proper action. Here we simulate blockage fault to show how severity of the fault affect decision making.

### C. Case Two: Blockage Fault Occurrence

Next, we simulate the system by assuming that the inter-connecting pipe $V_{12}$ is blocked completely. Responses of $h_1$ and $h_2$ in this case are shown in Fig. 2. Blocking $V_{12}$ results in an undesirable response, which emphasizes the necessity of control reconfiguration. In this situation the third tank enters as a hardware redundancy. The above fault can be easily detected by means of two preliminary residuals. The resulting responses are shown in Fig.3. Fig.4 also shows residuals. In this case to detect blockage we have no problem using these residuals; but when partial blockage occurs, the residuals would be as shown in Fig.5. As it is clear, it is somehow difficult to detect the right fault from difference between Fig.4and Fig.5. Therefore, supplementary residuals $e_{ji}$ shown in Fig.6-9 would be helpful.

## V. Conclusions

In this paper, we employed the concept of parameter estimation based on Laguerre series to solve the problem of reconfigurable fault tolerant control of nonlinear time delay systems. To achieve fast and reliable fault detection, diagnosis and correct reconfiguration, a new parameter estimation method using Laguerre series has been proposed and is combined with other residuals generation approaches. For this purpose, a model reference adaptive fault tolerant control based on Laguerre model for nonlinear time delay systems is proposed. Simulation results for a time delay form of COSY benchmark system in the presence of blockage fault has shown the effectiveness of proposed Laguerre-based FDD-FTCS approach. Note that, the proposed method is also simulated for the leakage and the other faults. Although the obtained simulation results show again its effectiveness, but these simulation results have been omitted due to the lack of space.

However, the robustness of the proposed method to parameter uncertainty and/or its extension for real-world case will be investigated in our future works.

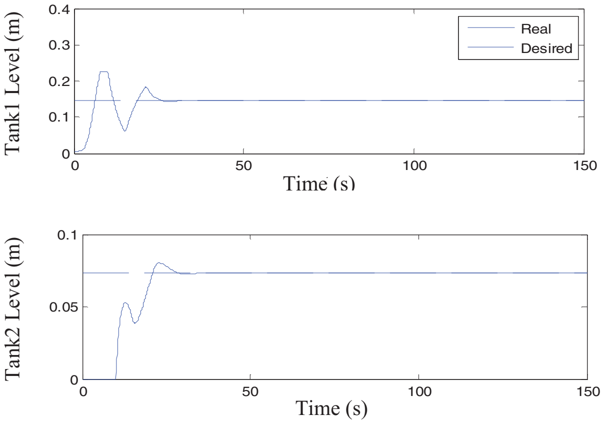


Figure 1. Output responses in normal operation

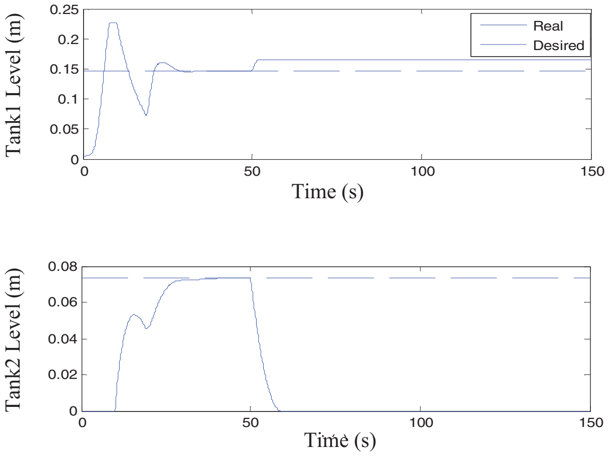
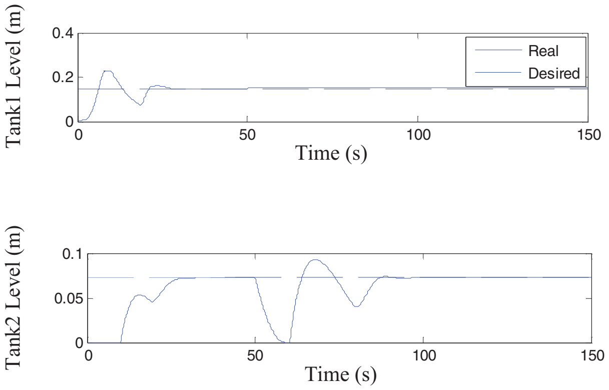


Figure 2. Uncontrolled output responses in presence of Blockage



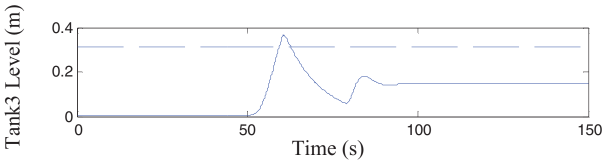


Figure 3. Controlled output responses in presence of Blockage

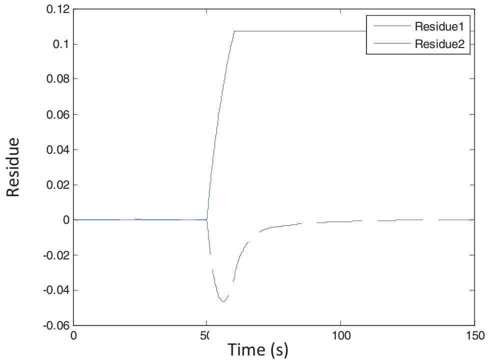


Figure 4. Residuals (output errors) in presence of full Blockage

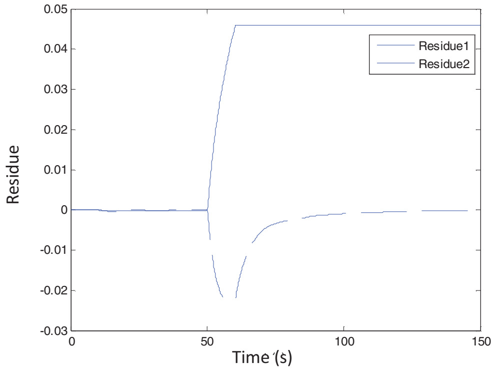


Figure 5. Residuals (output errors) in presence of partial Blockage

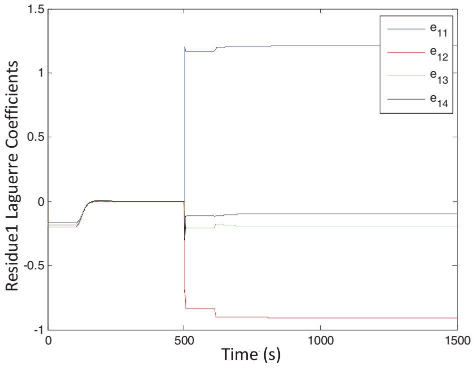


Figure 6. Laguerre coefficients of first output error in presence of full Blockage

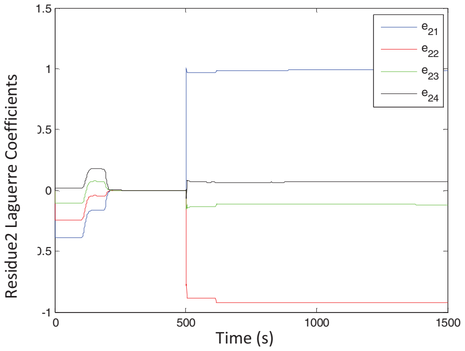


Figure 7. Laguerre coefficients of second output error in presence of full Blockage

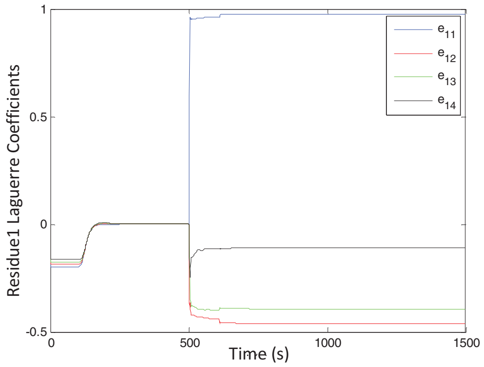


Figure 8. Laguerre coefficients of first output error in presence of partial Blockage

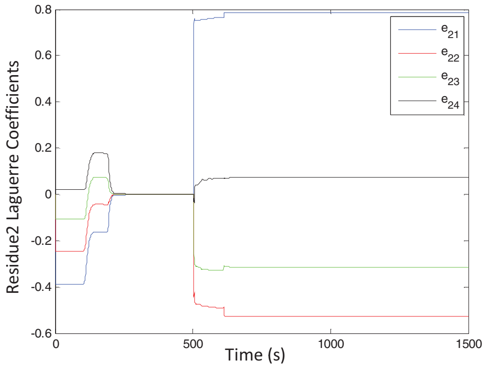


Figure 9. Laguerre coefficients of second output error in presence of partial Blockage

## References

[1] R. Isermann, "Model-based fault-detection and diagnosis – status and applications", Annual Reviews in Control, Vol. 29, pp. 71-85, 2005.

[2] S. Katipamula and M. R. Brambley, "Methods for fault detection, diagnostics, and prognostics for building systems— A review, part I", HVAC&R research, Vol. 11, No. 1, pp. 3-25, 2005.

[3] A. Akhenak, M. Chadli, D. Maquin and J. Ragot, "State estimation via multiple observer the three tank system", Proc. 5th IFAC Safeprocess, 2003.

[4] D. Theilliol, H. Noura and J.-C. Ponsart, "Fault diagnosis and accommodation of a three-tank system based on analytical redundancy", ISA Transactions, 41, 2002.

[5] D. Shields, S. Du, "An assessment of fault detection methods for a benchmark system", Proc. 4th IFAC Safeprocess, 2000.

[6] P. E. Poirot, J. Nogiec and S. Ren, "A Framework for Constructing Adaptive and Reconfigurable Systems", IEEE Trans. on Nuclear Science, Vol. 55, No. 1, pp. 284-289, 2008.

[7] M. Bodson and J. E. Groszkiewicz, "Multivariable Adaptive Algorithms for Reconfigurable Flight Control", IEEE transaction on control systems technology, Vol. 5, No. 2, pp. 217-229 , 1997.

[8] J. D. BoSkoviC, M. Gopinathan and R. K. Mehra, "Computer Aided Design of Failure Detection and Identification and Adaptive Reconfigurable Control Systems for Aerospace Applications", IEEE Proc. of International Symposium on Computer Aided Control System Design, pp. 188-193, 1999.

[9] J. D. BoSkoviC, S. M. Lee and R. K. Mehra, "Reconfigurable Flight Control Design Using Multiple Switching Controllers and On-line Estimation of Damage-Related Parameters", IEEE Proc. of International Conference on Control Applications, pp. 479-484, 2000.

[10] J. D. BoSkoviC, J. Redding and R. K. Mehra, "Robust Fault-Tolerant Flight Control using a New Failure Parameterization", IEEE Proc. of American Control Conference, pp. 5753-5758, 2007.

[11] W. Yu and M. Delasen, "A new adaptive control scheme with arbitrary nonlinear inputs", International Journal of systems science, Vol 29, No. 4, pp. 407-417, 1998.

[12] M. Bohm, M. A. Demetriou, S. Reich and I. G. Rosen," Model reference Adaptive control of distributed parameter systems", SIAM Journal of Control Optimization, Vol. 36, No. 1, pp. 33-81, 1998.

[13] R. B. McLain AND M. A. Henson, "Nonlinear Model Reference Adaptive Control with Embedded Linear Models", Industrial Engineering Chemistry Research, Vol. 39, No. 8, pp. 3007-3017, 2000.

[14] V. B. Kolmanovskii, S. I. Niculescu and K. Gu, "Delay effects on stability: A survey" IEEE Proc. of decision and control, pp. 1993–1998, 1999.

[15] S. S. Ge, F. Hong and T. H. Lee, "Robust adaptive control of nonlinear systems with unknown time delays", Automatica, pp. 1181–1190, 2005.

[16] S. Xu and G. Feng, "Further results on adaptive robust control of uncertain time-delay systems", IET Control Theory and Applications, Vol. 2, No. 5, pp. 402–408, 2008.

[17] H. Wu, "Adaptive robust control of uncertain dynamical systems with multiple time-varying delays", IET Control Theory and Applications, Vol. 4, No. 9, pp. 1775–1784, 2010.

[18] Y. Wang, S. X. Ding, H. Ye and G. Wang, "A New Fault Detection Scheme for Networked Control Systems Subject to Uncertain Time-Varying Delay ",IEEE Trans. on Signal Processing, Vol. 56, No. 10, pp. 5258-5268, 2008.

[19] N. Meskin and k. Khorasani, "Fault Detection and Isolation of Distributed Time-Delay Systems", IEEE Trans. on Automatic Control, Vol. 54, No. 11, pp. 2680-2685, 2009.

[20] G. Valencia-palomo and J.A. Rossiter, "Using Laguerre functions to improve efficiency of multi-parametric predictive control", IEEE Proc. of American Control Conference, pp. 4731-4736, 2010.

[21] H. Moodi and D. Bustan, "On Identification of Nonlinear Systems Using Volterra Kernels Expansion on Laguerre and Wavelet Function", IEEE Proc. of Chinese Control and Decision Conference, pp. 1141-1145, 2010.

[22] Y. Feng, L. Wang and W. Luo, "Laguerre Functions based Nonlinear Model Predictive Control using Multi-Model Approach", 34th Annual Conference of IEEE, pp. 247-252, 2008.

[23] P.A. Ioannou and J. Sun, Robust adaptive control, Prentice-Hall, New Jersey, 2000.

[24] Y. Zhang, and J. Jiang, " Bibliographical Review on Reconfigurable Fault-Tolerant Control systems", Proc. of the 5th IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes, pp. 265-276, 2003.

[25] K. J. Astrom, P. Albertos, M. Blanke, A. Isidori, R. Sanz and W. Schaufelberger, Control of Complex Systems, Springer Verlag, London, 2001.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Model of a Fusion Cryopumping System for Condition Monitoring

Nick Wright
School of Electrical, Electronic
and Systems Engineering
Loughborough University
Loughborough, Leics LE11 3TU
Email: n.wright@lboro.ac.uk

Roger Dixon
School of Electrical, Electronic
and Systems Engineering
Loughborough University
Loughborough, Leics LE11 3TU
Email: r.dixon@lboro.ac.uk

Roel Verhoeven
Culham Centre for Fusion Energy
Culham Science Centre
Oxfordshire OX14 3DB
Email: roel.verhoeven@ccfe.ac.uk

***Abstract*—Cryopumping systems provide an essential function in most magnetic confinement nuclear fusion experiments. The maintenance of an ultra-high vacuum in a magnetic confinement vessel is required for experiments to be conducted, without exception, and cryopumping systems are widely used to achieve this. As such, the availability of this type of nuclear fusion experiment depends in part upon the availability of its vacuum system. In order to reduce experimental time lost to unplanned maintenance, investigation and avoidable failures, a condition monitoring scheme targeted on a cryopumping vacuum system is proposed. A model of the cryopumping system deployed on the Joint European Torus is presented. The model is supported by a first principles derivation and is validated using historical data. Its application to a condition monitoring scheme is discussed. This paper contributes to the wider nuclear fusion development programme by addressing a key maintenance and reliability issue, which is an important step on the road to commercial fusion energy.**

***Index Terms*—Condition Monitoring, Cryopumping, Dynamic Modelling, Two Phase Systems**

## I. Introduction

All nuclear fusion experiments using magnetic confinement techniques rely on the maintenance of an ultra-high vacuum inside the reaction vessel. This includes all the other vessels with which it shares an atmosphere. Any event that causes a disruption or loss of vacuum will result in the experiment being stopped while the vacuum is regenerated and the cause of the event mitigated. At the leading nuclear fusion experiment, the Joint European Torus (JET), and indeed at most others, loss of vacuum events and disruptions to the vacuum pumping system are identified, diagnosed and isolated manually. This process can be lengthy, requiring many man hours of labour. Time expended on maintenance is a significant burden, and as such there is a strong motivation for its reduction. Condition monitoring techniques have been used successfully in many applications, and it is proposed that such a condition monitoring scheme would be a very useful tool for the engineering team at JET. Historically, most disruptive and loss of vacuum events have occurred in the neutral beam heating systems, therefore a condition monitoring scheme focused on this area would be useful. In this paper we discuss a model of the cryopumping system deployed in the neutral beam heating devices at JET and how it can be used in the development of a condition monitoring scheme.

The neutral injection box (NIB) cryopump can be considered to be a two phase system. Two excellent examples of modelling two phase systems can be found in T Phillip's book on modelling and simulation[1] and in KJ Astrom's paper on drum boiler dynamics[2]. The novel model presented here is inspired by these examples. G Duesing's 1987 paper[3] is often referred to in discussions of neutral beam injection systems, and more recently, Ciric[4] and Ciric et al[5] wrote about recent developments in JET's NIB cryopumping scheme. These sources, in addition to design documentation available from CCFE, provide detailed information on the system discussed here. For information on cryogenic and vacuum systems in general, R Barron's book[6], WM Rohsenow's comprehensive book[7] and DJ Hucknall's book[8] are all excellent sources. A two phase model of a cryopumping system, however, has not yet been presented in literature. Using sources together has allowed us to develop a novel analytical model of the JET NIB cryopumping system which can serve both as tool for the design of a condition monitoring scheme and simulating faults.

This paper is split into four sections. The second section describes the physical system with which we are concerned. A justification of the selection of the target subsystem is provided, and its important features are discussed. The third section goes through the analytical model of the NIB cryopumping system and the fourth section presents the a result of a simulation of the model compared to historical data. The final section summarises the key information presented in this paper and notes its future application to condition monitoring.

## II. The Physical Cryopump

Cryopumping reduces the pressure inside a vacuum vessel by condensing gasses onto very cold surfaces. There are several ways to achieve this; RA Haefer describes some in his 1989 book[9]. The NIB cryopumping system at JET works by cooling a series of extruded aluminium surfaces inset with capillaries with cryogenic fluid. Each NIB has ten cryopump elements on each side of the vessel. Fig. 1 is a plan view illustration of a single cryopump element.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

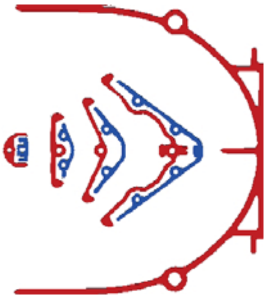


Fig. 1. A single cryopump element

Each cryopump element contains two types of surface: nitrogen surfaces and helium surfaces. The nitrogen surfaces are cooled to around 77K using liquid nitrogen. Similarly, the helium surfaces are cooled to around 4.2K, using liquid helium. In Fig. 1, the nitrogen surfaces have been drawn in red, the helium surfaces in blue. The purpose of the nitrogen surfaces is to shield the helium surfaces from thermal radiation and to pre-cool gas particles. The helium panels condense the majority of the gas. The capillaries embedded in the helium panels are filled from the bottom, and on the top they are tied-off onto a horizontal manifold and phase separator assembly.

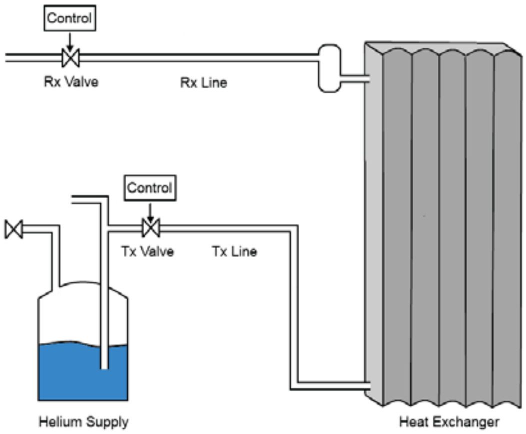


Fig. 2. Illustration of the NIB helium loop

Fig. 2 is a schematic of the NIB helium loop. The heat exchanger represents the two cryopumping walls. Also shown are the fluid/gas transmission and return line (labelled Tx and Rx lines), the transmission and return valves (labelled Tx and Rx valves), and the helium supply tank. The system boundaries are the helium source from which the helium tank is filled, and the helium return valve box, where gaseous helium evolved from the heat exchanger is diverted to either the helium liquefiers or helium collection balloons.

The helium transmission and return lines are part of the same assembly. Liquid helium, liquid nitrogen, gaseous helium and gaseous nitrogen are all carried in a single transmission line, comprised of several concentric cylinders. Together with a vacuum jacket, the cylinders are arranged in such a way as to minimise thermal losses. Between the helium supply tank and the helium return valve box are two valves: the helium supply and return valves. Each valve is designed to minimise thermal losses. They act as control valves, each with a standard feedback PI scheme. The control variable for the supply valve is the level of fluid in the phase separator and the control variable for the return valve is the (absolute) pressure in the helium return line. The helium supply tank is periodically filled, according to the fluid level. Helium inside the tank is distributed to four places: the two NIBs, a cooling system in the supply valve box and to a helium liquefier (to deal with tank losses). Helium inside the tank is maintained at a constant pressure.

A list of measured process variables on the helium loop is presented in Table I. While each of these process variables is available in real time, a recording is only made once an unscaled variable deviates two percent from its previously recorded value.

| Process variable | Unit |
|---|---|
| Phase separator level | % |
| Supply & return valve position | % |
| Supply & return line pressure | BarA and BarG |
| Capillary delta pressure | mBar |
| Helium supply temperature | K |
| Helium return temperature | K |
| Helium tank fill level | % |
| Helium tank fill volume | l |

TABLE I
LIST OF MEASURED PROCESS VARIABLES

The specific (latent) heat capacity of liquid helium is low and the helium panels are sensitive to heat load. This, in addition to the high level of instrumentation on the helium panel support system (the helium loop), means that information about the state of the NIB vacuum can be deduced from examination of the state of the helium loop. Furthermore, historically, the helium loop has experienced more maintenance issues than the nitrogen loop, and several (potential) loss of vacuum events relating to faults in the helium line were identified in a simplified FMEA process. For these reasons, it was decided that the initial focus for the condition monitoring scheme should be on the helium loop. It should also be noted that the cryopump has several operational modes. The mode focused on here is the “full cooldown” mode, which the pumping system is set to during regular operation (i.e. Most of the time outside of scheduled maintenance periods).

## III. The Model Cryopump

The mathematical model of the helium loop in the cryopumping system is split into nine component models, roughly corresponding with the components depicted in Fig. 2. In order to provide structure to the model, the component models have been categorised as either storage or resistive components, allowing a common interface between them and simplifying their analysis. Specifically, the storage components have a pressure associated with them; the resistive components, a flow rate. Fig. 3 is an illustration of this structure. A description of each of these blocks is presented below, starting with the supporting blocks and moving on to the main block, the heat exchanger.

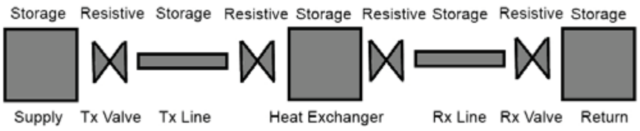


Fig. 3. Model structure

### A. Supporting Blocks

Starting with the supply and return (boundary) blocks, the following assumptions were made:

(a) The helium tank is always refilled at the appropriate times
(b) When the tank is being refilled, it is refilled at a constant, uninterrupted rate
(c) The pressure on the return side of the return valve is controlled and constant
(d) The helium supply pressure and temperature is constant
(e) The helium consumption in both NIBs is similar and converges with time
(f) Helium tank losses (and hence the amount of material sent to the liquefier) are negligible

Given these assumptions, the following equation described the the helium tank fluid volume:

$$\frac{dV_{ht}}{dt} = K_{kl}q_{tf} - K_{kl}\left(q_{sc} + q_{ts} + 2q_{tl}\right) \tag{1}$$

where $V_{ht}$ is the fluid volume in the helium tank, $q_{tf}$ is the helium tank fill rate, $q_{sc}$ is the flow rate to the valve box subcooler, $q_{ts}$ is the flow rate of tank losses, and $q_{tl}$ is the flow rate to the (single) NIB transmission line.

The transmission and return control valves are treated independently, because the former controls the flow of a (relatively, compared to helium gas) incompressible fluid and the latter, compressible helium gas. They are both resistive components. The flow rate through the transmission valve is described by:

$$q_{fl} = Y_{tx}C_v\sqrt{\frac{\Delta p}{v_i}} \tag{2}$$

where $C_v$ is the valve conductance, $\Delta p$ is the differential pressure across the valve, $Y_{tx}$ is the proportion the valve is open, normalised between zero and unity, and $v_i$ is the specific volume of the fluid at the valve inlet.

The return valve is described using an equation for valves transmitting compressible gasses referred to by Baumann[10]:

$$q_{rl} = Y_{rx}C_v 3.22\sqrt{\Delta p\left(p_1 + p_2\right)G_g} \tag{3}$$

where, with care to use US units for all the terms and converting afterwards, owing to the empirical scaling factor, $q_{rl}$ is the return valve flow rate and $G_g$ is the specific gravity of the gas.

The transmission lines are storage components. They are assumed to be of fixed volume, regardless of pressure (historically, pressure deviates no more than 10% from its mean). For the return line, making ideal gas assumptions, the pressure is given by the well known equation:

$$p_{rl} = \frac{nRT_{rl}}{V_{rl}} \tag{4}$$

where the subscript $rl$ referes to the return line.

The supply line pressure is given by:

$$p_{tl} = \int\left(q_{tl} - q_f\right)\frac{B}{V_{tl}} + p_{hyd} \tag{5}$$

where the subscript $tl$ refers to the transmission line, $B$ is the bulk modulus of the helium fluid. $p_{hyd}$ represents hydrostatic pressure, which is given by:

$$p_{hyd} = h\rho g \tag{6}$$

where $h$ the height of the fluid in the heat exchanger capillaries and phase separator and $g$ is acceleration due to gravity.

It should be noted that, whilst liquid helium is more compressible than many other fluids (with commonly assumed bulk modulus of 268 Bar[11], although this varies with temperature), a high pressure change (in the context of this system) is required to significantly reduce its volume. Under normal conditions, helium is typically compressed no more than 1%. Hemce it can be assumed that the pressure in the transmission line is equivalent to the pressure inside the heat exchanger plus the hydrostatic pressure component.

In Fig. 3, two resistive blocks are set either side of the storage component of the heat exchanger. These blocks are treated the same as the valve blocks described in (2) and (3), with $Y$ set to unity and a conductance representative of the constriction between the transmission line and heat exchanger.

### B. Heat Exchanger

Finally, the heat exchanger block is considered. This block requires a more in depth analysis for its behaviour to be described in sufficient detail. There are several different ways to model this section of the system; the technique used here is to take a mass and energy balance across this section's boundaries.

Fig. 4 is an illustration of the capillaries and phase separator as modelled. The top section of the diagram represents the phase separator, with the horizontal return manifold attached, the bottom section represents the capillaries. The distribution of the cryogenic fluid between the capillaries is not relevant and so the capillaries are treated as a lumped element. $Q$, $q_f$ and $q_s$ are the main inputs and outputs to the system. They are, respectively, heat load, fluid flow rate into the capillaries and vapour flow rate. The fluid level in the phase separator is labelled $l$ and the system pressure and temperature are represented by $T$ and $P$.

The following assumptions have been made: The capillaries (risers) are treated as a lumped construction, this entire

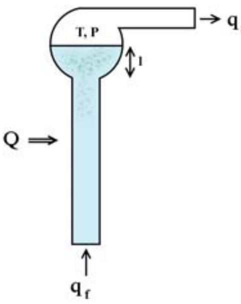


Fig. 4. Phase separator and capillary layout

section is in thermal equilibrium, and the phase separator contains a saturated vapour-fluid mixture. These assumptions are supported by historical process data and represent that real working of the system well. In the coming analysis, the terms $Q$, $q$, $h$, $V$ and $\rho$ represent, heat, flow rate, specific enthalpy and volume respectively. The subscripts '$s$', '$f$', '$w$' and '$t$' refer to vapour, feed, fluid and total respectively. SI units are used unless otherwise stated. This portion of the model is derived as follows.

*1) Global Mass and Energy Balance:* To begin with, the global mass balance of the system is described by:

$$\frac{d}{dt}\left[\rho_s V_{st} + \rho_w V_{wt}\right] = q_f - q_s \tag{7}$$

And the energy balance of the system is:

$$\frac{d}{dt}\left[\rho_s V_{st} u_s + \rho_w V_{wt} u_w + m_t C_p t_m\right] = Q + q_f h_f - q_s h_s \tag{8}$$

Now, the well known relationship between specific enthalpy and specific energy

$$u = h - \frac{p}{\rho} = h - pV \tag{9}$$

is substituted into equation 8 to give a global energy balance:

$$\begin{aligned}&\frac{d}{dt}\left[\rho_s V_{st} h_s + \rho_w V_{wt} h_w + m_t C_p t_m - pV_t\right]\\&= Q + q_f h_f - q_s h_s\end{aligned} \tag{10}$$

*2) Local Mass and Energy Balances:* To describe the distribution of vapour of fluid in the phase separator we start with the relationship between their volumes:

$$V_t = V_{st} + V_{wt} \tag{11}$$

And condensation enthalpy is represented by $h_c$.

With this property defined, the next step is to examine the mass and energy balance in the capillaries. In the analysis of other two-phase boiler systems (for example, power plant boilers), often the term "steam quality" is used to describe the proportion of vapour to fluid at a given location.

The mass fraction of vapour in the capillaries (steam quality) here is described by:

$$\alpha_m = \frac{QA}{qh_c V} z \tag{12}$$

If $\xi$ is the normalised length coordinate along the riser ($0 \leq \xi \leq 1$), then:

$$\alpha_m = \alpha_r \xi \tag{13}$$

Steam slip is a measure of the relative average velocities of liquid and gas phases in two phase flow. In this model it is assumed that steam slip is negligible, as its inclusion significantly increases the complexity of the analysis while not contributing significantly to the output of the model. For similar reasons, it is assumed the the boiling and vapour nucleation process begins at the bottom of the capillary tubes.

The average steam volume ratio in the capillaries is given by:

$$\begin{aligned}\overline{\alpha}_v = \frac{\rho_w}{\rho_w - \rho_s}\Bigg[1 - &\frac{\rho_s}{\left(\rho_w - \rho_s\right)\alpha_r}\\ &\ln\left(1 + \frac{\rho_w - \rho_s}{\rho_s}\alpha_r\right)\Bigg]\end{aligned} \tag{14}$$

Using this relation, the capillary mass balance is given by:

$$\frac{d}{dt}\left[\rho_s \overline{\alpha}_v V_r + \rho_w \left(1 - \overline{\alpha}_v\right) V_r\right] = q_f - q_r \tag{15}$$

Then from this and 9, the capillary energy balance is:

$$\begin{aligned}&\frac{d}{dt}\left[\rho_s h_s \overline{\alpha}_v V_r + \rho_w h_w \left(1 - \overline{\alpha}_v\right) V_r - pV_r + m_r C_p t_s\right]\\&= Q + q_f h_f - \left(\alpha_r h_c + h_w\right) q_r\end{aligned} \tag{16}$$

With the mass and energy balance in the capillaries accounted for, the dynamics of the phase separator and the distribution of vapour within it remain to be analysed.

The volume of vapour in the phase separator under the fluid level is given by:

$$\frac{d}{dt}\left[\rho_s V_{sd}\right] = \alpha_r q_r - q_{sd} - q_{cd} \tag{17}$$

where the condensation flow rate in the phase separator is given by:

$$\begin{aligned}q_{cd} = \frac{1}{h_c}\Bigg[&\rho_s V_{sd}\frac{dh_s}{dt} + \rho_w V_{wd}\frac{dh_w}{dt}\\&- \left(V_{sd} + Vwd\right)\frac{dp}{dt} + m_d C_p \frac{dt_s}{dt}\Bigg]\end{aligned} \tag{18}$$

The flow rate of vapour through the surface is calculated from the velocity and volume of vapour bubbles leaving the risers:

$$q_{sd} = \frac{\rho_s V_{sd}\left[1.53\frac{\sigma g(\rho_w - \rho_s)}{\rho_w^2}\right]^{\frac{1}{4}}}{l} \tag{19}$$

The volume of fluid inside the phase separator is obtained by subtracting the amount of fluid in the capillaries from the total amount of fluid in the system:

$$V_{wd} = V_{wt} - (1 - \overline{\alpha}_v)\, V_r \tag{20}$$

This allows the fluid level in the phase separator to be calculated:

$$\text{level}(\%) = 50 + \left[\arcsin\left(2\frac{V_{wd} + V_{sd}}{V_t}\right) - 1\right]\frac{50}{90} \tag{21}$$

where the trigonometric function is in degrees and the fill level of the phase separator is between 0 and 100 %.

*3) Dynamics of the Capillaries and Phase Separator:* In order to describe the mass flow rates of vapour and fluid through the capillaries and phase separator, first the capillary mass balance described in 15 is multiplied by $-(h_w + \alpha_r h_c)$:

$$\begin{aligned}&\frac{d}{dt}\left[\rho_s\overline{\alpha}_v V_r + \rho_w (1 - \overline{\alpha}_v)\, V_r\right]\left[-(h_w + \alpha_r h_c)\right] \\ &\quad = \left[q_f - q_r\right]\left[-(h_w + \alpha_r h_c)\right]\end{aligned} \tag{22}$$

Then this is added to the capillary energy balance, given in 16:

$$\begin{aligned}&\frac{d}{dt}\left(\rho_s\overline{\alpha}_v h_s V_r\right) - (h_w + \alpha_r h_c)\frac{d}{dt}\left(\rho_s\overline{\alpha}_v V_r\right) \\ &+ \frac{d}{dt}\left(\rho_w h_w (1 - \overline{\alpha}_v)\, V_r\right) \\ &- (h_w + \alpha_r h_c)\frac{d}{dt}\left(\rho_w (1 - \overline{\alpha}_v)\, V_r\right) - V_r\frac{dp}{dt} + m_r C_p\frac{dt_s}{dt} \\ &= Q + q_f h_f - (\alpha_r h_c + h_w)\, q_r - (q_f - q_r)(h_w + \alpha_r h_c) \\ &= Q + q_f h_f - q_f h_w - q_f\alpha_r h_c \\ &= Q + q_f (h_f - h_w - \alpha_r h_c)\end{aligned} \tag{23}$$

This simplifies to:

$$\begin{aligned}&h_c (1 - \alpha_r)\frac{d}{dt}\left(\rho_s\overline{\alpha}_v V_r\right) + \rho_w (1 - \overline{\alpha}_v)\, V_r\frac{dh_w}{dt} \\ &- \alpha_r h_c\frac{d}{dt}\left(\rho_w (1 - \overline{\alpha}_v)\, V_r\right) + \rho_s\overline{\alpha}_v V_r\frac{dh_s}{dt} \\ &- V_r\frac{dp}{dt} + m_r C_p\frac{dt_s}{dt} = Q + q_f (h_f - h_w - \alpha_r h_c)\end{aligned} \tag{24}$$

To derive an equation for the capillary flow rate, $q_r$, 15 is rearranged into terms of $p$ and $\alpha_r$:

$$\begin{aligned}q_r &= q_f - \frac{d}{dt}\left[\rho_s\overline{\alpha}_v V_r + \rho_w (1 - \overline{\alpha}_v)\, V_r\right] \\ &= q_f - V_r\frac{d}{dt}\left((1 - \overline{\alpha}_v)\,\rho_w + \overline{\alpha}_v\rho_s\right) \\ &= q_f - V_r\frac{d}{dp}\left[(1 - \overline{\alpha}_v)\,\rho_w + \overline{\alpha}_v\rho_s\right]\frac{dp}{dt} \\ &\quad + V_r (\rho_s - \rho_w)\frac{d\overline{\alpha}_v}{d\alpha_r}\frac{d\alpha_r}{dt}\end{aligned} \tag{25}$$

The final step is to derive an expression for the dynamics of the vapour in the phase separator. Substituting the mass equations for $q_{cd}$, $q_{sd}$ and $q_r$ (equations 18, 19 and 25) into the vapour balance equation (17) gives:

$$\begin{aligned}&\rho_s\frac{dV_{sd}}{dt} + V_{sd}\frac{d\rho_s}{dt} \\ &= \alpha_r\Big(q_f - V_r\frac{d}{dp}\left[(1 - \overline{\alpha}_v)\,\rho_w + \overline{\alpha}_v\rho_s\right]\frac{dp}{dt} \\ &+ V_r (\rho_w - \rho_s)\frac{d\overline{\alpha}_v}{d\alpha_r}\frac{d\alpha_r}{dt}\Big) \\ &- \frac{\rho_s V_{sd}\left[1.53\frac{\sigma g(\rho_w - \rho_s)}{\rho_w^2}\right]^{\frac{1}{4}}}{l} \\ &- \frac{1}{h_c}\Big(\rho_s V_{sd}\frac{dh_s}{dt} + \rho_w V_{wd}\frac{dh_w}{dt} \\ &- \left[V_{sd} + V_{wd}\right]\frac{dp}{dt} + m_d C_p\frac{dt_s}{dt}\Big)\end{aligned} \tag{26}$$

Which can be rearranged to:

$$\begin{aligned}&\rho_s\frac{dV_{sd}}{dt} + V_{sd}\frac{d\rho_s}{dt} \\ &- \alpha_r\Big(-V_r\frac{d}{dp}\left[(1 - \overline{\alpha}_v)\,\rho_w + \overline{\alpha}_v\rho_s\right]\frac{dp}{dt} \\ &+ V_r (\rho_w - \rho_s)\frac{d\overline{\alpha}_v}{d\alpha_r}\frac{d\alpha_r}{dt}\Big) \\ &+ \frac{1}{h_c}\Big(\rho_s V_{sd}\frac{dh_s}{dt} + \rho_w V_{wd}\frac{dh_w}{dt} \\ &- \left[V_{sd} + V_{wd}\right]\frac{dp}{dt} + m_d C_p\frac{dt_s}{dt}\Big) \\ &= \alpha_r q_f + \left(\frac{\rho_s V_{sd}\left[1.53\frac{\sigma g(\rho_w - \rho_s)}{\rho_w^2}\right]^{\frac{1}{4}}}{l}\right)\end{aligned} \tag{27}$$

*4) Non-linear State Variables:* Using four state variables, a full set of set of state equations that describe this section of the model can be developed. The four state variables are pressure, steam quality at the capillary-phase separator junction, total volume of fluid and volume of vapour under the fluid level ($p$, $\alpha_r$, $V_{wt}$ and $V_{sd}$). The state equations describing the heat exchanger take the following form, where the $e$ terms are found by collecting terms in the equations highlighted below. Saturated steam tables are used to evaluate the thermodynamic terms.

From 7, 10 and 11:

$$e_{11}\frac{dV_{wt}}{dt} + e_{12}\frac{dp}{dt} = q_f - q_s \tag{28}$$

$$e_{21}\frac{dV_{wt}}{dt} + e_{12}\frac{dp}{dt} = Q + q_f h_f - q_s h_s \tag{29}$$

and 24 and 27:

$$e_{32}\frac{dp}{dt} + e_{33}\frac{d\alpha_r}{dt} = Q + q_f (h_f - h_w - \alpha_r h_c) \tag{30}$$

$$e_{42}\frac{dp}{dt}+e_{43}\frac{d\alpha_r}{dt}+e_{44}\frac{dV_{sd}}{dt}=\alpha_r q_f + \left(\frac{\rho_s V_{sd}\left[1.53\frac{\sigma g(\rho_w-\rho_s)}{\rho_w^2}\right]^{\frac{1}{4}}}{l}\right) \tag{31}$$

## IV. Validation

This model is implimented in Matlab/Simulink using a combination of standard blocks and s-functions, and the output of the simulation was compared to data recorded at JET on the NIB4 cryopumping system. The two simulation results presented below are compared to data from the morning of February 21, 2012. The inputs to the simulation were historically measured supply valve position and a heat load corresponding to the experiment shots run that morning. The physical parameters (sizes, masses, conductances and boundary pressures) of the model were set according to design data provided by CCFE.

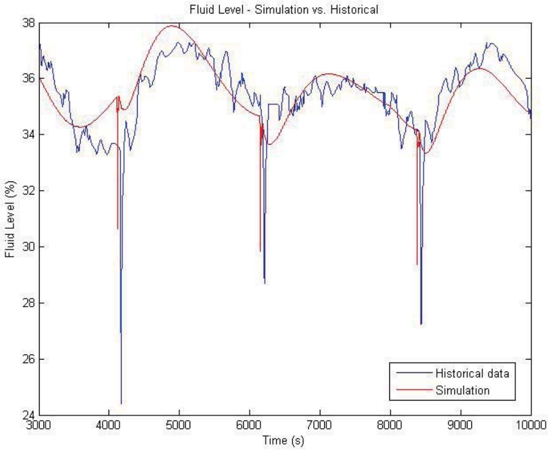


Fig. 5. Phase separator fluid level

Fig. 5 and 6 show the most interesting features of the simulation. The simulation tracks the historical process data to within 1% of the total variable range (1 to 100%) in both cases for most of the time shown. This suggests that the model describes the system well enough to simulate and predict the important system behaviors.

Each of the three salient points corresponds to the occurrence of a shot. A typical shot results in an additional heat load of around 20W on the helium panels for one or two seconds, followed by a lower heat load of around 3W for a further thirty seconds. Both the fluid level and return valve position appear to be sensitive to small heat loads on the cryopumping system, compared to the base load heat from the panel supports and thermal radiation of nearby components of close to 104W. As such, these are potentially useful variables to consider when designing a scheme to detect faults resulting in excessive heat load, such as loss of vacuum events.

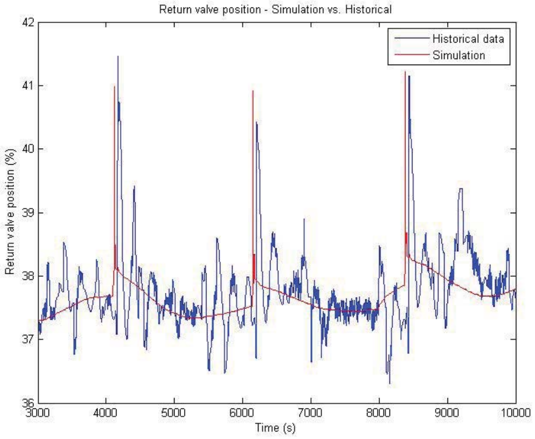


Fig. 6. Return valve position

## V. Conclusion

A novel analytical model of a large two phase cryopumping system has been derived from first principles and validated using historical data. Dividing the cryopump into resistive and storage components is a suitable technique for modelling such systems. From a Matlab/Simulink simulation it was found that phase separator fluid level and return valve position are highly sensitive to extra heat load on the cryopumping system. This suggests that the model could be used in the design of a condition monitoring scheme to detect loss of vacuum events.

Future work will see this model used to design a condition monitoring scheme, making use of the parity equations approach. This model could also be used as a basis for developing simulations of faults and could also be used as part of a model based control design process.

## Acknowledgment

This research is being conducted on behalf of CCFE and is funded by EPSRC, as a CASE studentship.

## References

[1] T. Philip, *Simulation of industrial processes for control engineers*. Butterworth-Heinemann, 1999.

[2] K. Aström and R. Bell, "Drum-boiler dynamics," *Automatica*, vol. 36, no. 3, pp. 363–378, Mar. 2000.

[3] G. Duesing, "The vacuum systems of the nuclear fusion facility JET," *Vacuum*, vol. 37, no. 3-4, pp. 309–315, 1987.

[4] D. Ciric, "Neutral Beam Heating Overview," Culham Centre for Fusion Energy, Culham, Tech. Rep., 2008.

[5] D. Ciric *et al.*, "Overview of the JET neutral beam enhancement project," *Fusion Engineering and Design*, vol. 82, no. 5-14, pp. 610–618, Oct. 2007.

[6] R. F. Barron, *Cryogenic heat transfer*. Taylor and Francis, 1999.

[7] W. Rohsenow, J. Hartnett, and E. Ganic, *Handbook of heat transfer fundamentals*, second edi ed. McGraw-Hill, 1985.

[8] D. Hucknall and A. Morris, *Vacuum technology: calculations in chemistry*. Royal Society of Chemistry, 2003.

[9] R. A. Haefer, *Cryopumping: theory and practice*. Clarendon Press, 1989.

[10] H. D. Baumann, *Control Valve Primer: A User's Guide*. ISA, 2009.

[11] E. R. Grilly, "Pressure-volume-temperature relations in liquid and solid4He," *Journal of Low Temperature Physics*, vol. 11, no. 1-2, pp. 33–52, Apr. 1973.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Thermal Modelling of an Alternator for Use in a Prediction System

James Graham*, Roger Dixon*, Keith Gregory*, John Pearson†

*School of Electronic, Electrical and Systems Engineering, Loughborough University,
Loughborough, Leicestershire, LE11 3TU
†BAE Systems SEIC, Holywell Park, Loughborough University,
Loughborough, Leicestershire, LE11 3TU
Contact: J.H.Graham@lboro.ac.uk

***Abstract*—On future UAVs it is envisaged that the power requirements of all on-board electrical systems will increase. Whilst, in most flight (mission) situations the installed generation capacity will have adequate capacity to operate the systems, it is possible that during certain abnormal situations the generators on-board may be forced to operate under very high load conditions. The main failure mechanism for a generator is overheating and subsequent disintegration of windings, hence the research problem being addressed here is that of modelling the thermal dynamics of a generator in such a way that the model can be used to predict future temperatures given knowledge of the future mission requirements. The temperature predictions will be used to allow prioritising of the mission actions in order to get the most out of a generator without overheating it.**

**The research presented here shows details the modelling of the generator, and presents some initial validation results showing good correlation between data taken from the rig and the simulation output of the model.**

## I. Introduction

The current trend of increased use of electrical power to run systems on-board an aircraft [1] [2] has lead to much research being undertaken to improve system performance, reliability and maintainability. Flight critical systems previously operated by other means, such as hydraulics or pneumatics, now need to be guaranteed a constant supply of electricity to be considered safe. This aim of this study is to look at power supply within aircraft, focusing particularly on power generation, aiming to develop methods of ensuring more reliable supply of power.

The ability of a generator to supply power is based upon its thermal state, assuming there are no other faults on the system. Under normal circumstances a well designed generator will easily operate within its thermal limits, but certain abnormal situations may force very high loading of the generator which could push it near or above its thermal limits.

The main failure mechanism for a generator is overheating and subsequent disintegration of windings, hence the research problem being addressed here is that of modelling the thermal dynamics of a generator in such a way that the model can be used to predict future temperatures given knowledge of the future mission requirements. The temperature predictions will be used to allow prioratising of the mission actions in order to get the most out of a generator without overheating it.

When looking at the literature available on thermal modelling of electrical machines it became apparent that it was almost exclusively focused on their design, and there was no work using this type of model as part of a health management system. The types and use of thermal models is summarised by Boglietti et.al. [3] in their survey paper.

When choosing the modelling method, the biggest factor in the decision was the execution speed. For the model to be useful in a prediction system it would have to be able to make predictions in real-time, this eliminated computational fluid dynamics (CFD), and finite element analysis (FEA) as possible methods due to their computational intensity. Instead a lumped parameter, thermal network type model was chosen.

Literature is available detailing this type of model. Example include the work of Perrez and Kassakian [4], and the work of Mellor et. al. [5]. This works on the principle that at its most simple an electrical machine is a lump of iron/steel, with copper windings. By modelling the iron parts of the generator as cylinders, parameters can be designed to calculate the average temperature of the various components to a good degree of accuracy, but using a fraction of the computing power of other methods.

The rest of this paper details the design of the thermal model for a generator, specifically one that is part of an experimental rig available to help validate the model. Section II. talks about the experimental rig. Section III. details the model design and choices, as well as the method of simulation. Section IV. details the current results, and section V. concludes the paper.

## II. Experimental Setup

The experimental rig setup is shown in Figure 1.

The rig itself consists of a motor driving the shaft of a 3-phase, 415V, 50Hz generator with a rated power of 5kVA, allowing the shaft speed to be controlled as necessary for any experiments. Table I lists the sensor measurements currently available, although more sensors are to be added in the near future to give a larger range temperature measurements from the generator. These include the temperature of the air flow at the input, output, and in some of the air spaces where possible. Sensors which enable a mass flow reading for the airflow to be determined are also to be installed.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

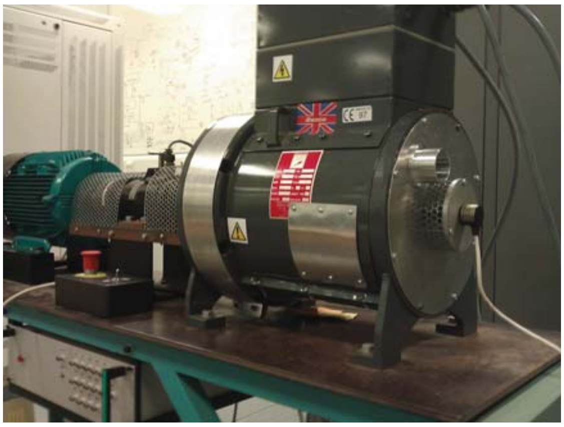


Fig. 1. Test Rig.

TABLE I
SENSOR ON THE RIG

| Variable | Signal Label | Transducer Range |
|---|---|---|
| Stator Phase U Current | Ua1 | ±230$A$ |
| Stator Phase U Current | Ua2 | ±15$A$ |
| Stator Phase V Current | Va1 | ±230$A$ |
| Stator Phase V Current | Va2 | ±15$A$ |
| Stator Phase W Current | Wa1 | ±230$A$ |
| Stator Phase W Current | Wa2 | ±15$A$ |
| Stator Phase U - Neutral Voltage | Uv | ±500$V$ |
| Stator Phase V - Neutral Voltage | Vv | ±500$V$ |
| Stator Phase W - Neutral Voltage | Wv | ±500$V$ |
| Exciter Winding Current | Ea | ±5$A$ |
| Exciter Winding Voltage | Ev | ±125$V$ |
| Auxiliary Winding Current | Xa | ±2.5$A$ |
| Auxiliary Winding Voltage | Xv | ±500$V$ |
| Stator Phase U Temperature | Ut | $-50$ to 300°$C$ |
| Stator Phase V Temperature | Vt | $-50$ to 300°$C$ |
| Stator Phase W Temperature | Wt | $-50$ to 300°$C$ |

## III. Model Design

The first stage in designing a lumped parameter model is to subdivide the generator into the various section, with each section then being modeled, as a subset of the full model. The subdivisions need to model the generator in sufficient detail that the temperature states of various parts of the generator can be found. This is balanced by the most important requirement, that the model is able to simulate the thermal state in real time so the number of nodes needs to be kept to the minimum possible.

Many authors have made different choices in how they subdivide their model Yangsoo et.al [6] divided the stator and rotor into multiple nodes in order to obtain the best possible accuracy, but this also lead to increased computation time. Kylander [7] showed that a thermal network model can still obtain high accuracy even when subdividing the machine in a modest number of areas. This was also applied by Okoro [8] giving steady state temperatures within ±5°$C$.

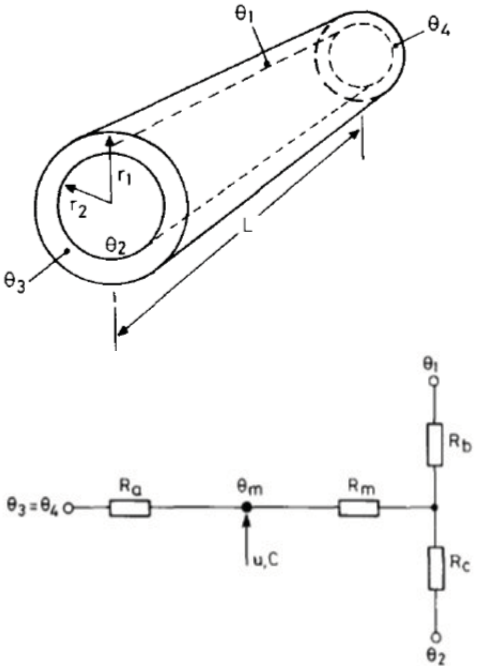


Fig. 2. Standard Cylinder and Thermal Network [5].

### A. Subdivisions within the Model

In the model the generator was broken down into 5 main areas, some of which were further subdivided. The 5 areas were the frame, stator, rotor, shaft and the air within the generator.

Unlike most current literature where the whole motor is considered in the same amount of detail, whether that be high or low, for this model the most important heat paths were identified and modeled in the most detail. These were: the stator, due to the fact that the stator winding will be the hottest part of the motor; and the air in the generator, the high airflow through the machine means that the forced convection created by it needs to be modeled carefully to get the best accuracy.

Two types of heat transfer considered in the model: conduction; and convection. The conductive heat transfer was modeled using a node at the centre representing the average temperature of the section, where any heat input, and the heat storage of the section are introduced, thermal resistance then describe the heat conduction over the component. The general model used to represent a cylinder is shown in Figure 2.

Due to the fact that the stator in the machine is only 54mm long, it was assumed that the axial temperature in the components of the model is constant, this is shown in figure 2 as a single thermal resistance $R_a$ representing the flow of heat from the mean temperature to the ends of the cylinder. Unlike many authors using this approximation though the decision was taken to model the whole length of the machine, not just half. This was due to considerations with the coolant network that are described later. The equations for the network from Mellor et. al. are shown below [5].

$$R_a = \frac{L}{6\pi k_a(r_1^2 - r_2^2)} \tag{1}$$

$$R_b = \frac{1}{2\pi k_r Ls}\left(1 - \frac{2r_2^2 log(\frac{r_1}{r_2})}{r_1^2 - r_2^2}\right) \tag{2}$$

$$R_c = \frac{1}{2\pi k_r Ls}\left(\frac{2r_1^2 log(\frac{r_1}{r_2})}{r_1^2 - r_2^2}\right) \tag{3}$$

$$R_m = \frac{1}{4\pi k_r Ls(r_1^2 - r_2^2)}\left(r_1^2 + r_2^2 - \frac{4r_1^2 r_2^2 log(\frac{r_1}{r_2})}{r_1^2 - r_2^2}\right) \tag{4}$$

Where $R_{a,b,c,m}$ are the resistance shown in Fig. 2, $L$ is the cylinder length, $r_1$ and $r2$ are the outer and inner radius respectively, $k_r$ and $k_a$ are the axial and radial thermal conductances, and $s$ is the stacking factor.

*1) Stator:* As stated earlier the stator is one of the two important areas in the model, and getting the necessary detail to model the winding temperature as well as the iron temperature was necessary. The stator was split into three areas, the stator back iron, stator teeth and stator winding. The stator iron and teeth were modeled as cylinders shown above, with the addition for the stator teeth of a resistance which gave the temperature of the inside of the slot touching the winding.

The stator winding was modeled by using the network for a rod (i.e. inner radius = 0) to represent a single slot, then multiplying this by the number slots on the stator.

*2) Rotor:* The rotor was modeled as a single cylinder, this choice was made because it was less important to have the same detail in the model that the stator has, as it would be much cooler during operation and not liable to overheat. It would also be extremely difficulty to measure the temperature of the rotor during operation making it hard to validate the results accurately, and the extra computation time required to simulate the nodes is unnecessary in this application.

*3) Shaft and Frame:* Both the shaft and frame were modeled as simple two resistance networks, which are more than adequate for this application.

### *B. Coolant Network*

The generator that the model is based upon (Fig. 1) has a large fan attached to the shaft designed to provide a large amount of airflow through the generator. This means that within this model it is not appropriate to assume that most heat transfers from the stator to rotor, or vice versa.

Yangsoo and Kauth [6] researched the modelling of an induction motor with forced cooling, using a two-network solution, one for the motor, one for the cooling. In this case both parts are modeled as a single network in order to maintain greater simplicity. The major difference between radial and axial heat flow is in how the resistance for axial heat flow is defined. Figure 3 shows a diagram of the thermal network for the air gap in the machine.

The thermal resistances $R_{con1}$ and $R_{con2}$ are calculated using the equation for convective heat transfer

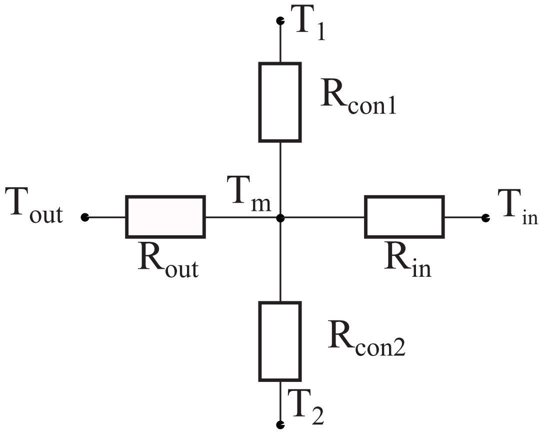


Fig. 3. Network for airflow.

$$R = \frac{1}{h_C A} \tag{5}$$

Where $h_C$ is the convective heat transfer co-efficient, and $A$ is the surface area.

The thermal resistance for $R_{in}$ and $R_{out}$ are based upon the mass flow rate through the air section

$$R = \frac{1}{\dot{m}C_p} \tag{6}$$

Where $\dot{m}$ is the mass flow rate, and $C_p$ is the thermal conductivity. As expected this leads to very low resistance values for axial heat flow through the air sections as expected with forced convection.

This thermal network is used to model the four air sections in the generator, the air gap, the two end air sections, and finally the air in the gap between the stator and the frame. The air flows into the end air at one side, splits between the air gap and top air, then re-combines on the other side. The size of the gap between stator and the frame is much larger than the stator/rotor air gap meaning the large proportion of the air flows through it cooling the stator and transferring some heat to the frame.

### *C. Solving the Thermal Network*

Figure 4 shows the complete thermal network of the generator, with Table II showing what each number in the diagram represents. From this a heat balance equation can be obtained for each node in the model, the equation for the central node of the stator iron is shown below

$$C_{si}\frac{d\theta_{sim}}{dt} = \frac{1}{R_{sia}}(\theta_{sia} - \theta_{sim}) + \frac{1}{R_{sir3}}(\theta_{sim'} - \theta_{sim}) + U_{si} \tag{7}$$

Where the subscript $si$ denotes that the value is from the stator back iron, with the letters $a$ and $r$, denoting axial and radial directions respectively, $R_{sia}$ for example is the stator iron axial resistance. $C_{si}$ is the heat storage for the stator back iron, $\theta_{sim}$ is the average temperature of the part, $U_{si}$ is the heat

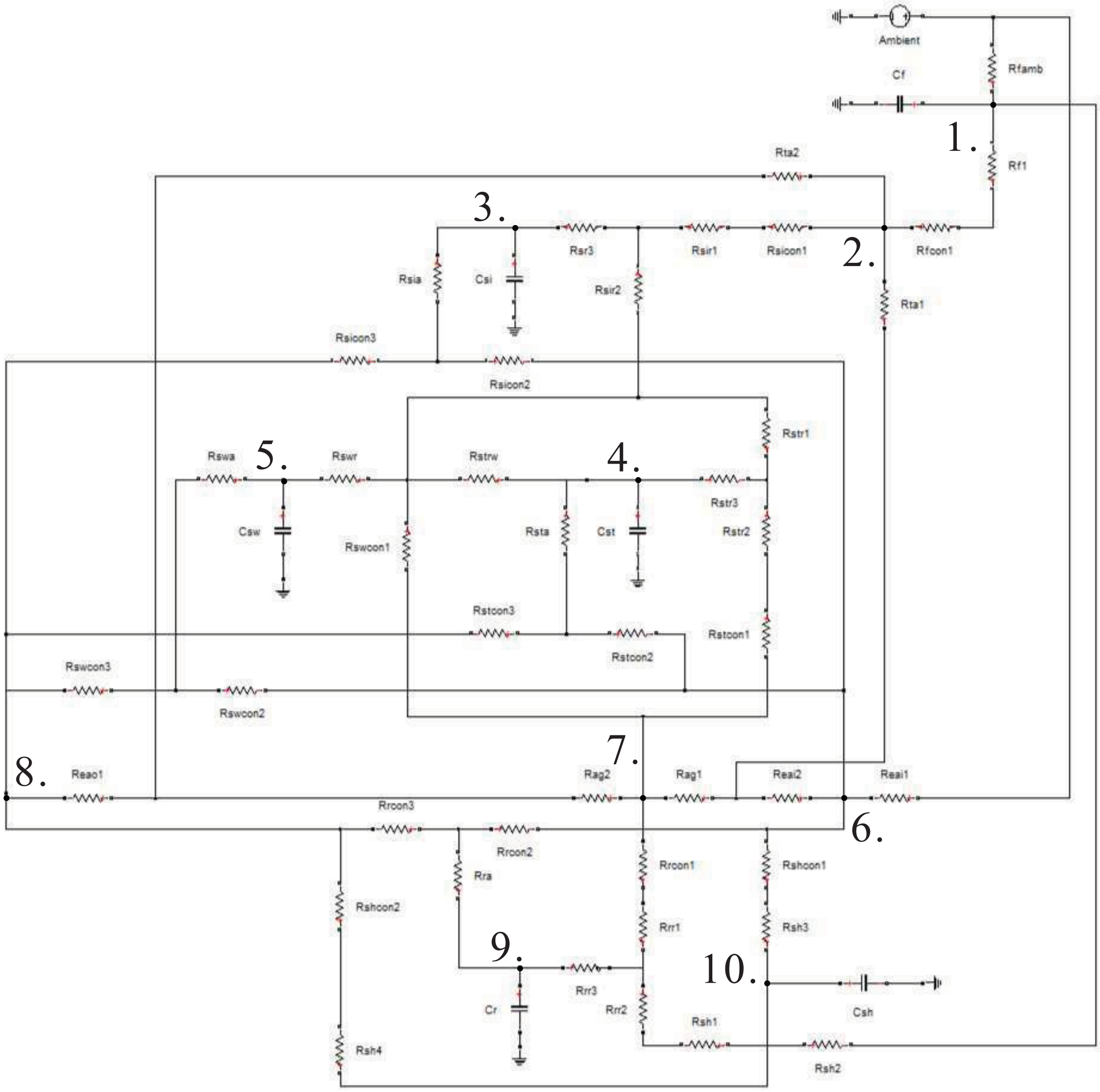


Fig. 4. Complete thermal network.

TABLE II
MODEL SUBDIVISIONS

| Fig. 4 No. | Generator Part |
|---|---|
| 1 | Frame |
| 2 | Top Air |
| 3 | Stator Back iron |
| 4 | Stator Teeth |
| 5 | Stator Winding |
| 6 | End Air In |
| 7 | Air Gap |
| 8 | End Air Out |
| 9 | Rotor |
| 10 | Shaft |

input to the stator iron. $\theta_{sia}$ and $\theta_{sim'}$ are the temperatures at the adjacent nodes, with $R_{sia}$ and $R_{sir3}$ being the thermal resistance between them.

Each node can be expressed in a similar manner to Eq. 7, these equations can then be expressed in matrix form

$$[C]\frac{d[\theta]}{dt} = [G][\theta] + [u] \tag{8}$$

where $[C]$ is a square matrix of thermal heat storage values, $[G]$ is a square matrix of thermal conductance between nodes, $[\theta]$ is the node temperature states, and $[u]$ is a column matrix of heat sources. The above differential equations allow the network to be solved for transient conditions.

TABLE III
SIMSCAPE EQUIVALENT COMPONENTS

| Electrical Component | Thermal Equivalent |
|---|---|
| Resistor | Thermal Resistance |
| Capacitor | Heat Storage |
| Current Source | Heat Input |
| Voltage Source | Constant Temp |

### D. Simulation in Simscape

When the initial work was undertaken to create this model, the first iterations were kept simple with few nodes to test the concept, complexity was then increased through iterations, with the addition of nodes. It was found that while it was easy to update the various resistances within the model using a simple Matlab script, the addition of nodes meant that often equations completely changed, as well as new ones added, and this meant that the system of equations would have to constantly be re-written, and then condensed, taking a lot of time.

In order to speed up the process, to avoid having to re-calculate the the matrices for Eq. 8 every time changes to the number of nodes are made, the Simscape package in the Matlab Simulink environment was used. Specifically the electrical component libraries. The electrical library was used instead of the thermal library due to the fact the thermal library doesn't have the flexibility required to simulate the cylindrical components used, nor can some of the interactions be modeled sufficiently between certain sub-divisions within the model.

The electrical components within the model are easily related to the equivalent thermal ones, Table III list the components and what they represent.

## IV. VALIDATION

Some initial data was collected from the alternator rig and was used to show that the model performed as expected. Currently it was only possible to extract temperature data for stator windings, as more temperature sensors are yet to be installed, but the results still give a good indication of the accuracy of the model until full validation is finished.

The temperatures were measured using PT 100 type platinum resistance thermometers (PRT) placed within each of the stator windings. The PRTs conform to the class B specification in BS 1904 : 1984 [9], the probes are 35mm long, and 4.8mm diameter, with a thin film core. The three measurements taken were then combined to give an average temperature of the stator windings as a whole.

Figure 5 shows a comparison of the measured stator winding temperature to the simulated one. The results show that the model simulates the final steady state temperature with good accuracy, but as is seen in work by others [8] the time constant of the model is faster than the actual system, while this doesn't lead to huge inaccuracy matching the time constant more accurately will provide better performance during rapid load transitions. Further data needs to be collected, giving information on other areas of the generator. Tests need to be undertaken which show the model simulating multiple steps for increasing power output, and finally tests which show the model performance during a cooling cycle.

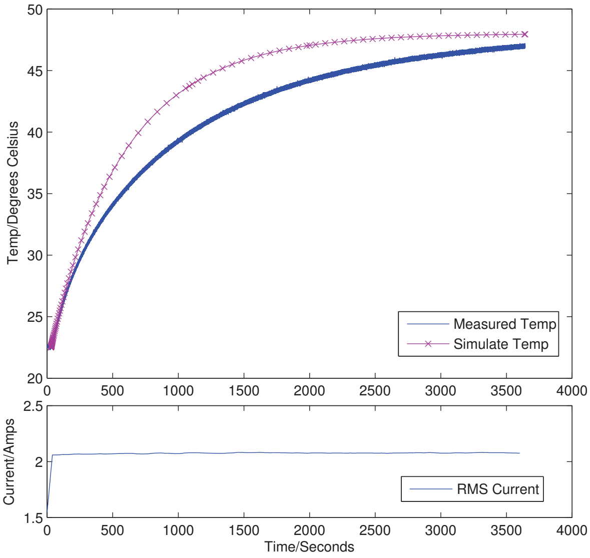


Fig. 5. Initial comparison of model to stator winding temperature.

Also included are graphs showing the simulated temperatures for the rest of the generator (Figures 6 and 7). Looking at the graph it can be seen that the simulation results meet some basic expectation. The stator winding and teeth are the hottest parts of the generator as expected, followed by the stator iron. It also shows that the heat flow in the air is modeled correctly, with the input air temperature much lower than the output air temperature.

## V. FURTHER WORK

The work described here covers just the initial modelling of the generator, with some initial validation work completed. The next stage is to complete the validation process, with the expectation of increasing the accuracy of the model further, especially refining the most important heat paths in the model around the stator and its winding, as well as the airflow through the generator.

There is also the second stage of the study to complete using the model as a prognostic tool to predict the generator's thermal state over time. This will include designing a system to set the initial model states as accurately as possible before prediction, a method of creating estimated load profiles, and the designing of scenarios to test the system at its limits.

## VI. CONCLUSION

This paper has presented the thermal model for a 3-phase generator. The execution is very fast, simulating 3614 seconds of operation in 2.59 seconds, therefore easily fulfilling the

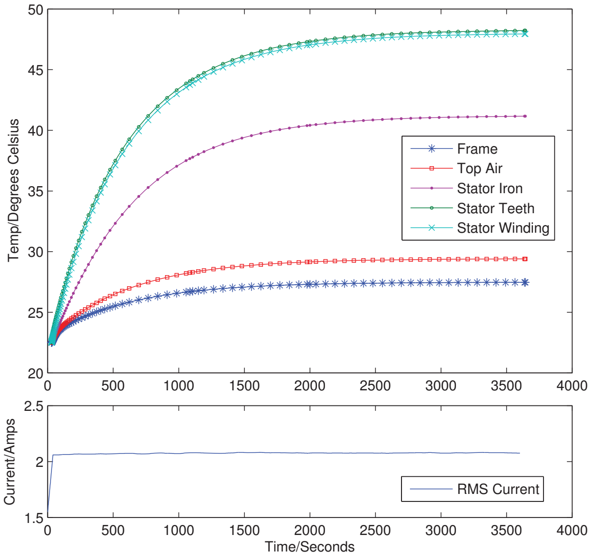


Fig. 6. Stator, Frame and Top Air temperatures with current over time.

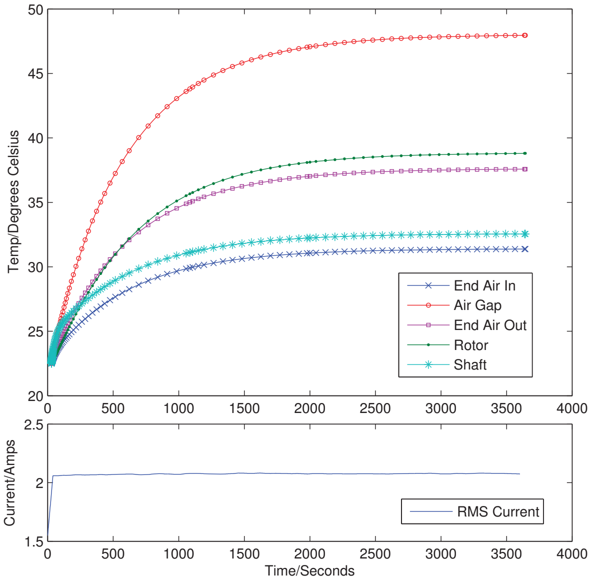


Fig. 7. Air and Rotor temperatures with current over time.

most important requirement of the model, to operate in real-time or faster.

Initial validation results show that the model is able to achieve a good degree of accuracy as well, simulating transient operation. Although there is still more testing to be undertaken with extra sensors added to test if this accuracy is maintained across all temperature states of the model.

## VII. Acknowledgments

The authors work like to thank the Engineering and Physical Sciences Research Council (EPSRC) for the case award for this work, as well as BAE Systems for the additional support provided.

## References

[1] I. Moir, “More-electric aircraft-system considerations,” *IEE Colloquium on Electrical Machines and Systems for the More Electric Aircraft*, 1999.

[2] J. Rosero, J. Ortega, E. Aldabas, and L. Romeral, “Moving towards a more electric aircraft,” *IEEE Aerospace and Electronic Systems Magazine*, vol. 22, 2007.

[3] A. Boglietti, A. Cavagnino, D. Staton, M. Shanel, M. Mueller, and C. Mejuto, “Evolution and modern approaches for thermal analysis of electrical machines,” *IEEE Transactions on Industrial Electronics*, vol. 56, no. 3, March 2009.

[4] I. Perez and J. Kassakian, “Stationary thermal-model for smooth air-gap rotating electric machines,” *Electrical Machines and Electromechanics*, vol. 3, no. 3 - 4, 1979.

[5] P. Mellor, D. Roberts, and D. Turner, “Lumped parameter thermal model for electrical machines tefc design,” in *IEE Proceedings-B Electric Power Applications*, vol. 138, 1991.

[6] Y. Lee, S. yop Hahn, and S. K. Kauh, “Thermal analysis of induction motor with forced cooling channels,” *IEEE Transactions on Magnetics*, vol. 36, no. 4, July 2000.

[7] G. Kylander, “Thermal modelling of small cage induction motors,” Ph.D. dissertation, School of Electrical and Computer Engineering Chalmers University of Technology, Gteborg, Sweden, 1995.

[8] O. I. Okoro, “Steady and transient states thermal analysis of a 7.5-kw squirrel-cage induction machine at rated-load operation,” *IEEE Transactions on Energy Conversion*, vol. 20, no. 4, December 2005.

[9] BSI, “British standard specification for industrial platinum resistance thermometer sensors,” British Standards Institution, BS 1904, 1984.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# *Fault Detection and Diagnosis using Principal Component Analysis of Vibration Data from a Reciprocating Compressor*

M Ahmed, M Baqqar, F Gu and A D Ball

Center for Diagnostic Engineering

University of Huddersfield,

Queensgate, Huddersfield HD1 3DH, UK

E-mail:M.Ahmed@hud.ac.uk

*Abstract—***This paper investigates the use of time domain vibration features for detection and diagnosis of different faults from a multi stage reciprocating compressor. Principal Component Analysis (PCA) is used to develop a detection and diagnosis framework in that the effective diagnostic features are selected from PCA of 14 potential features and a PCA model based detection method using Hotelling's $T^2$ and Q statistics is subsequently developed to detect various faults including suction valve leakage, inter-cooler leakage, loose drive belt, and combinations of discharge valve leakage with suction valve leakage, suction valve leakage with intercooler leakage and discharge valve leakage with intercooler leakage. A study of Q -contributions has found two original features: Histogram Lower Bound and Normal Negative log-likelihood which allow full classification of different simulated faults.**

***Keywords; Fault detection, Vibration, Reciprocating compressor, Principles component analysis, contribution plots.***

## I. INTRODUCTION

Principal component analysis (PCA) has been applied successfully in condition monitoring systems[1]. Statistical techniques for extracting process information from massive data sets and interpreting this information have been developed in various fields [2, 3] and PCA has been widely used to reduce the dimensionality of the original dataset by projecting it onto a lower dimensional space. Such a procedure was first proposed in 1933 by Hotelling [4] to solve the problem of de-correlating the statistical dependency between variables in multivariate statistical data derived from exam scores.

In the PCA approach, the first principal component corresponds to the direction in which the projected observations have the largest variance. The second component is then orthogonal to the first and again maximizes the variance of the data points projected on it. One approach that has proved particularly powerful for monitoring and diagnosis is the use of PCA in combination with $T^2$ charts, Q charts, and contribution plots [5]. Chemometric techniques for multivariate process monitoring have been described in several review papers [6]. Misra et al., applied PCA techniques to industrial data from a reactor system and compared its performance with that of a multi-scale PCA approach [7]. Some researchers have used different extensions of PCA such as nonlinear, multi-scale or exponentially weighted PCA [8]. Roskovic used PCA to analyse automatic fault detection and identification of process measurement equipment or sensors [9].

In this work, PCA is used not only as an approach for feature space dimensionality reduction but also for contribution plots.

A contribution plot shows the contribution of each process variable to the statistic calculated. A high contribution of a process variable usually indicates a problem with this specific variable. This approach has been used and works successfully in practice [10, 11] as it does not need historical information for the results. Kourti and MacGregor [12] applied contribution plots of quality and process variables to find faulty variables for a high-pressure low-density polyethylene reactor. They remarked that the contribution plots may not reveal the assignable causes of abnormal events; but, the group of variables contributed to the detected events will be identified for further investigation. Kano, et al., [13] presented the contribution of each process variable to the dissimilarity index used in DISSIM which can be used to identify the variables that contribute significantly to an out of control value of the dissimilarity index, and then the effectiveness of the contribution plot is evaluated. Qin et al., [14] decentralized a complex chemical process into several blocks; hierarchically investigating block and variable contributions to isolate faulty variables. Since the monitored variables were arranged into blocks according to the process, the fault isolation tasks were easier to perform than an investigation of all the variables. Yoon and MacGregor [15] comprehensively compared the model-based and data-driven approaches for fault detection and isolation, and concluded that the contribution plots provide for easy isolation of simple faults, but that additional information about the operating process is needed to isolate complex faults. This paper is organized as follows. Section 2 presents an overview of PCA for detection faults of the $T^2$ and Q statistics. In Section 3 the contribution plots Q statistic.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

## II. BASIC THEORY

### A. Data Modelling using PCA

A primary objective of PCA is for dimensionality reduction or data compression to achieve efficient data analysis. PCA forms a new smaller set of variables with minimal loss of information, compared with the original data. Based on this unique characteristic, PCA is used for classification of variables and hence early identification of abnormalities in the data structure, i.e. detection of faults.

The PCA creates a covariance matrix (or correlation matrix) by transforming the original correlated variables into a new set of uncorrelated variables. Let the variables describing the machine being investigated be the m–dimensional data set: $X = x1, x2, x3, \dots xm$, the PCA decomposes the observation vector, X, into a set of new directions P as [16]:

$$X = TP^T = t_1{P_1}^T + t_2{P_2}^T + \cdots + t_m{P_m}^T$$
$$= \sum_{i=1}^{m} t_i{P_i}^T \qquad (1)$$

Where $P_i$ is an eigenvector of the covariance matrix of X. P is defined as the principal component loading matrix and T is defined to be the score matrix of the principal components (PCs).

The loading matrix helps identify which of the variables contribute most to individual PCs, whilst the score provides information on sample clustering and identifies transitions between different operating conditions.

The expectation with PCA is that the original variables are sufficiently well correlated that the only a relatively small number of the new variables (PCs) account for most of the variance. In this case no essential information is lost by using only the first few PCs for further analysis and Equation (1) can be expressed as [17]:

$$X = TP^T + E = \sum_{i=1}^{k} t_i{p_i}^T + E \qquad (2)$$

Where E represents a residual error matrix. For example, if only the first three PCs represent a sufficiently large part of the total variance, E will be calculated by:

$$E = X - [t_1{p_1}^T + t_2{p_2}^T + t_3{p_3}^T] \qquad (3)$$

In certain applications such as process monitoring, when a plant malfunctions, original variables have minimal impact on the first few PCs, but dominate the higher orders. Thus in process engineering use of these higher order components may be needed to provide the necessary diagnostic information [16]. In this way $E$ can be very useful to measure these changes.

### B. PCA Model Based Detection

PCA based fault detection is usually based on two detection indices: Hotelling's $T^2$ statistic and Q statistic.

Hotelling's $T^2$statistic is a measure of the major variation of measurement variation and detects new data if the variation in the latent variables is greater than the variation explained by the model or baseline condition. For a new measurement feature vector x, the $T^2$ statistic detection can be found from:

$$T^2 = x^T P \lambda^{-1} P^T x \le T_\alpha^2 \qquad (4)$$

Where the $100(1-\alpha)\%$ control limit for $T_\alpha^2$ is calculated by means of a F-distribution as [18]:

$$T_\alpha^2 = \frac{k(m-1)}{m-k} F(k, m-1; \alpha) \qquad (5)$$

Where $F(k, m-1; \alpha)$ is an F-distribution with k and $(m-1)$ degrees of freedom, with chosen level of significance α, k is the number of PC vectors retained in the PCA model, and m is the number of samples used to develop the model. The Q statistic, also represented as SPE, is the squared prediction error. It is a measure of goodness of fit of the new sample to the model. The Q statistic based detection can be done by:

$$SPE = \|(I - PP^2)x\|^2 \le Q_\alpha \qquad (6)$$

The $100(1-\alpha)\%$ control upper limit $Q_\alpha$[12] is:

$$Q_\alpha = \theta_1 \left[\frac{h_0 c_\alpha \sqrt{2\theta_2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2}\right]^{\frac{1}{h_0}} \qquad (7)$$

Where:

$$\theta_i = \sum_{j=a+1}^{m} \lambda_j^i \qquad (8)$$

$$h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2} \qquad (9)$$

New events (faults) can be detected using the $T^2$or SPE; the Q-contribution plot represents the significance of each variable on the index as a function of the variable number for a certain sample, and can be used to diagnose the fault. When the $T^2$or SPE exceeds the threshold, the contribution of the individual variables to the $T^2$ or SPE can be identified, and the variable making a large contribution to the $T^2$ or SPE is indicated to be the potential fault source. In general, when an unusual event occurs and it produces a change in the covariance structure of the model, it will be detected by a high value of Q.

### C. Contribution Plots of Q Statistic

Once an abnormal factor has been detected, it is important to diagnosis the special event to find its cause. The contribution of the measurement variable and time periods to the deviation observed in the Q statistic can be used to help suggest an assignable cause. Using the distributions, confidence limits for the two statistics can be obtained. For the monitoring of new batches, the process data of the new batch $X_{new}(JK \times 1)$ is projected onto the model.

$$X_{new}^T = t_{new}^T P^T + e_{new}^T \qquad (10)$$

$$t_{new}^{T}P^{T} = X_{new}^{T}P(P^{T}P)^{-1}$$

$$e_{new}^{T} = X_{new}^{T} - t_{new}^{T}P^{T}$$

The $Q$-statistic for the new batch, $X_{new}$ is defined as follows:

$$Q_{new} = \sum_{ik=1}^{JK}(e_{new,jk})^2 \qquad (11)$$

### D. *Contribution of the Process Variables to the Q Statistic*

If, for a specific new batch, a disturbance was detected in the Q-chart of the residuals, then the contribution of the variables to the Q-statistic should be investigated. The contribution $c_{jk}^{Q}$ of process variable j at time k to the Q-statistic for this batch equals:

$$c_{jk}^{Q} = (e_{new,jk})^2 = (x_{new,jk} - \hat{x}_{new,jk})^2 \qquad (12)$$

Where $x_{new,jk}$ is the jkth element of $x_{new,jk}$(JK × 1), $\hat{x}_{new,jk}$ is the part of this element predicted by the model, and $e_{new,jk}$ is the residual. In order to find at disturbance occurred, all contributions $c_{jk}^{Q}$ can be plotted and examined [19].

## III. VIBRATION DATA AND FEATURE CALCULATION

### A. *Vibration Data Acquisition*

Vibration datasets were collected from a two-stage, single-acting Broom Wade TS9 reciprocating compressor, which has two cylinders, designed to deliver compressed air between 0.55MPa and 0.8MPa to a horizontal air receiver tank with a maximum working pressure of about 1.38MPa.

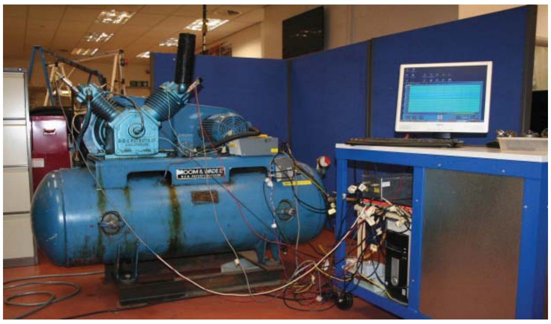


Figure 1 Reciprocating compressor test system.

As shown in Figure 1, the driving motor was a three phase, squirrel cage, air cooled, type KX-C184, 2.5kW induction motor. It was mounted on the top of the receiver and transfers its power to the compressor through a pulley belt system. The transmission ratio is 3.2, which results in a crank shaft speed of 440 rpm when the motor runs at its rated speed of 1420 rpm. The air in the first cylinder is compressed and passed to the higher pressure cylinder via an air cooled intercooler.

For characterising vibrations under different faults, four common faults were seeded into the compressor: a leaky discharge valve in the high pressure cylinder, suction valve leakage, a leaky intercooler, a loose drive belt. Different combinations of these faults were also used; discharge valve leakage combined with suction valve leakage, suction valve leakage combined with intercooler leakage and discharge valve leakage combined with intercooler leakage. In order these faults are denoted: fault 1, fault 2, fault 3, fault 4, fault 5, fault 6 and fault 7. These faults may have little effect on the pressures generated but a faulty compressor will consume more electrical energy than a healthy compressor.

Vibrations of the two-stage compressor were measured using two accelerometers mounted respectively on the low stage and high stage cylinder heads near the inlet and outlet valves. As shown in Figure 2. In addition, the pressures, temperatures and speed were also measured simultaneously for comparisons. The data segment collected is 30642 samples at different discharge pressures ranged from 0.2 to 1.2MPa in steps of 0.1MPa. As the sampling rate is 62.5 kHz, each segment of data includes more than three working cycles of the compressor, which is sufficient for obtaining stable results. In total, 8×11=88 data records were collected for the baseline and seven faults for each discharge pressure.

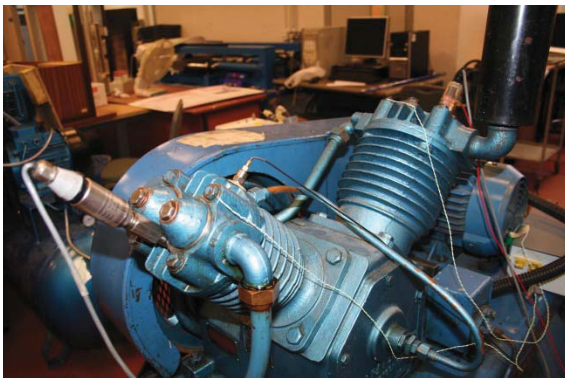


Figure 2 Vibration transducers

### B. *Time Domain Features*

Many features (statistical parameters) can be extracted from the raw vibration signals for fault detection and diagnosis. The parameters used in this study have been proved previously by many researchers as effective representation of vibration signals for CM.

The statistical parameters extracted from the raw vibration signals included, peak factor, RMS, histogram lower bound (HLB), histogram upper bound (HUB), entropy, crest factor, absolute value, shape factor, clearance factor, variance, skewness, kurtosis[20], normal negative log-likelihood value (Nnl) and Weibull negative log-likelihood value (Wnl) [21].

Weibull negative log-likelihood value and normal log-likelihood value were used recently for features extraction from vibration signals [21].

$$-LogL = -Log\prod_{i=1} f\left(a, \frac{b}{x_i}\right)$$

$$= -\sum_{i=1}^{n} log\, f(a,b\backslash x_i) \qquad (13)$$

Where $f(x_i, a, b)$ is the probabilty density function. For Weibull negative log-likelihood function and normal negative log-likelihood function, the *pdfs* are calculated as follows:

$$Weibull\, pdf\ f(x_i\backslash a,b) = \frac{b}{a}\left(\frac{x_i}{a}\right)^{b-1} exp^{-\left(\frac{x_i}{a}\right)^{b}}$$

$$Normal\, pdf f(x_i\backslash \mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} exp^{-(x_i-\mu/2\sigma^2)^2}$$

where $\mu$ and $\sigma$ denote the mean and standard deviation respectively.

## IV. DETECTION AND DIAGNOSIS RESULTS

### A. *PCA Model Development*

Figure 3 shows the relative variance of the fourteen variables selected for PCA. It also shows that seven of these account for 99% of the variance, and this means that the subspace composed of those seven PCs contains enough information on the variation of the original features for it to be sufficient to detect the faults in the reciprocating compressor.

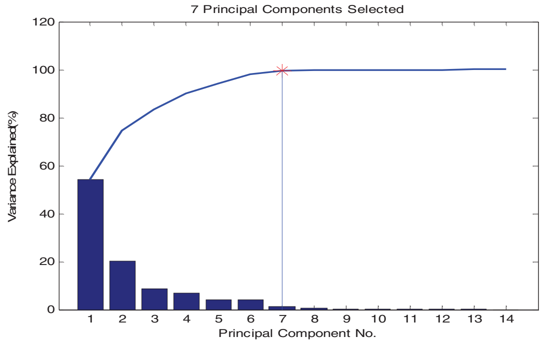


Figure 3 Principal component selection.

### B. *PCA Model Based Detection*

From Figure 4(a) and 4(b) it can be seen that most of both $T^2$ and $SPE$ are within the thresholds but there are three occasions, at samples 2, 30, 45, at which the threshold is exceeded. This can be due to the non-stationary behaviour of the vibration signal and the ability of PCA to detect the change which are acceptable from statistical analysis but means the confidence level is selected appropriately.

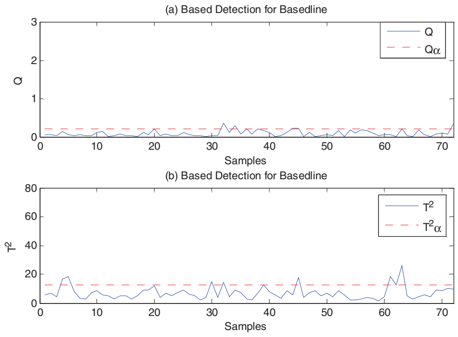


Figure 4 Model evaluation

On the Q-chart in Figure 5(a), there are some points at which the control limit is exceeded, and these indicate false alarms. However, the $T^2$ statistics detected a fault at the same points as shown in Figure 5(b), which shows too many contents reflected by the latent PCs and indicate the presence of a fault.

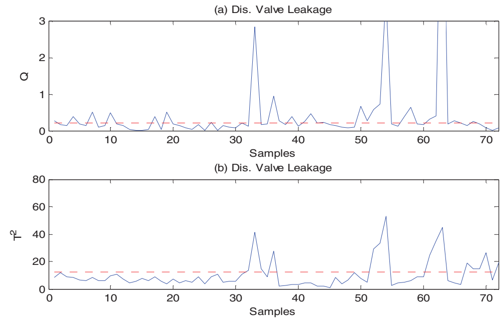


Figure 5 Discharge valve leakage detection by $T^2$ and $Q$statistics.

The performance of the $Q$ method with the leaky suction valve is shown in Figure 6(a). It can be seen that the $SPE$ value exceeds the threshold value many times which indicates the occurrence of major faults while the $T^2$ method crossed the control limits fewer times as can be seen from the Figure 6(b).

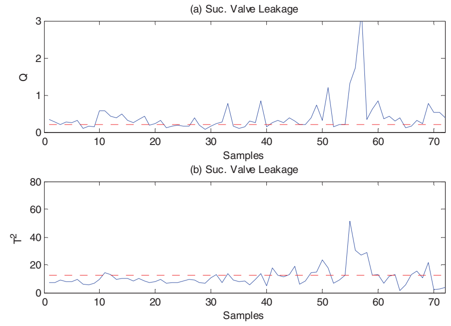


Figure 6 Suction valve leakage detection by $T^2$ and $Q$ statistics.

The result in Figure 7(a, and b) show the $Q$ and $T^2$ statistics for the intercooler leakage fault, and the values at which the threshold is crossed can be clearly seen in both plots but with larger deviation amplitude in the $T^2$ method.

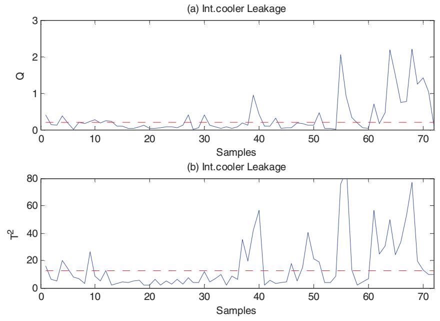


Figure 7 Intercooler detection by $T^2$ and $Q$ statistics.

Figure 8 depicts the performance $T^2$ and $Q$ methods of the loose belt fault. From the obtained result it can be seen that the $SPE$ values cross the threshold many times, which indicates the occurrence of the major faults. While the $T^2$-statistic has crossed the threshold in less points with varying amplitude.

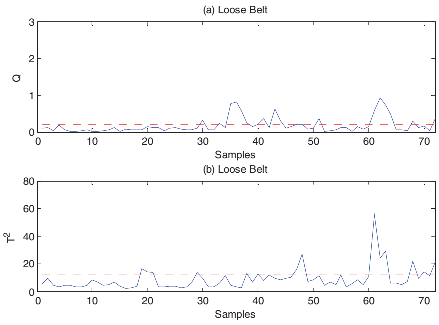


Figure 8 Loose belt detection by $T^2$ and $Q$ statistics.

The performance of $T^2$ and $Q$ statistics models with combined faults was also investigated. Figure 9 shows the results for combined discharge valve leakage and suction valve leakage. It can be seen that with the $\mathrm{T}^2$ method there are a number of occasions when the SPE value exceeds in threshold value. Similarly the $Q$ statistics clearly shown the $SPE$ plot crossed the threshold a large number of times which indicates the occurrence of major faults. This confirms the ability of the $T^2$ method to detect combined faults.

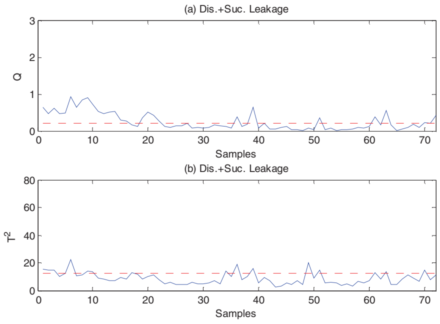


Figure 9 Combined discharge valve leakage and suction valve leakage detection by $T^2$ and $Q$ statistics.

For combined suction valve leakage and intercooler leakage, both $T^2$ and $Q$ statistics detected the faults as shown in Figure 10, where it can be clearly seen that many data points exceeds the threshold. Both models exhibited similar performance for detection of this fault with particularly high deviation amplitudes in the $Q$ statistics.

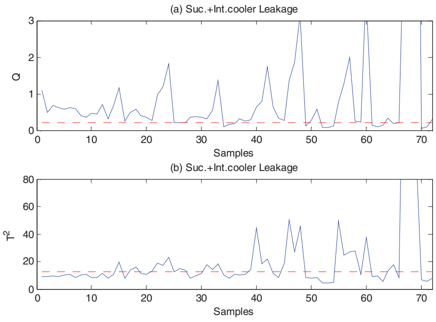


Figure 10 Combined suction valve leakage and intercooler leakage detection by $T^2$ and $Q$ statistics.

From the Figure 11, the combined discharge valve leakage and intercooler leakage provide many data points that exceed the threshold for the both $T^2$ and $Q$ statistics and hence indicate the presence of severe faults.

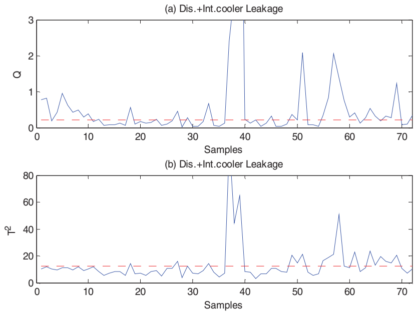


Figure11 Combined discharge valve leakage and intercooler leakage detection by $T^2$ and $Q$ statistics.

### C. PCA Model Based Diagnoses

Once a fault has been detected, it is important to identify an assignable cause. Identification of the source of the fault is facilitated by inspecting the plots showing the contributions of the various measurement variables to the deviations observed in the monitored metric. Such contribution or diagnostic charts can be immediately displayed on line by the system, as soon as the special event is detected. Although they may not provide an unequivocal diagnosis, they should at least clearly indicate the group of variables that are primarily responsible for the detected fault. The contribution plots obtained from the data in different cases as shown in Figure 12, the contribution of each variable is different. The major variables contributing in these deviations were mostly variables 10, 11 and 13 along with variables 2, 3, 4 and 14. The variables contributing most significantly to the Q-statistic are 10 and 13 because they are largest. This result implies that a fault or disturbance related to a pressure in the process occurs. On the other hand, the variables contributing significantly to the dissimilarity are 2, 3, 4, 11 and 14. These variables are slightly different from the variables contributing in the process occurs. Thus, the information obtained from the contribution plots is useful for investigating the cause of the fault.

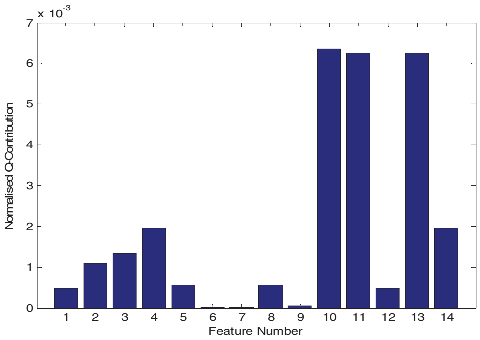


Figure 12 Overall $Q$ contribution charts for 14cases based on PCA model.

The results in Figure 13 show that variable 11 contributes most for the loose drive belt fault and combined discharge valve and suction valve leakage.

Variable 13 also recorded the highest contribution for combined discharge valve and suction valve leakage.

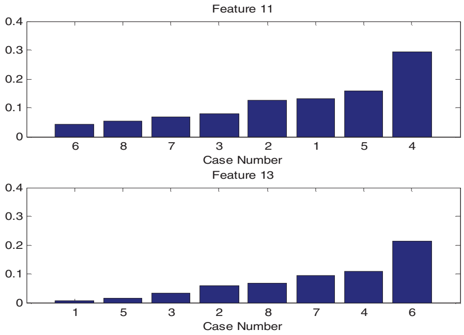


Figure 13 $Q$ contribution charts for fault classification based on feature 11 and 13.

We can therefore represent that faults as combinations of variables. Figure 14 presents a way to achieve separation between the normal operation and operation with any of the given faults. It provides the best combination of variables, with which to detect faults most effectively. It can be shown that the best combination of variables given by the $Q$-plots are variables 11 and 13. This combination gives a direction in the multivariate tool-state variable space, onto which the data can be projected, which can be used for detecting a specific class of fault. This is depicted in Figure 14. For each fault that is classified.

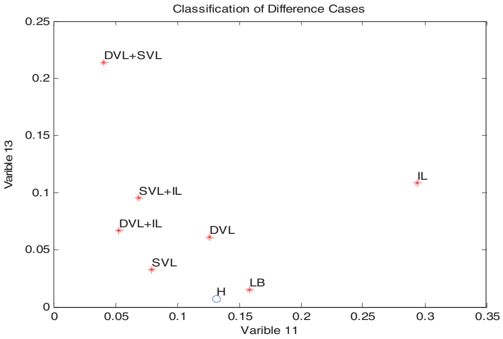


Figure 8 .Fault classifications based on feature 11 and 13 combination

## V. CONCLUSIONS

It has been demonstrated in this study that PCA based approaches allow the detection of single and multiple faults in a reciprocating compressor. The model developed from baseline consists of the seven most important PCs which explain nearly 99% of the variances from 14 original vibration features. The presence of faults can be detected by comparing the feature values from the time domain of the vibration signal with the $T^2$ and $Q$ statistics.

However the $Q$-statistic had better detection ability for all faults investigated. The contribution of the $Q$-plot, was presented in a way which allows it to be used with any latent variable component or regression model to detect a specific progress variable. The $Q$-contributions show that two particular variables, 10 and 13 gave the largest values of the minimum difference between different cases, thus these were used to detect and differentiate the given faults.

## REFERENCES

[1] L. H. Chiang and L. F. Colegrove, "Industrial implementation of on-line multivariate quality control," *Chemometrics and Intelligent Laboratory Systems,* vol. 88, pp. 143-153, 2007.

[2] M. Daszykowski, B. Walczak, and D. L. Massart, "Projection methods in chemistry," *Chemical Intelligence Laboratory Systems,* vol. 65, pp. 97– 112, 2003.

[3] R. A. Johnson, D. W. Wichern, and "Applied Multivariate Statistical Analysis," *Prentice-Hall,New Jersey,* 1992.

[4] J. E. Jackson, *A User's Guide to Principle Components*. New York: NY, 1991.

[5] J. Kresta, J. F. MacGregor, and T. E. Marlin, "Multivariate statistical monitoring of process operating performance," *Canadian Journal of Chemical Engineering,* vol. 69 pp. 35– 47, 1991.

[6] B. M. Wise and N. B. Gallagher, "The process chemometrics approach to process monitoring and fault detection," *Journal of Process Control,* vol. 6, pp. 329–348, 1996.

[7] M. Mirsa, H. H. Yoo, S. J. Qin, and C. Ling, "Multivariate Process monitoring and fault diagnosis by multi-scale PCA " *Computers and chemical Engineering,* vol. 26, pp. 1281-1293, 2002.

[8] S. Lane, E. B. Martin, A. J. Morris, and P. Gower, "Application of exponentially weighted principal component analysis for the monitoring of a polymer film manufacturing process," *Transactions of the Institute of measurement and control,* vol. 25, pp. 17-35, 2003.

[9] A. Roskovic, R. Grbic, and D. Sliskovic, "Fault tolerant system in a process measurement system based on the PCA method," in *MIPRO, 2011 Proceedings of the 34th International Convention*, 2011, pp. 1646-1651.

[10] P. Miller, R. Swanson, and C. Heckler, "Contribution plot: a missing link in multivariate quality control " *Applied Math and Computer Science,* vol. 8, pp. 775-792, 1998.

[11] P. Nomikos, "Detection and diagnosis of abnormal batch operations based on multi-way principal comonent analysis," *ISA Trans,* vol. 35, pp. 259-266, 1996.

[12] T. Kourti and J. F. MacGregor, "Multivariate SPC Methods for Process and Product Monitoring," *J.Qual. Technol,* vol. 28, p. 409, 1996.

[13] M. Kano, S. Hasebe, and I. Hashimoto, "Contribution Plots for Fault Identcation Based on the Dissimilarity of Process Data," *AIChE,* 2000.

[14] J. S. Qin, S. Valle, and M. J. Piovoso, "On Unifying Multiblock Analysis with Application to Decentralized Process Monitoring," *J. Chemom,* vol. 15, p. 715, 2001.

[15] S. Yoon and J. F. MacGregor, "Statistical and Causal Model-Based Approaches to Fault Detection and Isolation," *AIChE J,* vol. 46, p. 1813, 2000.

[16] J. F. MacGregor and T. Kourti, "Statistical Process Control of Multivariate Processes," *Control Engineering Practice,* vol. 3, pp. 403-414, 1995.

[17] B. M. Wise and N. B. Gallagher, "The Process Chemometrics Approached to process and fault detection," *J. of Process Control,* vol. 6, pp. 329-348, 1996.

[18] J. F. MacGregor and T. Kourti, "Statistical process control of multivariate processes," *Control Engineering Practice,* vol. 3, pp. 403-414, 1995.

[19] J. A. Westerhuis, S. P. Gurden, and A. K. Smilde, "Generalized contribution plots in multivariate statistical process monitoring," *Chemometrics and Intelligent Laboratory Systems,* vol. 51, pp. 95-114, 2000.

[20] M. Ahmed, F. Gu, and A. Ball, "Feature Selection and Fault Classification of Reciprocating Compressors using a Genetic Algorithm and a Probabilistic Neural Network," in *9th International Conferenceon Damage Assessment of Structures (DAMAS2011)*, Oxford, UK, 2011.

[21] M. Yadav and D. S. Wadhwani, "Vabration analysis of bearing for fault detection using time domain features and neural network," *International journal of Applied Reserch in Mechanical Engineering,* vol. 1, p. 6, 2011.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# A Robust LPV Fault Detection Approach Using Parametric Eigenstructure Assignment

Fengming Shi and Ron J. Patton
Department of Engineering
University of Hull
Hull, HU6 7RX, UK
shi_fengming@163.com and R.J.Patton@hull.ac.uk

**Abstract --In this paper, an eigenstructure assignment fault detection approach to linear time invariant (LTI) systems is extended to Linear Parameter Varying (LPV) systems. Fault detection filter design algorithms using eigenstructure assignment have been widely studied for LTI systems. However, LPV strategies are very useful for systems which have no unique equilibrium and are difficult to linearize. The parametric eigenstructure assignment approach is used to design an observer as a residual generator by viewing the varying parameters as fixed parameters in the design procedure. The residual observer feedback structure is implemented using a measured scheduling parameter An example is given of actuator fault detection of a two-link manipulator system.**

***Key words:* *Fault detection; LPV systems; Eigenstructure assignment, Fault residual generation***

## I. INTRODUCTION

Safety and reliability are very important in control systems and these demanding requirements must be ensured at a reasonable level. Fault detection (FD) methodologies and techniques are important topics in systems engineering from the viewpoint of improving plant safety and reliability. The FD literature is vast and the topics addressed are essentially related to the different design methodologies proposed to tackle the FD problem [1-4]. Model-based FD techniques are the most popular and are receiving considerable attention. The ideas are to derive a mathematical model of the plant and to compute additional artificial signals that are checked, during the on-line operations, with the corresponding measured quantities. State observers are often considered as the role of on-line residual generation because of the fast detection rate [5].

The eigenstructure assignment approach to robust FD was first demonstrated in [6]. It has been shown that a well-defined residual signal can be completely de-coupled from the disturbance by assigning a suitable eigenstructure to an observer. In this way, robust fault detection is achieved. Parametric eigenstructure assignment approaches [7-14] opened a wide field to use the design freedom of eigenvalue placement to achieve other desired performance, such as structured disturbance decouple. Some optimization approaches were also considered in the FD methods with pole placement [15-17].

Many real systems cannot be modeled by linear models, for example when no unique equilibria exist. A feasible approach to handle the nonlinearity of such systems is to use linear parameter-varying (LPV) models to approximate the dynamic nonlinearity. The LPV strategy was first introduced in [18, 19] and the big advantage of LPV modeling is that powerful linear design tools for stability and performance can be extended and applied [20, 21], LMI methods for multiple-model FD have been studied in [22, 23]. Traditional multi-model eigenstructure assignment approaches use iterative methods based on optimizing the worst case performance and the initial condition is calculated e.g. by a linear quadratic regulator (LQR) or by an H infinity method [24, 25].

In this paper, a non-iterative robust fault detection approach is presented based on a state observer structure within the LPV framework. The LPV fault detection approach is an extension of the approach in LTI case. Using parametric eigenstructure assignment, the varying parameters are viewed as fixed parameters in the design procedure and the observer law is implemented with the varying parameters measured or estimated on line.

The remainder of the paper is organized as follows: Section II recalls the parametric eigenstructure assignment approach to LPV systems and FD approach for LTI system. A design procedure for robust FD is also proposed in this Section; Section III demonstrates the usefulness of the proposed approach by a means of a two-link manipulator example. Conclusions are given in Section IV.

## II. PARAMETRIC EIGENSTRUCTURE ASSIGNMENT TO LPV SYSTEMS

Consider a stable LPV system in the following form:

$$\dot{x}(t) = A\big(\theta(t)\big)x(t) + B\big(\theta(t)\big)u(t) + B_f\big(\theta(t)\big)f(t) + Ed(t) \quad (1)$$

$$y(t) = C\big(\theta(t)\big)x(t) + D\big(\theta(t)\big)u(t), \quad (2)$$

where $x(t) \in R^n, u(t) \in R^r and\ y(t) \in R^m$ are the state vector, the input vector and measured output vector, respectively. And $f(t)$ and $d(t)$ are the fault vector and disturbance signal, respectively. $E$ is a known constant matrix and $A(.), B(.), C(.), D(.), B_f(.)$ are known

This work is sponsored by the Hull China and CSC scholarships awarded to Fengming Shi during 2010-2013

continuous functions of a time-varying parameter vector $\theta(t)$ which satisfies:

$$\theta(t) = \left[\theta_1(t), \cdots \theta_{n_\theta}(t)\right]^T \in \Theta, \forall\, t \geq 0$$

where $\Theta$ is a compact set. Hereafter, the subscript $t$ is omitted without causing confusion.

The observer dynamics used by the residual generator are described by:

$$\left.\begin{array}{l} \dot{\hat{x}} = \big(A(\theta) - K(\theta)C(\theta)\big)\hat{x} + \big(B(\theta) - K(\theta)D(\theta)\big)u + K(\theta)y \\ \hat{y} = C(\theta)\hat{x} + D(\theta)u \\ r = Q(\theta)(y - \hat{y}) \end{array}\right\} \tag{3}$$

where $r \in R^p$ is the residual vector, $\hat{x}$ and $\hat{y}$ are state and output estimation vectors. The matrix $Q(\theta) \in R^{p\times m}$ is the residual weighting factor.

### *A. LPV Parameric Eigenstrucutruer assignent*

Observer design is dual of state feedback controller design. Left Eigenstructure assignment of observer is dual of the right Eigenstructure assignment of state feedback controller. That means if the desired left eigenvector matrix of $A(\theta) - K(\theta)C(\theta)$ is $R^T(\theta)$, then the desired right Eigenvector matrix of $A^T(\theta) - C^T(\theta)K^T(\theta)$ is $R(\theta)$. That is to say the desired left Eigenvector of $A(\theta) - K(\theta)C(\theta)$can be assigned by assigning the right eigenvector of $A^T(\theta) - C^T(\theta)K^T(\theta)$. By the definition, if the right eigenvector matrix of $A^T(\theta) - C^T(\theta)K^T(\theta)$ is $R(\theta)$, the eigenvalue matrix is $F(\theta)$. It follows that:

$$A^T(\theta)R(\theta) - C^T(\theta)K^T(\theta)R(\theta) = R(\theta)F(\theta) \tag{4}$$

Let $K^T(\theta)R(\theta) = W(\theta)$, Equation (4) can be rewritten as

$$A^T(\theta)R(\theta) - C^T(\theta)W(\theta) = R(\theta)F(\theta) \tag{5}$$

Hence, the problem to assign desired closed-loop eigenstructure to a system using a residual generator is to find a solution of (4).

Now, without proof, a theorem of parametric solution of Sylvester equation is introduced, and the proof details can be found in [8, 9, 26].

**Theorem 1**

Let $[A(\theta)\ B(\theta)]$ be controllable, and the matrix $B(\theta)$ be of full-column rank. If the desired closed-loop self-conjugate eigenvalue set be described as $\Lambda = \{\lambda_i(\theta): \lambda_i(\theta) \in \mathbb{C}, i = 1,2,..\tilde{n},\ 1 \leq \tilde{n} \leq n\}$. The algebraic and geometric multiplicities of the eigenvalue $\lambda_i$are denoted by $q_i$and $r_i$, respectively and $p_{ij}$, $q_i$ *and* $r_i$satisfy the relations:

$$\textstyle\sum_{j=1}^{r_i} p_{ij} = q_i, \sum_{i=1}^{\tilde{n}} q_i = n$$

Then all the solutions of the Sylvester matrix equation [10]:

$$A(\theta)R(\theta) + B(\theta)W(\theta) = R(\theta)F(\theta)$$

are given by:

$$\begin{bmatrix} R_{ij}^k \\ w_{ij}^k \end{bmatrix} = \begin{bmatrix} N\big(\theta, \lambda_i(\theta)\big) & \cdots & \frac{1}{(k-1)!}\frac{d^{k-1}}{d\lambda^{k-1}} N\big(\theta, \lambda_i(\theta)\big) \\ M\big(\theta, \lambda_i(\theta)\big) & \cdots & \frac{1}{(k-1)!}\frac{d^{k-1}}{d\lambda^{k-1}} M\big(\theta, \lambda_i(\theta)\big) \end{bmatrix} \begin{bmatrix} f_{ij}^k(\theta) \\ \vdots \\ f_{ij}^1(\theta) \end{bmatrix} \tag{6}$$

$$\mathrm{k} = 1, 2, \ldots., \mathrm{p}_{ij}, \mathrm{j} = 1,2, \ldots.\mathrm{q}_i, \mathrm{i} = 1,2, \ldots.\mathrm{n}'$$

where the $f_{ij}^k \in C^r$ are arbitrarily chosen from parameter vectors. $N\big(\theta, \lambda(\theta)\big)$ and $M\big(\theta, \lambda(\theta)\big)$ are right co-prime matrix polynomials satisfying:

$$[\lambda(\theta)I - A(\theta)]^{-1}B(\theta) = N\big(\theta, \lambda(\theta)\big)M^{-1}\big(\theta, \lambda(\theta)\big) \tag{7}$$

Then the observer gain can be calculated by:

$$K(\theta) = R^{-1}(\theta)W(\theta)$$

From the above theorem, it can be known that the desired eigenvectors and generalized eigenvectors can be parameterized by (6). By specially choosing the free parameters given in (6), solutions with desired properties can be obtained.

### *B. LPV fault detection*

The FD design must ensure that the residuals are close to zero in the fault-free situation whilst suitably deviating from zero in the presence of faults. A necessary condition for achieving disturbance de-coupling design is [4, 14].

$$Q(\theta)C(\theta)E = H(\theta)E = 0$$

If $C(\theta)E = 0$, any residual weighting matrix can satisfy this necessary condition.

The basic principle to assign the left eigenstructure for LTI case is given in [4, 12]. The theorem is introduced here.

**Theorem 2**

The sufficient conditions for satisfying the disturbance de-coupling requirement $G_{rd} = QC(sI - A + KC)^{-1}Ed = 0$ are:

(1) $QCE = 0$

(2) All rows of the matrix $H = QC$ are left eigenvectors of $(A - KC)$ corresponding to any eigenvalues.

A similar result for the LPV case is now given.

**Theorem 3**

The sufficient conditions for satisfying the disturbance de-coupling requirements for the system.

$$G_{rd} = Q(\theta)C(\theta)(sI - A(\theta) + K(\theta)C(\theta))^{-1}Ed(s) = 0$$

are:

(1) $Q(\theta)C(\theta)E = 0$

(2) All rows of the matrix $H(\theta) = Q(\theta)C(\theta)$ are left eigenvectors of $(A(\theta) - K(\theta)C(\theta))$ corresponding to any eigenvalues.

Noting that the above result is intuitively an extension of the LTI case, the proof is omitted here.

### C. Design procedure

Following the previous arguments, a design procedure is proposed to design a robust residual generator to LPV system.

**Step 1**: Select the desired eigenvalues for the observer which can be parametric to obtain more design freedoms.

**Step 2:** Calculate the $N(\theta, \lambda(\theta))$ and $M(\theta, \lambda(\theta))$ using elementary transformation and the rational matrix factorization method.

**Step 3:** Check rank$(\mathrm{C}(\theta)\mathrm{E})$, choose a basis for $LKer(\mathrm{C}(\theta)\mathrm{E})$.

**Step 4:** Check $\mathrm{Q}(\theta)\mathrm{C}(\theta)$, set the desired *left* eigenvectors with some of the eigenvectors in a parametric form to keep the design freedom.

**Step 5:** Project the desired eigenvectors into the achievable subspace to get the achieved eigenvector matrix.

**Step 6:** Calculate the observer gain by $K(\theta) = W(\theta)R(\theta)^{-1}$. To simplify the structure, some parameters are chosen at this step.

**Step 7:** Verify the achieved eigenvalues and eigenvectors, and chose the remaining parameters based on the performance specifications.

## III. AN EXAMPLE

A two-link robotic manipulator is considered to rotate in the vertical plane, whose position can be described by a 2-vector $\varphi = (\varphi_1, \varphi_2)^T$ of joint angles, and whose actuator inputs consist of a vector $u = (u_1, u_2)^T$ of torques applied at the manipulator joints as shown in Fig. 1. Using the vectors $\dot{\varphi}$ and $\ddot{\varphi}$ to denote the joint velocities and accelerations, respectively. The dynamics of this simple manipulator can be written in the more general form [27] as:

$$\Xi(\varphi)\ddot{\varphi} + O(\varphi, \dot{\varphi})\dot{\varphi} + \mathrm{g}(\varphi) = u$$

where: $\Xi(\varphi) \in R^{2\times 2}$ is the manipulator inertia tensor matrix, $O(\varphi, \dot{\varphi})\dot{\varphi} \in R^2$ is the vector function containing the Centripetal and Carioles torques, i.e. $O(\varphi, \dot{\varphi})) \in R^{2\times 2}$ and $\mathrm{g}(\varphi) \in R^2$ are the gravitational torques. The details of equations of motion and physical parameters as outlined in Table I are described in [23].

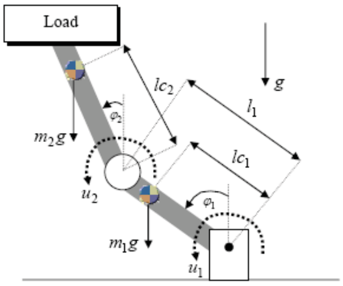


Figure 1. Two-link manipulator structure

TABLE I. PARAMETER VALUES FOR THE TWO LINK MANIPULATOR SYSTEM

| Parameters | $I_1$ | $I_2$ | $l_1$ | $lc_1$ | $lc_2$ | $m_1$ | $m_2$ | g |
|---|---|---|---|---|---|---|---|---|
| Values | 0.833 | 0.417 | 1.0 | 0.5 | 0.5 | 10.0 | 5.0 | 9.80 |
| Units | Kg*$m^2$ | Kg*$m^2$ | m | m | m | Kg | Kg | $m/s^2$ |

### A. LPV model of two link manipulator

A polytopic LPV representation of this model system is used in this example. It is important to note that in this study the quadratic terms $O(\varphi, \dot{\varphi})$ are not considered because they are not bounded. As shown later by the Simulink result, it turns out that the two-link manipulator works well, even if these bounds are not known *a priori*. Considering this limitation the system dynamics can be described as:

$$\Xi(\varphi)\ddot{\varphi} + \mathrm{g}(\varphi) = u$$

where

$$\Xi(\varphi) = \begin{bmatrix} m_{11} & m_{12}\cos(\varphi_1 - \varphi_2) \\ m_{21}\cos(\varphi_1 - \varphi_2) & m_{22} \end{bmatrix}$$

$$\mathrm{g}(\varphi) = \begin{bmatrix} k_{11}\sin(\varphi_1) \\ k_{12}\sin(\varphi_2) \end{bmatrix}$$

$$m_{11} = m_1 lc_1^2 + m_2 l_1^2 + I_1, m_{12} = m_2 l_1 lc_2,$$

$$m_{21} = m_2 l_1 lc_2, m_{22} = m_2 lc_2^2 + I_2$$

$$k_{11} = -[m_1 lc_1 + m_2 l_1]g, k_{12} = -m_2 g lc_2$$

The nonlinear term in $\Xi(\phi)$ is clearly a bounded function:

$$\rho_1 = \cos(\varphi_1 - \varphi_2), -1 \le \rho_1 \le 1$$

To facilitate a state-space formulation, the vector field $\mathrm{g}(\varphi)$ with $\varphi \in R^2$ can bearranged in the form of $G(\varphi)\varphi$ and function $\phi_2(\varphi)$ can now be defined which is bounded,

$$\sin(\varphi_1) = \left(\frac{\sin(\varphi_1)}{\varphi_1}\right)\varphi_1 = \rho_2(\varphi).\varphi_1$$

$$-0.2 \le \rho_2 \le 1.$$

Similarly,

$$\sin(\varphi_2) = \left(\frac{\sin(\varphi_2)}{\varphi_2}\right)\varphi_2 = \rho_3(\varphi)\,.\,\varphi_2$$

$$-0.2 \leq \rho_3 \leq 1$$

To define the two-link system state space representation, let:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} \varphi_1 \\ \varphi_2 \\ \dot{\varphi}_1 \\ \dot{\varphi}_2 \end{bmatrix} \text{ and } W_b = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

The LTV state space equation is as follows:

$$\dot{x} = A(\varphi)x + B(\varphi)u + Ed + R_1 f$$

where

$$A(\varphi) = \Pi^{-1}\begin{bmatrix} 0 & I \\ -\mathrm{G}(\varphi) & 0 \end{bmatrix}, B(\varphi) = \Pi^{-1}W_b,$$

$$\Pi = \begin{bmatrix} I & 0 \\ 0 & \Xi(\varphi) \end{bmatrix}, \Xi(\varphi) = \begin{bmatrix} m_{11} & m_{12}\rho_1 \\ m_{21}\rho_1 & m_{22} \end{bmatrix}$$

$$\mathrm{G}(\varphi) = \begin{bmatrix} k_{11}\rho_2 & 0 \\ 0 & k_{12}\rho_3 \end{bmatrix}$$

$$-1 \leq \rho_1 \leq 1, -0.2 \leq \rho_2 \leq 1, -0.2 \leq \rho_3 \leq 1$$

Assume that only $\varphi_1(t)$ and $\varphi_2(t)$ are measurable, so that:

$$C(\varphi) = C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

Assume further that the system is disturbed by a zero-mean Gaussian random disturbance $d(t)$ with variance magnitude and with disturbance distribution vector:

$$E = [1 \quad 0 \quad 0 \quad 0]^T$$

The considered fault is an actuator fault acting on the second actuator, so that the fault distribution vector is

$$B_f = [0 \quad 0 \quad 0 \quad 1]^T$$

The proof that $\Pi$ is non-singular follows from $\Xi(\varphi)$. As $\Pi$ is block diagonal, its determinant is given by $\Xi(\varphi)$. It is thus only required to show that $m_{11}m_{22} \neq m_{12}m_{21}$ . $m_{12} = m_{21}$ (by symmetry) and $m_{11} > m_{22}$since $I_1 > I_2$ and $m_1 > m_2$, hence $\Pi$ is non-singular.

*B. Observer based residual Generator Design*

Following the proposed procedure, the residual generator design is shown in this subsection.

**Step 1**: The desired observer eigenvalues are set to be parametric to obtain more design freedom as:

$$\Lambda = \{\lambda_1 \quad \lambda_2 \quad \lambda_3 \quad \lambda_4\}.$$

**Step 2:** Using elementary transformation and the rational matrix factorization method, the following are obtained:

$$N(\theta, s) = \begin{bmatrix} s & 0 \\ 0 & s \\ 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$M(\theta, s) = \begin{bmatrix} s^2 + \dfrac{6.5346\rho_2}{0.0833 + \rho_1^2 - \rho_1} & \dfrac{-9.8(2\rho_1 - 1)\rho_2}{0.0833 + \rho_1^2 - \rho_1} \\ \dfrac{-2.45 * (2\rho_1 - 1)\rho_3}{0.0833 + \rho_1^2 - \rho_1} & s^2 + \dfrac{2.45\rho_3}{0.0833 + \rho_1^2 - \rho_1} \end{bmatrix}.$$

**Step 3:** Note that

$$CE = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$

So $p = 1$, and a basis for $LKer(CE)$may be taken as

$$\xi^T = [0 \quad 1]$$

Hence $Q = [0 \quad \alpha]$

**Step 4:** Then noting that:

$$QC = [0 \quad \alpha \quad 0 \quad 0],$$

one desired left eigenvector is $[0 \quad \alpha \quad 0 \quad 0]^T$and other left eigenvectors can be chosen arbitrarily only to satisfy $det(L) \neq 0$ . So, other parameters are given in a parametric way from:

$$f_i^T = [x_{i1} \quad x_{i2}], i = 2,3,4$$

Using the parametric eigenstructure assignment approach, the first desired eigenvector is projected into the allowable subspace by setting:

$$f_1^T = [N^T(\lambda_1)N(\lambda_1)]^{-1}N^T(\lambda_1)C^T\alpha\xi = \begin{bmatrix} 0 \\ \alpha\dfrac{\lambda_1}{\lambda_1^2 + 1} \end{bmatrix}$$

**Step 5:** The achieved eigenvector matrix can be obtained by (6) as given in **Theorem 1**.

**Step 6:** To simplify the calculation, some parameters are chosen at this step. If the parameters are chosen as:

$$x_{21} = 1, x_{22} = 0, x_{31} = 0,$$

$$x_{32} = 1, x_{41} = 0, x_{42} = 1.$$

Using$K(\theta) = W(\theta)R(\theta)^{-1}$. The observer gain is obtained as in (8).

$$K = \begin{bmatrix} (\lambda_3 + \lambda_2) & \frac{(\lambda_1 - \lambda_3)(\lambda_3 - \lambda_4)}{\lambda_2 - \lambda_3} \\ 0 & (\lambda_1 + \lambda_4) \\ -\lambda_3\lambda_2 + \frac{6.53446\rho_2}{K_o} & \lambda_3\lambda_2 + \frac{-2.45(2\rho_1 - 1)\rho_3}{K_o} \\ \frac{-9.8(2\rho_1 - 1)\rho_2}{K_o} & \lambda_1\lambda_4 + \frac{8.06642\rho_3}{K_o} \end{bmatrix} \quad (8)$$

$$K_o = (-0.764 + 2.5\rho_1^2 - 2.5\rho_1)$$

**Step 7:** The achieved closed-loop eigenvalues are:

$$\Lambda = \{\lambda_1 \quad \lambda_2 \quad \lambda_3 \quad \lambda_4\}.$$

To stabilize the observer system, let $real(\lambda_i) < 0, i = 1,2,3,4$. The required transient response performance can be achieved by suitably choosing the parametric eigenvalues.

The achieved transfer function matrix between the residual and disturbance is:

$$G_{rd}(s) = [0]$$

It is apparent that the disturbance is completely de-coupled because the transfer function matrix between the residual and disturbance is zero.

The transfer function between the residuals and faults is:

$$G_{rf}(s) = \frac{\alpha}{s^2 - (\lambda_1 + \lambda_4)s + \lambda_1\lambda_4}$$

So, the steady-state gain matrix $G_{rf}(0)$

$$G_{rf}(0) = \frac{\alpha}{\lambda_1\lambda_4}$$

From the above it is easy to choose suitable values of $\lambda_1$ *and* $\lambda_4$ to achieve a desired transient response and set $\alpha = \lambda_1\lambda_4$ to obtain good steady-state fault estimation.

The above analysis shows that the disturbances are decoupled completely and the residual is sensitive to the fault. This implies that the proposed design approach has achieved the desired goals. In the next subsection, some Simulink results are given.

### *C. Simulation result*

The open-loop two-link manipulator is unstable. Therefore, a constant controller is designed first using an observer state feedback structure while the estimated state is provided by the designed residual generator. The LPV system is simulated with a step and sinusoidal signals as shown in Figs. 2 & 3. Noting that the initial estimation is not good because both the two-link manipulator and observer systems are not in steady state and the state estimation error is large during the transient phase. The fault estimation is close to the real fault signal after 2 seconds as shown both in Figs. 2 & 3. The LPV fault estimator can provide good estimation performance when the real system is in a steady state.

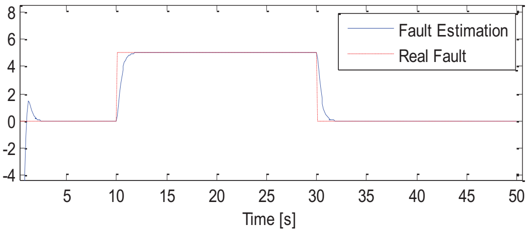


Figure 2. Fault estimation result for step fault signal

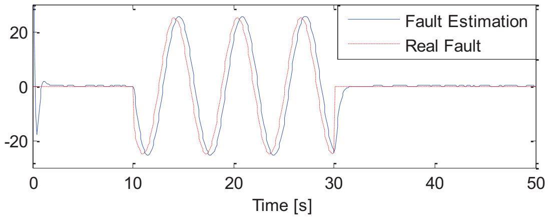


Figure 3. Fault estimation result for sinusoidal signal

## IV. Conclusion

This paper proposes an LPV fault detection approach using eigenstructure assignment which is robust in the sense of disturbance decoupling. The disturbances which can be considered to represent modeling uncertainty can be completely decoupled if the disturbance distribution and output matrix satisfy a rank condition. If this is not the case the disturbance can be decoupled as much as possible by suitable choice of design freedom. A two-link manipulator case is studied to show the usefulness of the proposed design procedure. The Simulink results show that for one fault case, the designed residual generator works well. Future studies will be concerned with multi-fault cases and how to use the detected fault information to accommodate faults and improve the system reliability.

## V. References

[1] R. J. Patton, "Robustness in model-based fault diagnosis: The 1995 situation," Annual Reviews in Control, vol. 21, pp. 103-123, 1997.

[2] H. Inseok, K. Sungwan, K. Youdan, and C. E. Seah, "A Survey of Fault Detection, Isolation, and Reconfiguration Methods," IEEE Trans. on Control Systems Technology, vol. 18, pp. 636-653, 2010.

[3] S. X. Ding, P. Zhang, T. Jeinsch, E. L. Ding, P. Engel, and W. Gui, "A survey of the application of basic data-driven and model-based methods in process monitoring and fault diagnosis," in IFAC World Congress, Milano, Italy, 2011.

[4] J. Chen and R. J. Patton, Robust model-based fault diagnosis for dynamic systems. London: Kluwer Academic Publishers, 1999.

[5] Y. Zhang and J. Jiang, "Bibliographical review on reconfigurable fault-tolerant control systems," Annual Reviews in Control, vol. 32, pp. 229-252, 2008.

[6] R. J. Patton and J. Chen, "Robust fault detection using eigenstructure assignment: a tutorial consideration and some new results," in the 30th IEEE Conference on Decision and Control, 1991, pp. 2242-2247.

[7] G. Duan, G. W. Irwin, and G. Liu, "Disturbance attenuation in linear systems via dynamical compensators: a parametric eigenstructure assignment approach," IEE Proceedings on Control Theory and Applications, vol. 147, pp. 129-136, 2000.

[8] G. Cai, C. Hu, and G. Duan, "Eigenstructure assignment for linear parameter-varying systems with applications," Mathematical and Computer Modelling, vol. 53, pp. 861-870, 2011.

[9] G. Duan, "Solution to matrix equation AV + BW = EVF and eigenstructure assignment for descriptor systems," Automatica, vol. 28, pp. 639-642, 1992.

[10] G. R. Duan, "Solutions of the equationAV+BW=VF and their application to eigenstructure assignment in linear systems," IEEE Trans. on Automatic Control, vol. 38, pp. 276-280, 1993.

[11] G.-R. Duan, "On the solution to the Sylvester matrix equation AV+BW=EVF," Automatic Control, IEEE Transactions on, vol. 41, pp. 612-614, 1996.

[12] R. J. Patton and J. Chen, "On eigenstructure assignment for robust fault diagnosis," International Journal of Robust and Nonlinear Control, vol. 10, pp. 1193-1208, 2000.

[13] G. P. Liu and R. J. Patton, "Parametric state feedback controller design of multivariable systems," International Journal of Control, vol. 61, pp. 1457-1464, 1995/06/01 1995.

[14] G. P. Liu and R. Patton, Eigenstructure Assignment for Control System Design: John Wiley& Sons, Inc., 1998.

[15] H. Wang, J. Wang, and J. Lam, "Worst-Case Fault Detection Observer Design: Optimization Approach," Journal of Optimization Theory and Applications, vol. 132, pp. 475-491, 2007.

[16] H. Wang and J. Lam, "Robust fault detection for uncertain discrete-time systems " Journal of Guidance, Control and Dynamics, vol. 25, pp. 291-301, 2002.

[17] H. Wang, J. Wang, and J. Lam, "An optimization approach for worst-case fault detection observer design," in American Control Conference, 2004.

[18] J. S. Shamma and M. Athans, "Analysis of gain scheduled control for nonlinear plants," IEEE Transactions on Automatic Control, vol. 35, pp. 898-907, 1990.

[19] W. J. Rugh, "Analytical framework for gain scheduling," IEEE Trans. on Control Systems, vol. 11, pp. 79-84, 1991.

[20] W. J. Rugh and J. S. Shamma, "Research on gain scheduling," Automatica, vol. 36, pp. 1401-1425, 2000.

[21] D.J.Leith and W. E.Leithead, "Survey of Gain-Scheduling Analysis and Design," Journal of Control, vol. 73, pp. 1001--1025, 1999.

[22] L. Chen and R. J. Patton, "Polytope LPV estimation for non-linear flight control," presented at the World Congress, Milano, Italy, 2011.

[23] R. Patton, L. Chen, and S. Klinkhieo, "An LPV pole-placement approach to friction compensation as an FTC problem," J. of Applied Mathematics and Computer Science, vol. 22, pp. 149-160, 2012.

[24] J. Magni, "Multimodel eigenstructure assignment in flight-control design," Aerospace Science and Technology, vol. 3, pp. 141-151, 1999.

[25] C. Döll, Y. Le Gorrec, G. Ferreres, and J. F. Magni, "A robust self-scheduled missile autopilot: design by multi-model eigenstructure assignment," Control Engineering Practice, vol. 9, pp. 1067-1078, 2001.

[26] G. Duan, G. Wu, and W. Huang, "Eigenstructure Assignment for Linear Time-Varying Systems," Science in China (Series A, English Eition), vol. 34, pp. 246-256, 1991.

[27] S. B. Niku, Introduction to Robotics: Prentice Hall Professional Technical Reference, 2001.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Bilinear approach employed for modelling of continuous stirred tank reactor processes

Tomasz Larkowski*, Leszek Koszalka† and Keith J. Burnham*†
* Control Theory and Applications Centre
Coventry University, CV15FB Coventry, UK, Email: t.larkowski@coventry.ac.uk
† Computer Systems and Networks
Wroclaw University of Technology, Wroclaw, Poland,

***Abstract*—In this paper feasibility of modelling approach based on a bilinear system approximation is demonstrated on one of the most frequently met processes in chemical engineering, namely a continuous stirred tank reactor. Selected examples of such systems from the literature are considered and modelled with a use of dynamic bilinear systems. Advantages of this approach are presented and discussed.**

## I. Introduction

Chemical reactor is often described as 'the most important unit operation in a chemical process', see [1]. The task of its modelling and control is commonly encountered in the literature representing a practical nonlinear industrial problem, see [2], [3], [4]. A popular model of a chemical reactor is the so-called continuous stirred tank reactor (CSTR). CSTR consists of a closed tank to which an input stream is fed in and the output stream fed out in a continuous manner, whilst a content is constantly stirred. Modelling of CSTRs is challenging mainly due to i) possibility of occurrences of rapid reactions (the so-called ignition-extinction phenomena), hence fast changing process gain and dynamics, and also due to ii) nonlinear steady-state behaviour, see [1] and [5].

This paper demonstrates that the CSTRs can be modelled by employing an approach based on bilinear system (BS) description. In order to increase modelling flexibility of BSs, an extension consisting of a static nonlinearity that transforms the input signal is proposed. Such a structure, referred to as a Hammerstein-bilinear system (HBS), see [6], [7], is considered and compared to BS and Hammerstein system (HS) models.

The use of BS based approach is motivated twofold. First, BSs retain a close structural relationship with linear models, hence standard well understood notions from classical linear system theory such as system time constants, damping/natural frequency and steady-state gain are to large extent retained. This follows from the property that BS structure can be interpreted as a linear time-varying system, which also greatly facilitates the control design. Second, BSs preserve linearity w.r.t. the parametrisation, which aids in their identification by allowing for standard parameter estimation methods to be used.

In this paper three different CSTR models are considered. Two of the models are isothermal, whilst the third model is an example of a diabatic CSTR, see [1].

## II. Model structures

HBS structure belongs to a sub-class of so-called output affine models, i.e. models that retain affinity w.r.t. the output signals, see [8]. It comprises of a cascade connection of a static (memoryless) nonlinearity followed by a dynamic time-invariant affine BS and is given by

$$y_k = \sum_{j=1}^{n_a} a_j y_{k-j} + \sum_{i=1}^{n_b} b_i v_{k-i} + \sum_{j=1}^{n_a}\sum_{i=1}^{n_b} \eta_{ij} v_{k-i} y_{k-j} + c \quad (1)$$

$$v_k = f(u_k) \quad (2)$$

where $a_j$, $b_i$, $\eta_{ij}$ and $c$ are model parameters. The bilinearity is defined as a product between system output $y_k$ and the intermediate input variable $v_k$, and $f(\cdot)$ denotes a general scalar static nonlinear function. Note that not all bilinear coefficients must necessarily be present in (1), hence a particular structure can be obtained by setting selected $\eta_{ij}$ to zero.

The HBS can be interpreted as a generalisation of both of its constituent subsystems, i.e. HS and BS models. In particular, a BS is obtained from (1)-(2) by setting $u_k = v_k$, i.e. by selecting $f(x) = x$, which gives

$$y_k = \sum_{j=1}^{n_a} a_j y_{k-j} + \sum_{i=1}^{n_b} b_i u_{k-i} + \sum_{j=1}^{n_a}\sum_{i=1}^{n_b} \eta_{ij} u_{k-i} y_{k-j} + c \quad (3)$$

Similarly, a HS is obtained by setting $\eta_{ij} = 0 \ \forall \ i, j$ in (1)-(2), which leads to

$$y_k = \sum_{j=1}^{n_a} a_j y_{k-j} + \sum_{i=1}^{n_b} b_i v_{k-i} + c \quad (4)$$

$$v_k = f(u_k) \quad (5)$$

Also, a linear (or more precisely an affine) structure is obtained by imposing both restrictions simultaneously, i.e. $\eta_{ij} = 0 \ \forall i, j$ and $u_k = v_k$.

In this paper, for simplicity, it is assumed that the input static nonlinearity is modelled as a polynomial of order $n_\alpha$, i.e.

$$f(x) = \sum_{l=1}^{n_\alpha} \alpha_l x^l \quad (6)$$

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

Consequently, a particular HBS structure is given by a quadruplet that defines the number of $a$, $b$, $\eta$ and $\alpha$ coefficients, i.e. HBS($n_a$,$n_b$,$n_\eta$,$n_\alpha$). Moreover, the sum of all coefficients plus unity (to account for an offset), corresponds to the total number of degrees of freedom (DoF) in a given structure. An analogous notation is used w.r.t. other structures that can be derived from the HBS model.

## III. Parameter estimation

The HBS structure is bilinear in terms of parametrisation, due to the products between $\alpha$ and $a$, and also between $\alpha$ and $\eta$ coefficients. A well known approach to solve such problems is to use the so-called bilinear parametrisation method (BPM), see [9]. The BPM solves the estimation problem in a two step manner, where in the first step $\alpha$ parameters are fixed and $a$, $b$, $\eta$ parameters are calculated, whilst in the second step $a$, $b$ and $\eta$ remain fixed and $\alpha$ parameters are computed [10]. Because the two subproblems separately are linear w.r.t. the unknowns, the ordinary least squares algorithm [9] can be applied, which renders the overall procedure numerically efficient, and this is the approach used here. An analogous technique can also be applied to HS models, see [11], whilst the parameters of affine and BS structures can be estimated by using a single ordinary least squares technique.

## IV. Simulation studies

### A. Performance criteria and experimental setup

In order to quantify the accuracy of models obtained, two performance criteria are used. Namely, the coefficient of determination and the (normalised) integral of absolute error, defined, respectively, as follows

$$\mathrm{R}_\mathrm{T}^2 = 100\left(1 - \frac{\|y-\hat{y}\|_2^2}{\|y-\bar{y}\|_2^2}\right) \tag{7}$$

$$\mathrm{IAE} = \frac{1}{N}\sum_{k=1}^{N}|y_k - \hat{y}_k| \tag{8}$$

where $y$ and $\hat{y}$ denote vectors composed of the measured (noisy) system outputs and outputs generated by the estimated model, respectively, and $\bar{y}$ is the mean value of $y$. The notation $\|\cdot\|_2$ denotes the Euclidean norm.

Because the main interest of experiments lies in the modelling capabilities of the model structures considered, relatively long input-output data consisting of $N = 20,000$ samples are used. Three data sets are considered, i.e. identification and two validation data sets. The sampling time is chosen as $0.1$s. In the case of the identification data set and the first validation data set, the input signal is generated as a series of uniformly distributed steps between the minimal and maximal range for a given system. The probability of transition to a different level is selected randomly with a uniform switching probability of $10\%$, providing a reasonable compromise between the content of transient and steady-state data. Additionally, to ensure that the input is sufficiently exciting, a normally distributed, white and zero-mean noise sequence of comparatively small variance is added. To render the identification experiment more realistic, the measured output is assumed to be contaminated with an additive, normally distributed, white and zero-mean disturbances such that the resulting signal-to-noise ratio is approximately 37dB. The second validation data set comprises of a monotonic staircase input, which allows the performance of the identified models to be evaluated with the emphasis placed on the steady-state behaviour. Also, to provide an indication of the complexity of the models, the corresponding DoF are considered.

### B. Isothermal CSTR with a first-order irreversible reaction

*1) System description:* The first isothermal CSTR model considered, referred to as the CSTR1, is given by the following equations, see [1] for details, i.e.

$$\frac{dC_A(t)}{dt} = \frac{F(t)}{V}C_{Af} - \left(\frac{F(t)}{V} + k\right)C_A(t) \tag{9}$$

$$\frac{dC_B(t)}{dt} = -\frac{F(t)}{V}C_B(t) + kC_A(t) \tag{10}$$

and describes a first-order irreversible reaction $A \xrightarrow{k} B$ where $k$ is the reaction rate per unit volume. The remaining variables are: $C_A(t)$, $C_B(t)$ - concentrations of substances $A$ and $B$ inside the tank of a constant volume $V$, respectively, $F(t)$ - inflow/outflow mass rate, $C_{Af}$ - inflow concentration of substance $A$. Only the substance $A$ is present in the inflow stream, and inflow and outflow mass rates are equal. The actual units are unimportant and hence are not included. It is assumed that the manipulated variable is $F(t)$ and that $C_B(t)$ is the output of interest. The task consists of identifying a model between $F(t)$ and $C_B(t)$.

The values of the parameters were chosen as $V = 1$, $k = 0.2$, $C_{Af} = 1$ and the initial states of the process as $C_{a0} = C_{b0} = 0.5$, where the subscript zero denotes the initial value. The input $F(t)$ is in range of $(0,\ 1]$.

It is observed that a product, i.e. bilinearity, between the input $F(t)$ and the state $C_A(t)$ occurs in (9) and that an analogous product between the input $F(t)$ and the state $C_B(t)$ is also present in (10). The steady-state characteristic of the CSTR1 model is plotted in Figure 1, where it is observed that the curve resembles the type of steady-state characteristics typical of BSs. Therefore, these observations substantiate the usage of a bilinear based modelling approach.

*2) Identification results:* Selected identification results are given in Table I. First, it is observed that the model is considerably nonlinear because the third order affine structure resulted in $\mathrm{R}_\mathrm{T}^2$ below 90% for all three data sets. Further increase of the order of the affine structure does not lead to any significant improvements in modelling performance. HS models show clear improvement, allowing for $\mathrm{R}_\mathrm{T}^2$ of approximately 98% in the case of the HS(1,1,4) to be achieved for all three data sets. Further increase of the order of the input polynomial does not lead to significant improvements in model fitting. This is due to the fact that the nonlinear steady-state characteristic is not a complex function, and it is rather the changing system dynamics that is not captured by the HS type structures. When considering the results obtained from BS models, an evident

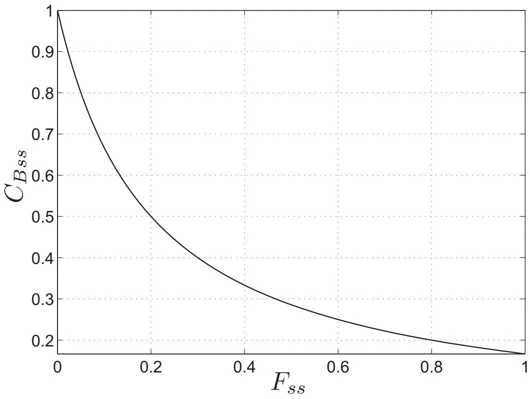


Fig. 1. CSTR1 - A steady-state characteristic.

| structure | | Id. data set | | Val. data set 1 | | Val. data set 2 | |
|---|---|---|---|---|---|---|---|
| | DoF | $R_T^2$ | IAE $\times 10^{-3}$ | $R_T^2$ | IAE $\times 10^{-3}$ | $R_T^2$ | IAE $\times 10^{-3}$ |
| affine(3,3) | 7 | 80.40 | 54.49 | 80.08 | 55.27 | 88.11 | 74.10 |
| HS(1,1,2) | 5 | 96.70 | 24.04 | 96.10 | 25.13 | 94.78 | 31.00 |
| HS(1,1,3) | 6 | 97.81 | 15.51 | 97.83 | 17.16 | 96.92 | 19.39 |
| HS(1,1,4) | 7 | 98.10 | 13.75 | 98.10 | 15.68 | 97.49 | 17.57 |
| BS(1,1,1) | 4 | 99.97 | 2.265 | 99.97 | 2.276 | 99.97 | 2.716 |
| BS(2,1,1) | 5 | 99.98 | 1.845 | 99.98 | 1.851 | 99.99 | 1.983 |
| HBS(1,1,1,2) | 6 | 99.97 | 2.275 | 99.97 | 2.304 | 99.97 | 2.748 |
| HBS(1,1,1,3) | 7 | 99.99 | 1.421 | 99.97 | 2.260 | 99.99 | 1.751 |

TABLE I
CSTR1 - QUANTIFIED IDENTIFICATION AND VALIDATION RESULTS FOR MODEL STRUCTURES CONSIDERED.

improvement in the approximation performance is noted, i.e. the $R_T^2$ of almost 100% is obtained with the IAE criterion decreasing by approximately 7 times, when compared to the best HS model. This means that both, the nonlinear steady-state characteristic and the changing system dynamics are approximated well by the BS structures with only 4 or 5 DoF. Because the only source of nonlinearity in the underlying process equations arises from product terms, cf. (9)-(10), this result could have been anticipated.

Representative graphical results of the identification are given in Figures 2 and 3, showing the performances of the selected estimated models on arbitrarily chosen intervals of the validation data sets 1 and 2, respectively. It is observed that in the case of both figures the actual system output is virtually undistinguishable from that generated by the identified BS(2,1,1) structure. Consequently, it is concluded that a BS structures are appropriate for modelling the CSTR1 process.

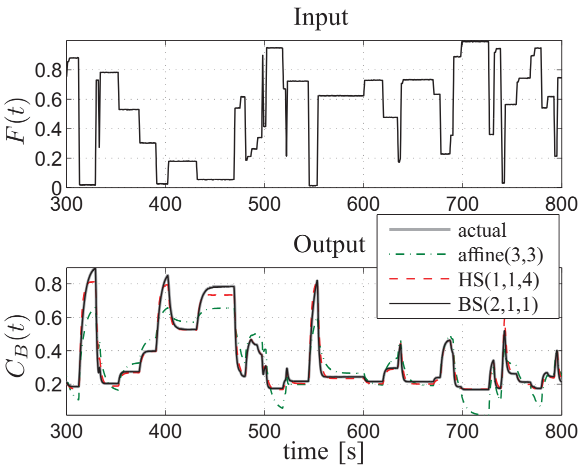


Fig. 2. CSTR1 - Selected representative results of identification on validation data set 1 in the interval $[300,\ 800]$s.

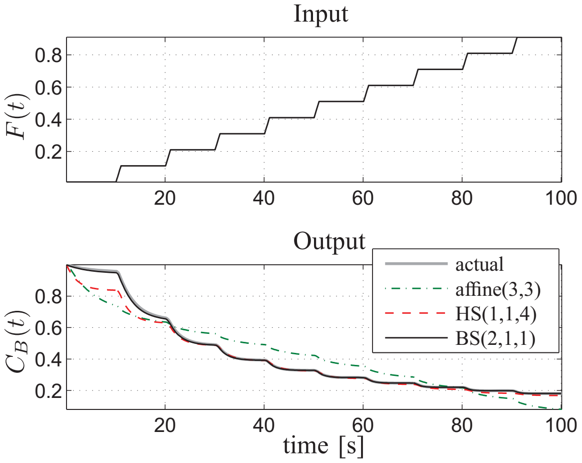


Fig. 3. CSTR1 - Selected results of identification on validation data set 2.

### C. *Isothermal CSTR with the Van de Vusse reaction*

*1) System description:* The second isothermal CSTR model considered, see [1], referred to as the CSTR2, is given by

$$\frac{dC_A(t)}{dt} = \frac{F(t)}{V}\big(C_{Af} - C_A(t)\big) - k_1 C_A(t) - k_3 C_A^2(t) \quad (11)$$

$$\frac{dC_B(t)}{dt} = -\frac{F(t)}{V} C_B(t) + k_1 C_A(t) + k_2 C_B(t) \quad (12)$$

$$\frac{dC_C(t)}{dt} = -\frac{F(t)}{V} C_C(t) + k_2 C_B(t) \quad (13)$$

$$\frac{dC_D(t)}{dt} = -\frac{F(t)}{V} C_D(t) + \frac{1}{2} k_3 C_A^2(t) \quad (14)$$

with the behaviour governed by the so-called Van de Vusse reaction kinetics. The reactions

$$A \xrightarrow{k_1} B \xrightarrow{k_2} C \quad (15)$$

$$2A \xrightarrow{k_3} D \quad (16)$$

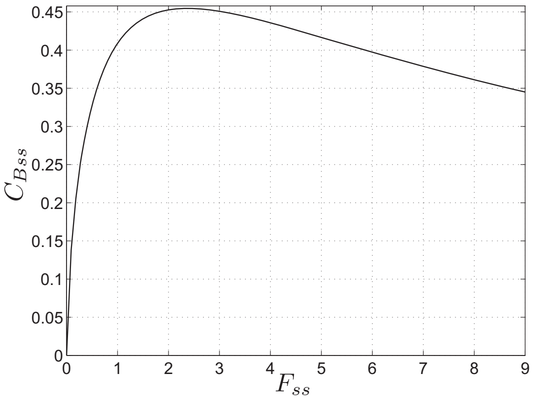


Fig. 4. CSTR2 - A steady-state characteristic.

| structure | | Id. data set | | Val. data set 1 | | Val. data set 2 | |
|---|---|---|---|---|---|---|---|
| | DoF | $\mathrm{R}_{\mathrm{T}}^2$ | IAE $\times 10^{-3}$ | $\mathrm{R}_{\mathrm{T}}^2$ | IAE $\times 10^{-3}$ | $\mathrm{R}_{\mathrm{T}}^2$ | IAE $\times 10^{-3}$ |
| affine(3,3) | 7 | 10.92 | 44.68 | 7.724 | 46.06 | 5.539 | 54.19 |
| HS(1,1,4) | 7 | 84.07 | 15.67 | 84.03 | 16.86 | 89.50 | 25.10 |
| HS(1,1,5) | 8 | 91.48 | 10.70 | 91.16 | 12.38 | 94.40 | 16.13 |
| HS(1,1,6) | 9 | 94.99 | 8.552 | 94.56 | 8.803 | 97.32 | 11.28 |
| HS(1,1,7) | 10 | 96.43 | 6.591 | 96.09 | 6.800 | 98.61 | 8.118 |
| HBS(1,1,1,3) | 7 | 92.96 | 10.72 | 92.78 | 11.43 | 98.23 | 11.31 |
| HBS(1,1,1,4) | 8 | 98.11 | 4.335 | 98.11 | 4.726 | 99.86 | 3.000 |
| HBS(1,1,1,5) | 9 | 98.34 | 3.257 | 98.33 | 3.551 | 99.60 | 3.747 |

TABLE II
CSTR2 - QUANTIFIED IDENTIFICATION AND VALIDATION RESULTS FOR MODEL STRUCTURES CONSIDERED.

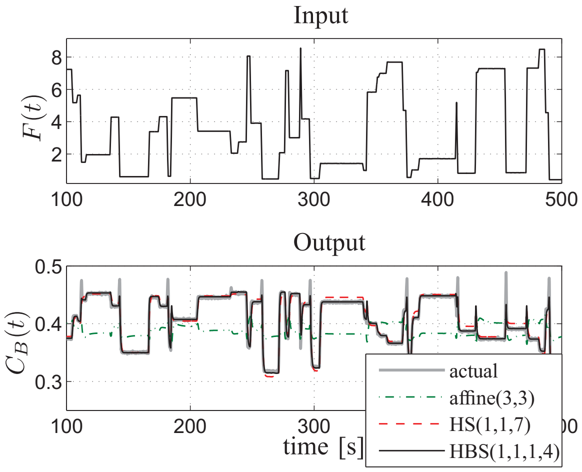


Fig. 5. CSTR2 - Selected representative results of identification on validation data set 1 in the interval [100, 500]s.

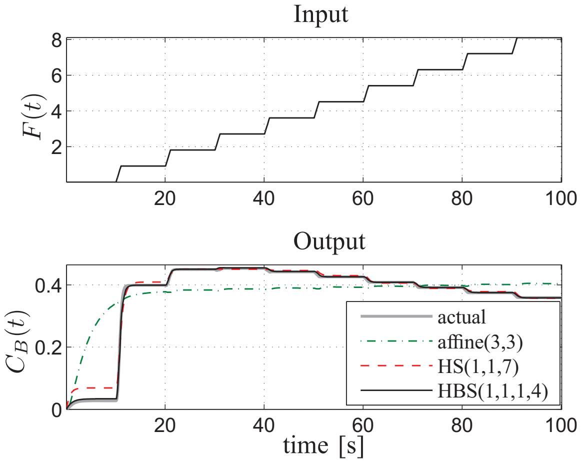


Fig. 6. CSTR2 - Selected results of identification on validation data set 2.

are irreversible and described by the reaction rate constants $k_1$, $k_2$ and $k_3$. The remaining variables are: $C_A(t)$, $C_B(t)$, $C_C(t)$, $C_D(t)$ - concentrations of substances $A$, $B$, $C$ and $D$ inside the tank of constant volume $V$, respectively, $F(t)$ - inflow/outflow mass rate, $C_{Af}$ - inflow concentration of substance $A$. Only the substance $A$ is present in the inflow stream, and inflow and outflow mass rates are equal. It is assumed that the manipulated variable is $F(t)$, whilst $C_B(t)$ is the output of interest. Therefore, the modelling task consists of identifying a model between $F(t)$ and $C_B(t)$.

The values of the parameters were chosen as: $V = 1$, $k_1 = 5/6$, $k_2 = 3$, $k_3 = 10/6$, $C_{Af} = 10$ and the initial states of the process are $C_{a0} = C_{b0} = 0$. The input $F(t)$ is postulated to vary between $(0, 9]$.

It is observed that a bilinearity between the input $F(t)$ and the state $C_A(t)$ is present in (11) and an analogous bilinearity between the input $F(t)$ and the state $C_B(t)$ appears in (12). However, in contrast to (9) corresponding to the CSTR1 model, here the product term is not the only source of nonlinearity in the evolution of $C_A(t)$. This is due to the presence of the expression $C_A^2(t)$ in (11), whose influence is controlled by the rate constant $k_3$.

Consequently, it is concluded that although the underlying system equations indicate that a bilinear type behaviour is present, it may be infeasible to model the process by a BS only over the entire range of its operation. This hypothesis is confirmed by considering the steady-state characteristic of the CSTR2 process given in Figure 4, where it is observed that the system exhibits the input multiplicity (IM) property [3]. Such static behaviour cannot be captured by a BS, due to its inherent structural limitations. This stands in contrast to HS and HBS that are structurally capable of IM, hence are anticipated to be more appropriate for modelling of the CSTR2 process.

*2) Identification results:* The identification results are given in Table II, where a low value of $\mathrm{R}_{\mathrm{T}}^2$ for all three data sets in the case of the affine model indicates that the overall process is considerably nonlinear. Further increase of the order of the affine structure does not result in noticeable improvements in fitting. The performance criteria of BS structures are not included in the table because it was not possible to fit such models, i.e. the corresponding $\mathrm{R}_{\mathrm{T}}^2$ values were negative. This result confirms the inappropriateness of using BS structures for approximating processes exhibiting IM. A considerable improvement is achieved in the case of HS structures, when compared to the affine model, with $\mathrm{R}_\mathrm{T}^2$ of about 97% on average for all three data sets obtained by HS(1,1,7) with 10 DoF. The results obtained by the HBS(1,1,4) and HBS(1,1,5), with 8 and 9 DoF, respectively, show further fitting improvements with both models achieving $\mathrm{R}_\mathrm{T}^2$ of approximately 98% for the identification data set and the validation data set 1, and almost 100% for the validation data set 2. Compared to the HS(1,1,7) the IAE was reduced by approximately threefold in the case of the first two data sets, and fourfold in the case of the third data set.

Representative graphical results obtained from the identification procedure are presented in Figures 5 and 6, showing the performances of the selected estimated models on arbitrarily chosen intervals of the validation data set 1 and 2, respectively. Whilst in the case of Figure 5 slight discrepancies between the actual output and that generated by the estimated HBS(1,1,1,4) model are observed, the two corresponding curves are virtually indistinguishable in the case of Figure 6. Also, it is noted that the system exhibits a non-minimum phase behaviour when $F(t)$ is high and changes to a lower value. This behaviour, which is observed to be manifested by spikes in the $C_B(t)$ signal, increases the difficulty of obtaining an acceptable approximation. Consequently, despite such a challenging task, the modelling results obtained by the HBS(1,1,1,4) structure can be treated as very satisfactory in overall.

### D. *Diabatic CSTR*

*1) System description:* The considered diabatic CSTR model, see [1] and [5], referred to as the CSTR3, is given by

$$\frac{dC_A(t)}{dt} = \frac{F(t)}{V}\big(C_{Af} - C_A(t)\big) - k_0 r(t) C_A(t) \tag{17}$$

$$\frac{dT(t)}{dt} = \frac{F(t)}{V}\big(T_f - T(t)\big) + \frac{-\Delta H}{\rho c_p} k_0 r(t) - \frac{US}{V\rho c_p}\big(T(t) - T_j(t)\big) \tag{18}$$

where the first order reaction rate per unit volume is given by the so-called Arrhenious expression, i.e.

$$r(t) = \exp\left(\frac{\Delta E}{RT(t)}\right) \tag{19}$$

The other variables are: $C_A(t)$ - concentration of substance $A$ inside a tank of the constant volume $V$, $F(t)$ - inflow/outflow mass rate, $C_{Af}$ - inflow concentration of substance $A$, $k_0$ - pre-exponential factor, $R$ - ideal gas constant, $\Delta E$ - activation energy, $T(t)$ - reactor temperature, $T_f$ - inflow (feed) temperature, $T_j(t)$ - jacket temperature, $U$ - overall heat transfer coefficient, $-\Delta H$ - heat of reaction, $\rho$ - density, $S$ - area for heat exchange, $c_p$ - heat capacity. Only the substance $A$ is present in the inflow stream, and inflow and outflow mass rates are equal. It is assumed that the manipulated variable is $T_j(t)$, whilst $C_A(t)$ is the output of interest. Therefore, the modelling task consists of identifying a model between $T_j(t)$ and $C_A(t)$.

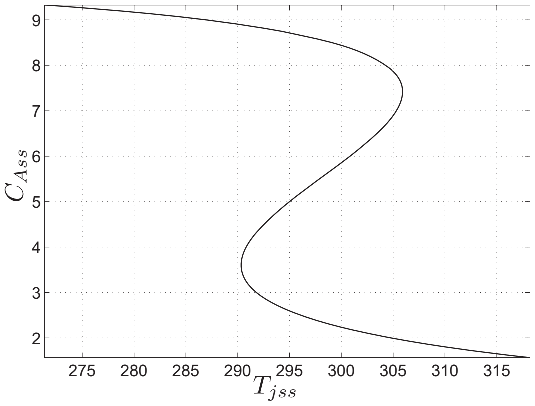


Fig. 7. CSTR3 - A steady-state characteristic.

| structure | | Id. data set | | Val. data set 1 | | Val. data set 2 | |
|---|---|---|---|---|---|---|---|
| | DoF | $\mathrm{R}_\mathrm{T}^2$ | IAE $\times 10^{-3}$ | $\mathrm{R}_\mathrm{T}^2$ | IAE $\times 10^{-3}$ | $\mathrm{R}_\mathrm{T}^2$ | IAE $\times 10^{-3}$ |
| affine(3,3) | 7 | 88.86 | 99.53 | 88.84 | 99.38 | 90.73 | 96.42 |
| HS(1,1,3) | 6 | 96.92 | 49.26 | 96.88 | 49.52 | 96.65 | 54.41 |
| HS(1,1,5) | 8 | 97.31 | 45.51 | 97.30 | 45.63 | 97.20 | 49.03 |
| HS(1,1,7) | 10 | 97.89 | 39.41 | 97.86 | 39.74 | 98.00 | 40.15 |
| HS(2,2,3) | 8 | 97.19 | 45.61 | 97.15 | 45.88 | 98.46 | 40.48 |
| HS(2,2,5) | 10 | 97.61 | 41.52 | 97.57 | 41.80 | 98.79 | 36.90 |
| HS(2,2,7) | 12 | 98.16 | 35.46 | 98.12 | 35.77 | 99.16 | 31.27 |
| BS(1,1,1) | 4 | 98.98 | 27.52 | 98.96 | 27.09 | 99.20 | 23.82 |
| BS(2,1,1) | 5 | 98.90 | 28.35 | 98.88 | 28.75 | 99.14 | 25.08 |
| BS(2,2,2) | 7 | 99.16 | 23.35 | 99.13 | 23.90 | 99.58 | 16.72 |
| HBS(1,1,1,3) | 7 | 99.13 | 25.63 | 99.11 | 26.00 | 99.15 | 24.77 |
| HBS(1,1,1,5) | 9 | 99.19 | 24.61 | 99.18 | 24.98 | 99.21 | 23.58 |
| HBS(1,1,1,7) | 11 | 99.28 | 22.50 | 99.27 | 22.97 | 99.36 | 20.16 |
| HBS(2,2,2,3) | 10 | 99.39 | 20.96 | 99.37 | 21.40 | 99.52 | 18.68 |
| HBS(2,2,2,5) | 12 | 99.46 | 19.53 | 99.44 | 20.05 | 99.58 | 16.96 |
| HBS(2,2,2,7) | 14 | 99.56 | 17.25 | 99.54 | 17.84 | 99.70 | 13.52 |

TABLE III
CSTR3 - QUANTIFIED IDENTIFICATION AND VALIDATION RESULTS FOR MODEL STRUCTURES CONSIDERED.

Considering equations (17) and (18), it is noted that in each case bilinearities are present, i.e. products between $F(t)$ and the state $C_A(t)$ in the first equation and between $F(t)$ and the state $T(t)$ in the second equation. These, however, are clearly not the only contributions that render nonlinearity of the overall behaviour. This is due to the presence of nonlinear relationships involving an exponent of $T(t)$ that appear in both equations.

A steady-state characteristic of the CSTR3 model, given in Figure 7, shows the presence of the output multiplicity (OM) property [3]. Because non of the model structures investigated in this paper is structurally capable of OM, see [8], only a restricted range of the process operation is considered, i.e. the range of $T_j \in [273,\ 306)$ within which OM is absent. Consequently, because the operating range is limited, it might be possible that a bilinear type behaviour will, in fact, be prevailing.

The values of the parameters were chosen as: $F(t) = 1$,

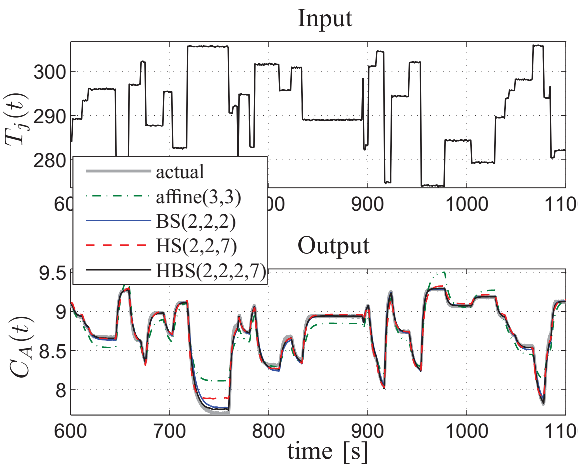


Fig. 8. CSTR3 - Selected representative results of identification on validation data set 1 in the interval $[600,\ 1100]$s.

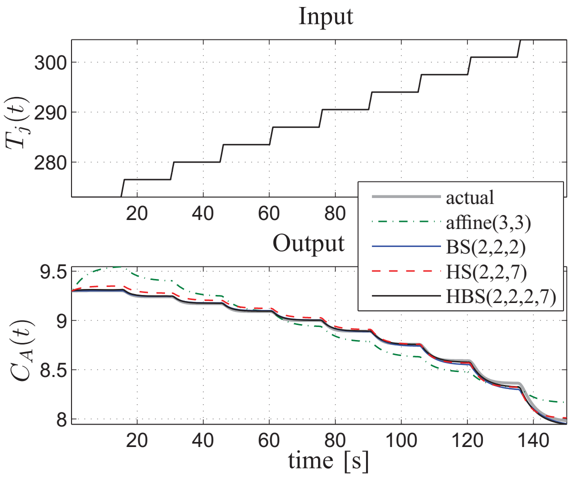


Fig. 9. CSTR3 - Selected results of identification on validation data set 2.

$V = 1$, $k_0 = 9703 \times 3600$, $-\Delta H = 5960$, $\rho c_p = 500$, $US = 150$, $\Delta E = 11843$, $R = 1.987$ and the initial states of the process are $C_{A0} = 8.5$ and $T_0 = 305$.

*2) Identification results:* The identification results obtained are collected in Table III, from where it is observed that the affine model achieved reasonable results close to 90% in terms of $\mathrm{R}_\mathrm{T}^2$ in the case of all three data sets. Further increase of the order of the affine structure does not provide any considerable improvements. First order HS structures yielded results that are better by approximately 8% on average in terms of the $\mathrm{R}_\mathrm{T}^2$ criterion and approximately twice on average in terms of the IAE criterion. This indicates a clear improvement and justifies the need for a nonlinear model structure. The best fitting among HS structures is obtained for a second order HS structure, i.e. HS(2,2,7) with a seventh order polynomial and 12 DoF in total. It is interesting to notice that these values are close to those produced by a relatively simple BS structure, i.e. BS(1,1,1), with only 4 DoF. A boundary of 99% in terms of the $\mathrm{R}_\mathrm{T}^2$ criterion is exceed by a second order BS model with 7 DoF. The fitting is improved, if at all, only slightly by first order HBS structures, and it is the second order HBS model, i.e. HBS(2,2,2,7), in the case of which the results improve more significantly. However, this comes at the cost of 14 DoF, when compared to only 7 DoF in the case of BS(2,2,2).

Representative graphical results generated from the estimated models are depicted in Figures 8 and 9, and demonstrate the performances of the selected models on arbitrarily chosen intervals of the validation data set 1 and 2, respectively. It is observed that the outputs of the estimated models BS(2,2,2) and HBS(2,2,2,7) are both virtually undistinguishable from the actual system output in both figures. Consequently, by taking into account the corresponding DoF and a pragmatic point of view, it is the second order BS that appears to be a preferable choice in this case.

## V. Conclusions

The paper has demonstrated feasibility of BS based modelling approach for approximating CSTRs. It has been shown that BS models are capable of capturing both, i.e. the dynamic and static behaviour of the exemplary CSTR systems considered. In the case of the CSTR process exhibiting the IM property, a BS structure with an additional nonlinear memoryless element transforming the input, i.e. a HBS structure, has shown to be an appropriate choice.

## References

[1] W. Bequette, *Process Dynamics: Modeling, Analysis and Simulation*, 1st ed., ser. Series in the Physical and Chemical Engineering Sciences. Prentice Hall PTR, 1998.

[2] E. Meadows and J. Rawlings, *Model predictive control.* Prentice Hall, Englewood Cliffs, NJ, 1997, ch. 5, pp. 233–310.

[3] M. Pottmann and R. K. Pearson, “Block-oriented NARMAX models with output multiplicities,” *American Institute of Chemical Engineers*, vol. 44, no. 1, pp. 131–140, 1998.

[4] K. Murakami and D. E. Seborg, “Constrained parameter estimation with applications to blending operations,” *J. of Process Control*, vol. 10, no. 2–3, pp. 195–202, 2000.

[5] G. Stephanopoulos, *Chemical Process Control: An Introduction to Theory and Practice.* Prentice Hall, 1983.

[6] I. Zajic, T. Larkowski, M. Sumislawska, D. Hill, and K. J. Burnham, “Modelling of an air handling unit: a Hammerstein-bilinear model identification approach,” in *Proc. of 21st Int. Conf. on Systems Engineering*, Las Vegas, USA, 2011, pp. 59–61.

[7] I. Zajic, T. Larkowski, D. Hill, and K. J. Burnham, “Energy consumption analysis of HVAC system with respect to zone temperature and humidity set-point,” in *Proc. of 18th IFAC World Congress*, Milan, Italy, 2011, pp. 4576–4581.

[8] R. K. Pearson, “Selecting nonlinear model structures for computer control,” *J. of Process Control*, vol. 13, no. 1, pp. 1–26, 2003.

[9] L. Ljung, *System Identification - Theory for the User*, 2nd ed. New Jersey, USA: Prentice Hall PTR, 1999.

[10] T. Larkowski and K. J. Burnham, “Bilinear approach to modelling of continuous stirred tank reactor process,” in *Proc. of 9th European Workshop on Advanced Control and Diagnosis*, Budapest, Hungary, 2011, pp. 168–173.

[11] F. Le, I. Markovsky, C. T. Freeman, and E. Rogers, “Identification of electrically stimulated muscle models of stroke patients,” *Control Engineering in Practice*, vol. 18, no. 4, pp. 396–407, 2010.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# The benefits of nonlinear cubic viscous damping on the force transmissibility of a Duffing-type vibration isolator

Carmen Ho, Zi-qiang Lang, Stephen A. Billings
Department of Automatic Control Systems Engineering
University of Sheffield, Mappin Street, Sheffield, UK
Email: {carmen.ho, z.lang, s.billings}@sheffield.ac.uk

*Abstract*—**Vibration isolation systems with nonlinear stiffness under sinusoidal excitation exhibit unwanted jump phenomena and superharmonics when they are lightly damped. These characteristics can be suppressed by linear viscous damping but the force transmissibility over the high frequency range increases as a result. In this study, nonlinear viscous damping will be chosen to solve this problem with the aid of a single-degree-of-freedom model with cubic stiffness. Simulation results show that nonlinear viscous damping can reduce the resonant peak as well as suppressing the adverse properties of nonlinear stiffness, jumps and harmonics, without compromising the transmissibility over the high frequency range. Nonlinear damping preserves the benefits of linear damping while removing the undesirable effects over the non-resonant regions and therefore improves the overall performance.**

**Keywords: nonlinear damping, nonlinear spring, jump phenomena, duffing, OFRF**

## I. Introduction

Passive vibration isolation systems are employed in many engineering applications where objects require protection from undesired vibration force or displacement. They are usually installed between the source of the disturbance and equipment requiring protection in order to reduce the level of vibration transmitted to the object. A single-degree-of-freedom (sdof) vibration isolator, modelled by a linear mass-spring-damper system, has been well-studied by many authors [1]–[9]. The need for nonlinear vibration isolation and the recent developments are summarised by Ibrahim [10].

The nonlinearity usually takes the form of either a nonlinear spring force or a nonlinear damping force. There are a number of works focusing on Duffing-type isolators where the restoring spring force is a cubic function of the displacement [11]–[14]. Compared to the linear case, the nonlinear stiffness term could lower the force transmissibility around the resonant frequency range. In some cases, the resonant frequency can be reduced, resulting in a larger frequency range of vibration isolation. One major drawback of nonlinear stiffness, however, is the jump phenomena, where the transmissibility suddenly jumps up or down when there is a small change in the frequency of the excitation signal. The inherent instability and the resulting high level of higher harmonics may cause some concerns. The effects of the nonlinear spring force on an sdof system was compared with that of the nonlinear damping force by Zhang et al. [15]. Their results showed that nonlinear damping is preferable as a significant reduction of the peak transmitted force can be achieved without bifurcations in the system. The study by Lang et al [16] showed that the nonlinear viscous damping force could complement the drawbacks of linear viscous damping by modifying the force transmissibility in the resonant region without causing any detrimental effects over the non-resonant regions.

The application of nonlinear stiffness to vibration isolation has a longer history than the use of nonlinear viscous damping. Because of the different pros and cons of these two nonlinear forces, this paper aims to examine the benefits brought by nonlinear viscous damping when it is incorporated into an isolator with nonlinear stiffness. The nonlinear stiffness may be an internal property of the isolator or it may be an inherit property of the system that requires protection from vibration. The isolation system with cubic stiffness is represented by an sdof mass-spring-damper model. The transmissibility performance over a range of frequencies for different levels of linear and nonlinear damping obtained by simulation will be displayed. The comparison of the transmissibility curves will reveal the benefits of nonlinear viscous damping over linear viscous damping. The jump phenomena and the harmonics caused by the spring nonlinearity can be suppressed by either linear or nonlinear damping but only nonlinear damping can maintain the performance over high frequency regions simultaneously.

An sdof mass-spring-damper model with nonlinear spring and damping forces under sinusoidal excitation is described in Section II. The influence of nonlinear damping on this model are given in Section III. Simulation results and discussions are provided in Section IV and finally a conclusion is given in Section V.

## II. SDOF VIBRATION ISOLATORS WITH SPRING AND DAMPING NONLINEARITY

Consider a single-degree-of-freedom (sdof) vibration isolation system under a sinusodial excitation force $f_{in}(t)$ as shown in Figure 1, where

$$f_{in}(t) = A\sin(\bar{\omega}t) \tag{1}$$

with magnitude $A$ and frequency $\bar{\omega}$. The force transmitted to the immobile base, $f_{out}(t)$, is related to $f_{in}(t)$ by the equations of motion of the nonlinear vibration isolation system given by

$$\begin{cases} M\ddot{x}(t) + C_1\dot{x}(t) + C_2[\dot{x}(t)]^3 \\ +K_1x(t) + K_2[x(t)]^3 = f_{in}(t) = A\sin(\bar{\omega}t) \\ \\ f_{out}(t) = C_1\dot{x}(t) + C_2[\dot{x}(t)]^3 + K_1x(t) + K_2[x(t)]^3 \end{cases} \tag{2}$$

where $x(t)$ is the displacement of the moving mass, $M$ is the mass, $C_1$ the viscous damping constant, $C_2$ the cubic viscous damping constant, $K_1$ the linear spring constant, $K_2$ the cubic spring constant. When $K_2, C_2 = 0$, System (2) is identical to a well-studied linear system. When $K_2 \neq 0$ and $C_2 = 0$, it becomes a Duffing-type vibration isolator which exhibits jump-up and jump-down phenomena as discussed in many publications [13], [17], [18].

For the purpose of general analysis, System (2) is reduced to a non-dimensional form which is non-specific to chosen values of $M$ and $K_1$. By denoting the resonant frequency $\omega_0 = \sqrt{K_1/M}$, the first equation in System (2) becomes

$$\ddot{y}(\tau) + \xi_1\dot{y}(\tau) + \xi_2[\dot{y}(\tau)]^3 + y(\tau) + \gamma[y(\tau)]^3 = \sin(\Omega\tau) \tag{3}$$

where

$$\begin{aligned} \tau &= \omega_0 t, && (4)\\ \Omega &= \frac{\bar{\omega}}{\omega_0}, && (5)\\ y(\tau) &= \frac{K_1}{A}x(t), && (6)\\ \gamma &= \frac{A^2K_2}{K_1^3}, && (7)\\ \xi_1 &= \frac{C_1}{\sqrt{K_1M}}, && (8)\\ & && (9) \end{aligned}$$

and

$$\xi_2 = \frac{C_2A^2}{\sqrt{(K_1M)^3}}. \tag{10}$$

System (2) can be viewed as a single-input two-output system by substituting in

$$y_1(\tau) = y(\tau) \tag{11}$$

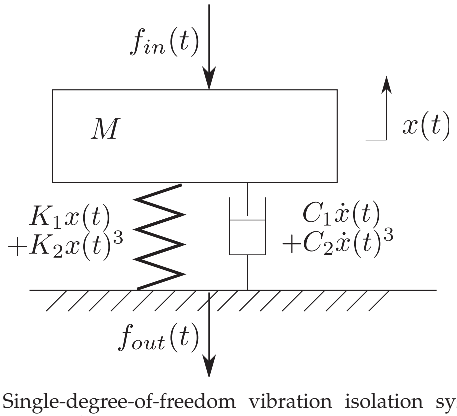


Fig. 1. Single-degree-of-freedom vibration isolation system with nonlinear cubic stiffness and nonlinear cubic viscous damping

and

$$y_2(\tau) = y_1(\tau) + \gamma[y_1(\tau)]^3 + \xi_1\dot{y}_1(\tau) + \xi_2[\dot{y}_1(\tau)]^3 \tag{12}$$

hence the model becomes

$$\begin{cases} \ddot{y}_1(\tau) + y_2(\tau) = u(\tau) = \sin(\Omega\tau) \\ y_2(\tau) = y_1(\tau) + \gamma[y_1(\tau)]^3 + \xi_1\dot{y}_1(\tau) + \xi_2[\dot{y}(\tau)]^3. \end{cases} \tag{13}$$

The force transmissibility $T_1(\Omega)$, the ratio of the magnitude of the transmitted force at the excitation frequency to that of the excitation force, can be deduced from equations (2), (4) - (10) and (13). In the time domain, this gives

$$\begin{aligned} \frac{f_{out}(t)}{A} &= \frac{C_1\dot{x}(t) + C_2[\dot{x}(t)]^3 + K_1x(t) + K_2[x(t)]^3}{A} && (14)\\ &= y_1(\tau) + \gamma[y_1(\tau)]^3 + \xi_1\dot{y}_1(\tau) + \xi_2[\dot{y}_1(\tau)]^3 && (15)\\ &= y_2(\tau). && (16) \end{aligned}$$

By applying Fourier transform to $y_2(\tau)$ and evaluating at $\omega = \Omega$ gives

$$T_1(\Omega) = \Big|Y_2(j\omega)|_{\omega=\Omega}\Big| = \Big|Y_2(j\Omega)\Big|. \tag{17}$$

Equation (17) implies that the force transmissibility of the system can be evaluated by examining the spectrum of the second output of System (13).

Because of the system's nonlinearity, the second output $y_2(\tau)$ contains the fundamental frequency of the input signal as well as harmonics. As System (13) has cubic stiffness and cubic damping, the third harmonics of the output may be significant. This harmonic effect can be described by $T_3(\Omega)$ defined as

$$T_3(\Omega) = \Big|Y_2(j\omega)|_{\omega=3\Omega}\Big|. \tag{18}$$

The numerical simulation results of System (13) will be presented in Section III. The effects of nonlinear viscous damping on the performance of vibration isolation will be shown by $T_1(\Omega)$ and $T_3(\Omega)$.

## III. The effects of nonlinear viscous damping on a Duffing-type vibration isolator

From analysis and observation, the following conclusions can be made.

**Proposition 1**

*(i)*

When $\Omega \approx 1$, there exists a $\bar{\xi_2} > 0$ such that

$$\left.\frac{d[T_1(\Omega)]^2}{d\xi_2}\right|_{\Omega\approx 1} < 0 \tag{19}$$

if $0 < \xi_2 < \bar{\xi_2}$ for $0 \le \gamma < \bar{\gamma}$ where $\bar{\gamma}$ is the maximum value of $\gamma$ beyond which jump phenomena will be observed in System (13).

*(ii)*

When $\Omega \ll 1$ or $\Omega \gg 1$, there exists a $\bar{\xi_2} > 0$ such that

$$T_1(\Omega)|_{\xi_2>0} \approx T_1(\Omega)|_{\xi_2=0} \tag{20}$$

if $0 < \xi_2 < \bar{\xi_2}$ for $0 \le \gamma < \bar{\gamma}$ where $\bar{\gamma}$ is the maximum value of $\gamma$ beyond which jump phenomena will be observed in System (13).

*(iii)*

For fixed values of $\gamma, \xi_1 > 0$, if System (13) exhibits jump phenomena, there exists a $\bar{\xi_2}$ such that the jumps are eliminated.

The complete proof is still under study but a brief description of the idea is provided here. System (13) is a polynomial form nonlinear differential equation model, the input and the two outputs of which can be represented by a Volterra series around the zero equilibrium point in the time domain [19]. The force transmissibility (i.e. the spectrum of the second output) can be related to the input spectrum using the output frequency response function (OFRF) concept [20]. The OFRF defines the relationship between the output spectrum (the force transmissibility) and the parameters which characterise the nonlinearity of a Volterra system ($\gamma$ and $\xi_2$) for a given input and fixed values of the linear terms. This concept is applied to a vibration isolation system with linear stiffness and nonlinear damping in a theoretical study by Lang et al. [16]. The same approach can be applied to System (13) which has an extra nonlinear stiffness term. Its force transmissibility $T_1(\Omega)$ can then be expressed as a polynomial function of $\gamma$ and $\xi_2$. The derivative of $[T_1(\Omega)]^2$ with respect to $\xi_2$ can be obtained for $\Omega \approx 1$ to prove conclusion (i) and the values of $T_1(\Omega)$ for $\xi_2 > 0$ and $\xi_2 = 0$ are compared to reach conclusion (ii).

Conclusion (iii) states that the jump phenomena can be eliminated by nonlinear damping. As the Volterra model cannot capture the jump phenomena, the OFRF concept cannot be applied here. An alternative approach, the harmonic balance method, can be used for this analysis [11], [12], [21]–[27]. The solution of the second output of System (13) is assumed to be of the form of a truncated Fourier series. Substituting this into System (13) and equating coefficients related to each harmonic components yields an expression for $T_1(\Omega)$. This well-known method has been the main tool for the analysis of Duffing systems and can now be extended to the study of a Duffing system with an additional nonlinear damping term.

The three conclusions of Proposition 1 are based on some previous analyses and observations from numerical examples given in Section IV. It is worth pointing out that for System (13), the following remark about the harmonics in the system output can also be made.

**Remark 1**

When $\Omega \approx 1$, there exists a $\bar{\xi_2} > 0$ such that

$$\left.\frac{d[T_3(\Omega)]^2}{d\xi_2}\right|_{\Omega\approx 1} < 0 \tag{21}$$

if $0 < \xi_2 < \bar{\xi_2}$ for $0 \le \gamma < \bar{\gamma}$ where $\bar{\gamma}$ is the maximum value of $\gamma$ beyond which jump phenomena will be observed in System (13).

The present study extends the findings of Lang et al. [16] to a vibration isolation system with nonlinear stiffness. The effects of nonlinear damping are concluded by Proposition 1 and Remark 1. Equation (19), the first conclusion of Proposition 1, suggests that an increase in the value of cubic damping leads to a reduction in the force transmissibility around the resonant frequency range and the second conclusion, summarised by Equation (20), indicates that the the transmissibility over the frequency range far away from the resonant frequency remains unaffected by nonlinear damping. These findings conclude that the ideal properties of nonlinear damping found by Lang et al. [16] also hold true for systems with nonlinear stiffness for certain values of $\gamma$. The jump phenomena is addressed by the third conclusion which shows that the cubic damping, when it reaches a high enough level, can remove the jump phenomena caused by the nonlinear spring thus greatly improving the system stability. Remark 1 focuses on the superharmonics produced by the nonlinear spring. The highest level of superharmonics is produced when the excitation frequency is near the resonant frequency. The derivative of $[T_3(\Omega)]^2$ with respect to $\xi_2$ is negative so Equation (21) indicates that the third harmonic can be reduced by nonlinear damping. These are the main advantages of including nonlinear viscous damping on a Duffing-type system.

## IV. Simulation results and discussion

Numerical simulation studies were conducted using the standard MATLAB solver for ordinary differential equations (`ode45`) to demonstrate the effects of nonlinear viscous damping on a Duffing-type vibration isolation system. The results are presented in Figures 2 to 6.

The advantage of nonlinear viscous damping over conventional linear damping on an sdof vibration isolation system with linear stiffness, as described in the study by Lang et al. [16], are summarised by Figures 2 and 3. In Figure 2, the system has linear stiffness and linear damping with $\xi_1$ = 0.1, 0.2 and 0.4. As the linear damping parameter $\xi_1$ increases, $T_1(\Omega)$ around the resonant region decreases but there is an undesirable increase in $T_1(\Omega)$ in the high frequency region. In Figure 3, the value of $\xi_1$ is kept constant at 0.1 while the cubic damping parameter $\xi_2$ takes the values 0, 0.2 and 0.4. In this case, the resonance at $\Omega \approx 1$ is suppressed by the increase of $\xi_2$ while $T_1(\Omega)$ over the frequency ranges of $\Omega \ll 1$ and $\Omega \gg 1$ is unaffected.

The simulation is repeated for System (13) where the nonlinear stiffness is taken into account (i.e. $\gamma > 0$). Figure 4 shows the effects of linear damping with the presence of nonlinear stiffness. The contrast of the effects of linear and nonlinear viscous damping can be observed by comparing Figure 4 with Figures 5 and 6, where the cubic viscous damping constant $\xi_2$ takes the values 0.2, 0.4 and 0.6 for $\gamma = 0.1$ and $\gamma = 0.2$ respectively. Similar to the case with pure linear stiffness, the cubic viscous damping again modifies the resonant region without causing detrimental effects in the non-resonant regions. It is clear that the increase of $\xi_1$ leads to an increase of $T_1(\Omega)$ over the high frequency range whereas the increase of $\xi_2$ does not. Thus, nonlinear viscous damping has a significant advantage over linear viscous damping on vibration isolation even with the presence of nonlinear stiffness. These observations are summarised by the first two conclusions of Proposition 1.

Nonlinear viscous damping may also address the well-known jump-up and jump-down phenomena of systems with nonlinear stiffness. To ensure stability, jump avoidance should be an important feature of a vibration isolator. The study by Ravindra and Mallik [18] found that the jump phenomena could be eliminated by linear viscous damping. With pure linear viscous damping, the simulation results in Figure 4 shows a jump occurring at $\Omega \approx 1.4$ when $\xi_1 = 0.1$ for $\gamma = 0.2$ and $\xi_2 = 0$ but the jump no long exists when the level of linear damping increases to $\xi_1 = 0.2$. The trade-off of applying linear damping to remove the jump phenomena is the adverse effects on the transmissibility over the high frequency range. This problem can be overcome by employing nonlinear viscous damping as discussed above. In Figures 5 and 6, the jumps disappear when $\xi_2$ increases from 0 to 0.2 while the shape of the force transmissibility curve remains unchanged for $\Omega \ll 1$ and $\Omega \gg 1$. This is the conclusion (iii) of Proposition 1.

Under the excitation of a sinusoidal input force, the output of an sdof system with nonlinear stiffness exhibits some harmonics. To achieve a good level of isolation, these harmonics, which are transmitted to the base, should be minimised. The spectrum of the second output of System (13) contains a strong component at the excitation frequency $\Omega$ plus the higher harmonics at $n\Omega$, where $n = 3, 5, 7, \cdots$. The effects of the fifth and higher harmonics are neglected as their magnitudes are small. The simulation results of $T_3(\Omega)$, the third harmonics, are illustrated in Figures 7 and 8. Consider the case when $\gamma = 0.2$ and $\xi_1 = 0.1$, two peaks, one at $\Omega \approx 0.4$ and another at $\Omega \approx 1.1$, are observed in Figure 7. The magnitudes of both peaks decrease as the value of nonlinear viscous damping increases. This is the conclusion of Remark 1. While $T_1(\Omega)$ measures the force transmissibility at the fundamental frequency, $T_3(\Omega)$ indicates the level of output force at 3Ω. The sum of $T_1(\Omega)$ and $T_3(\Omega)$ (and the higher harmonics terms) measures the overall energy transmissibility. Nonlinear damping can reduce the peaks occurring in both $T_1(\Omega)$ and $T_3(\Omega)$ so it can also remove energy at frequencies $\Omega \approx 1$ from an sdof system with nonlinear stiffness effectively.

The effects of linear damping on the third harmonics $T_3(\Omega)$ are included in Figure 8 for completeness. The results suggest that linear damping reduces $T_3(\Omega)$ over the whole frequency range, as opposed to just the $\Omega \approx 1$ region in Figure 7. This means that linear damping is better at suppressing the higher harmonics than nonlinear damping. However, when the total energy transmitted is considered, an increase in $\xi_2$ leads to a far greater rise in $T_1(\Omega)$ than the fall in $T_3(\Omega)$ for $\Omega \gg 1$. In the nonlinear damping case, $T_3(\Omega)$ rises from 0.003 (-52 dB) to 0.008 (-42 dB) when $\xi_2$ increases from 0.2 to 0.6 at $\Omega = 3$. This increase is very small in absolute terms. Nonlinear damping therefore remains the preferred choice over linear damping on a vibration isolation system with nonlinear stiffness.

## V. Conclusions

A vibration isolation system with nonlinear stiffness and linear viscous damping has been studied by many researchers. The force transmitted contains harmonics of the excitation frequency. The jump-up and jump-down phenomena occurs when the level of linear damping is small. This current study shows that the overall performance of vibration isolation is enhanced by the introduction of nonlinear viscous damping based on an sdof model described in Section II . The simulation results reveal important features of nonlinear viscous damping on a Duffing-type system outlined in Section III. The nonlinear damping parameter can eliminate the jump phenomena and reduce the third harmonics of the transmitted force when the excitation frequency is close to the resonant frequency. These may also be achieved by linear viscous damping but the nonlinear

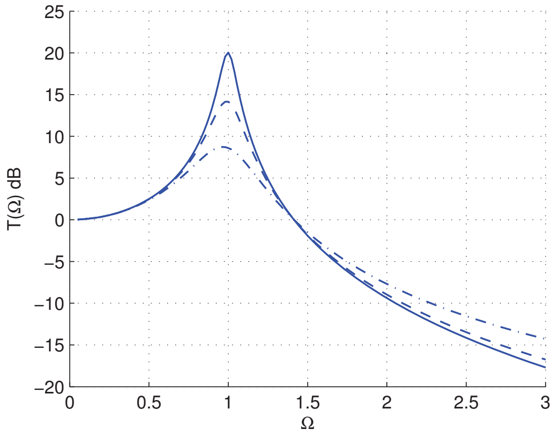


Fig. 2. The force transmissibility of System (2) with linear stiffness and linear viscous damping, where $\gamma, \xi_2 = 0$. Solid: $\xi_1 = 0.1$; Dashed: $\xi_1 = 0.2$; Dot-dashed: $\xi_1 = 0.4$.

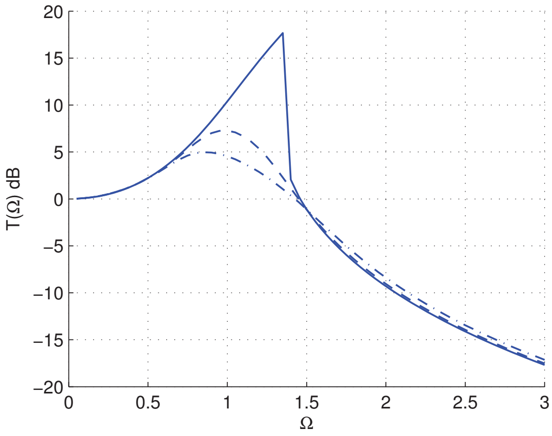


Fig. 5. The force transmissibility of System (2) with nonlinear stiffness and nonlinear viscous damping, where $\gamma, \xi_1 = 0.1$. Solid: $\xi_2 = 0$; Dashed: $\xi_2 = 0.2$; Dot-dashed: $\xi_2 = 0.4$.

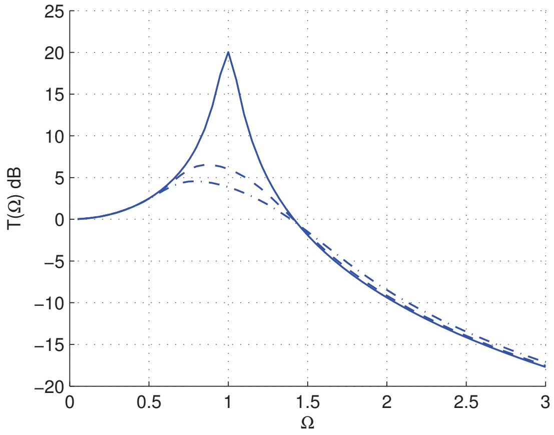


Fig. 3. The force transmissibility of System (2) with linear stiffness and nonlinear viscous damping, where $\gamma = 0$ and $\xi_1 = 0.1$. Solid: $\xi_2 = 0$; Dashed: $\xi_2 = 0.2$; Dot-dashed: $\xi_2 = 0.4$.

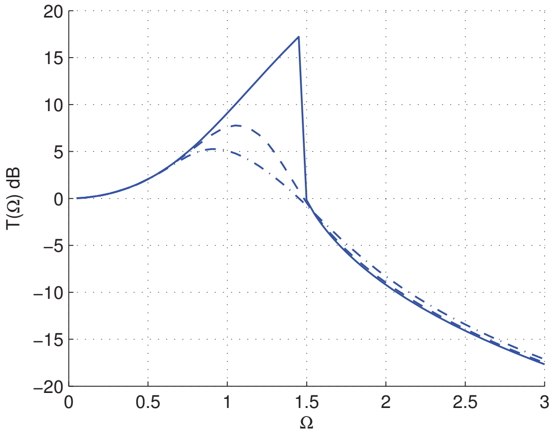


Fig. 6. The force transmissibility of System (2) with nonlinear stiffness and nonlinear viscous damping, where $\gamma = 0.2$ and $\xi_1 = 0.1$. Solid: $\xi_2 = 0$; Dashed: $\xi_2 = 0.2$; Dot-dashed: $\xi_2 = 0.4$.

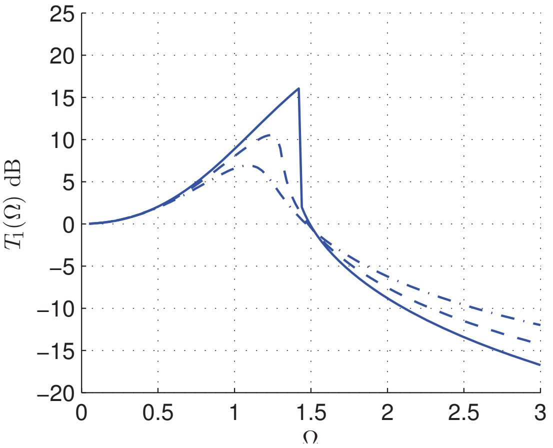


Fig. 4. The force transmissibility of System (2) with nonlinear stiffness and linear viscous damping, where $\gamma = 0.2$, $\xi_2 = 0$. Solid: $\xi_1 = 0.1$; Dashed: $\xi_1 = 0.2$; Dot-dashed: $\xi_1 = 0.4$.

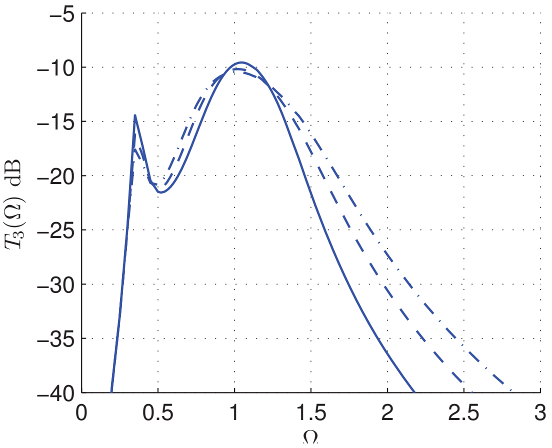


Fig. 7. The third harmonic of the force transmissibility of System (2) with nonlinear stiffness and nonlinear viscous damping, where $\gamma = 0.1$ and $\xi_1 = 0.1$. Solid: $\xi_2 = 0.2$; Dashed: $\xi_2 = 0.4$; Dot-dashed: $\xi_2 = 0.6$.

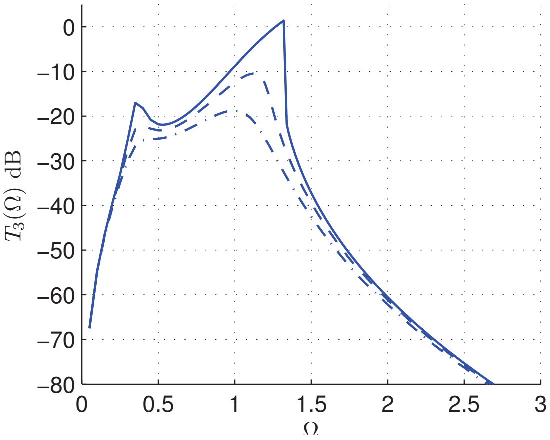


Fig. 8. The third harmonic of the force transmissibility of System (2) with nonlinear stiffness and linear viscous damping, where $\gamma = 0.1$ and $\xi_2 = 0$. Solid: $\xi_1 = 0.2$; Dashed: $\xi_1 = 0.4$; Dot-dashed: $\xi_1 = 0.6$.

damping parameter has a major additional advantage - it lowers the force transmissibility over the frequencies range around the resonant region while keeping the high frequency range unaffected. These features bring significant benefits to passive vibration isolation when nonlinear stiffness is present.

Theoretical analyses including a rigorous proof for Proposition 1 using the OFRF approach and the harmonic balance method will be provided in a later publication. Further studies will focus on the practical applications of this concept on different engineering designs. There are some recent research results on achieving nonlinear viscous damping by controlling the output of Magnetorheological dampers [28]–[30]. The authors will continue to explore this idea and apply the concept of nonlinear viscous damping to areas such as car suspension control and shock absortion for structures.

## Acknowledgment

The authors would like to thank the EPSRC and the ERC for funding this work and the reviewers for their helpful comments.

## References

[1] C. E. Crede, *Vibration and shock isolation*. Wiley, 1951.

[2] J. C. Snowdon, *Vibration and shock in damped mechanical systems*. J. Wiley, 1968.

[3] J. B. Vernon, *Linear vibration theory*. Wiley, 1967.

[4] J. C. Snowdon, “Vibration isolation: use and characterization,” *The Journal of the Acoustical Society of America*, vol. 66, p. 1245, 1979.

[5] E. Suhir, “Shock protection with a nonlinear spring,” *Components, Packaging, and Manufacturing Technology, Part A, IEEE Transactions on*, vol. 18, no. 2, pp. 430–437, 1995.

[6] E. I. Rivin, *Passive vibration isolation*. Amer Society of Mechanical, 2003.

[7] P. Thureau, D. Lecler, and J. Grosjean, *An introduction to the principles of vibrations of linear systems*. Thornes, 1981.

[8] D. J. Mead, *Passive vibration control*. John Wiley & Sons Inc, 1999.

[9] M. J. Crocker, *Handbook of noise and vibration control*. Wiley, 2007.

[10] R. A. Ibrahim, “Recent advances in nonlinear passive vibration isolators,” *Journal of Sound and Vibration*, vol. 314, no. 3-5, pp. 371–452, 2008.

[11] L. Liu, J. P. Thomas, E. H. Dowell, P. Attar, and K. C. Hall, “A comparison of classical and high dimensional harmonic balance approaches for a duffing oscillator,” *Journal of Computational Physics*, vol. 215, no. 1, pp. 298–320, 2006.

[12] H. Hu and J. H. Tang, “Solution of a duffing-harmonic oscillator by the method of harmonic balance,” *Journal of sound and vibration*, vol. 294, no. 3, pp. 637–639, 2006.

[13] M. J. Brennan, I. Kovacic, A. Carrella, and T. P. Waters, “On the jump-up and jump-down frequencies of the duffing oscillator,” *Journal of Sound and Vibration*, vol. 318, no. 4-5, pp. 1250–1261, 2008.

[14] A. Carrella, M. J. Brennan, T. P. Waters, and V. Lopes, “Force and displacement transmissibility of a nonlinear isolator with high-static-low-dynamic-stiffness,” *International Journal of Mechanical Sciences*, 2011.

[15] B. Zhang, S. A. Billings, Z. Q. Lang, and G. R. Tomlinson, “Suppressing resonant vibrations using nonlinear springs and dampers,” *Journal of Vibration and Control*, vol. 15, no. 11, p. 1731, 2009.

[16] Z. Lang, X. Jing, S. Billings, G. Tomlinson, and Z. Peng, “Theoretical study of the effects of nonlinear viscous damping on vibration isolation of sdof systems,” *Journal Of Sound And Vibration*, vol. 323, no. 1-2, pp. 352–365, 2009.

[17] T. Fang and E. H. Dowell, “Numerical simulations of jump phenomena in stable duffing systems,” *International journal of non-linear mechanics*, vol. 22, no. 3, pp. 267–274, 1987.

[18] B. Ravindra and A. K. Mallik, “Hard duffing-type vibration isolator with combined coulomb and viscous damping,” *International journal of non-linear mechanics*, vol. 28, no. 4, pp. 427–440, 1993.

[19] S. Boyd and L. Chua, “Fading memory and the problem of approximating nonlinear operators with volterra series,” *Circuits and Systems, IEEE Transactions on*, vol. 32, no. 11, pp. 1150–1161, 1985.

[20] Z. Q. Lang, S. A. Billings, R. Yue, and J. Li, “Output frequency response function of nonlinear volterra systems,” *Automatica*, vol. 43, no. 5, pp. 805–816, 2007.

[21] G. Schmidt and A. Tondl, *Non-linear vibrations*. Cambridge Univ Pr, 1986, vol. 66.

[22] M. N. Hamdan and T. D. Burton, “On the steady state response and stability of non-linear oscillators using harmonic balance,” *Journal of sound and vibration*, vol. 166, no. 2, pp. 255–266, 1993.

[23] N. MacDonald, “Choices in the harmonic balance technique,” *Journal of Physics A: Mathematical and General*, vol. 26, p. 6367, 1993.

[24] H. P. W. Gottlieb, “Harmonic balance approach to limit cycles for nonlinear jerk equations,” *Journal of Sound and Vibration*, vol. 297, no. 1, pp. 243–250, 2006.

[25] Z. K. Peng and Z. Q. Lang, “The effects of nonlinearity on the output frequency response of a passive engine mount,” *Journal of Sound and Vibration*, vol. 318, no. 1-2, pp. 313–328, 2008.

[26] R. E. Mickens, *Truly nonlinear oscillations: harmonic balance, parameter expansions, iteration, and averaging methods*. World Scientific Pub Co Inc, 2010.

[27] Z. K. Peng, G. Meng, Z. Q. Lang, W. M. Zhang, and F. L. Chu, “Study of the effects of cubic nonlinear damping on vibration isolations using harmonic balance method,” *International Journal of Non-Linear Mechanics*, 2011.

[28] N. D. Sims, R. Stanway, D. J. Peel, W. A. Bullough, and A. R. Johnson, “Controllable viscous damping: an experimental study of an electrorheological long-stroke damper under proportional feedback control,” *Smart materials and structures*, vol. 8, p. 601, 1999.

[29] N. D. Sims, R. Stanway, A. R. Johnson, D. J. Peel, and W. A. Bullough, “Smart fluid damping: shaping the force/velocity response through feedback control,” *Journal of intelligent material systems and structures*, vol. 11, no. 12, pp. 945–958, 2000.

[30] H. Laalej and Z. Q. Lang, “Numerical investigation of the effects of mr damper characteristic parameters on vibration isolation of sdof systems under harmonic excitations,” *Journal of Intelligent Material Systems and Structures*, vol. 21, no. 5, p. 483, 2010.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Stabilisation of Multi-Input Nonlinear Systems Using Associated Angular Approach

Zahra Sangelaji
Department of Engineering Systems
University of Greenwich
UK
Email: z.sangelaji@greenwich.ac.uk

***Abstract*—In this paper the stabilisation of multi-input nonlinear systems is studied using the associated angular approach. In this method, a nonlinear system is converted into two associated subsystems, the so-called radial and spherical subsystems. For a single input nonlinear system, the control is designed using the one dimensional radial system to stabilise the radial and consequently the original nonlinear system. For multi-input systems, the control is also designed based on the radial subsystem, however, the method is not straightforward in comparison with single input systems. The control law includes a weighting function which is determined based on the system performance and control action. Some examples are presented to illustrate the effect of various scenarios of using the proposed method.**

## I. Introduction

In recent years, various methods have been developed to design a control for many classes of nonlinear systems including linearisation [1], optimal control [2], [3], $H_\infty$ control [4], [5], sliding mode control (SMC) [6] using quantised feedback [7] and adaptive control [8]-[9]. Output feedback control design is also a method for stabilisation of a broad class of nonlinear systems which has been studied in the last two decades. This method is mainly used when the system output is measurable and some states are not available or they are very difficult to measure [1]. Full state-feedback control design methods are utilized to stabilise a nonlinear system globally, particularly, when the states are measurable. These methods include sliding mode control [10], backstepping [11], zero dynamics based on high gain [12] and neural network [13]. However many established methods only guarantee the local stability [1] or ultimate boundedness of the states [14]. Sangelaji and Banks [15]-[16] have proposed associated angular approach for the global stabilisation of a general class of single-input nonlinear systems by using the angular form. In this method, the system is converted into two nonlinear subsystems. The trajectories of a subsystem which move on a sphere is termed the spherical subsystem and the other, a scalar nonlinear system is called the radial subsystem. The method straightforwardly yields a controller when there is no singularity except the origin in the input map function. For single-input systems, the control law is generally simple for many cases and the method is applicable to a large class of nonlinear systems. Whenever the input map of the radial subsystem is zero, the radial control is not definable. In this case, some mild conditions are proposed to guarantee the system stability [17] or the radial control or control design method should be modified such that the designed control is definable everywhere within the operating region and also stabilises the system. The radial control can be continuous or discontinuous depending on the structure of the input map. The method was originally established for single input nonlinear systems [15], [16]. In this paper the method is extended to multi-input nonlinear systems which are not straightforward, because the control can not be driven using the associated angular subsystems as proposed for single input systems. The proposed control stablises the multi-input nonlinear system. The presented method can be applied to any nonlinear system while most of the existing methods are applicable to particular classes of nonlinear systems. The drawback of the angular method is the singularity points in which the control is not definable. Similar methods as established for single input nonlinear systems are required to remove the singularities.

This paper is organised as follows: In Section II the associated angular method is studied. In Section III the multi-input system is presented. In Section IV, the special cases of the design weighting matrix, are considered. Examples illustrating the control design process are presented in Section V. Finally conclusions are presented in Section VI.

## II. Associated angular method for single input nonlinear systems

Consider the nonlinear system:

$$\dot{x} = A(x) + B(x)u \tag{1}$$

where $x \in \mathbb{R}^n$ is the state, $u$ is the scalar control, $A(x) \in \mathbb{R}^n$ and $B(x) \in \mathbb{R}^n$.

Let $\mathbb{S}^n \subseteq \mathbb{R}^n$ be the unit $n$-ball, i.e. $S^n = \{z \in \mathbb{R}^n : \|z\| = 1\}$ and $\mathbb{R}^+$ be the set of positive real numbers. The map

$$\varphi : \mathbb{R}^n - \{0\} \to \mathbb{R}^+ \times \mathbb{S}^n$$
$$x \to \big(\|x\|, \frac{x}{\|x\|}\big)$$

is a diffeomorphism from $\mathbb{R}^n - \{0\}$ onto $\mathbb{R}^+ \times \mathbb{S}^n$. Note that even as $x$ tends to zero, $\dfrac{x}{\|x\|}$ $(= z)$ is on the unit ball. The origin is removed from the domain of the function $\varphi$; otherwise the origin corresponds to infinity pair $(0, z)$ where $z$ is any point in $\mathbb{S}^n$. This obstacle can be removed if a unique pair say $(0, z_0)$ with $z_0 = (1, \ldots, 0)$ corresponds to the origin. Using diffeomorphism $\varphi$ the system (1) is converted into the associated radial and spherical subsystems as presented in the following Lemma.

**Lemma 1.** *The system (1) can be written in the form*

$$\dot{r} = \lambda_A + \lambda_B u \tag{2}$$

$$\dot{z} = \frac{1}{r}\big(\bar{A}(r,z) + \bar{B}(r,z)u\big) \tag{3}$$

*from which the following control is obtained*

$$u = -\frac{\lambda_A + \alpha r}{\lambda_B} \tag{4}$$

*where*

$$\bar{A}(r,z) = A(r,z) - z^T A(r,z) z$$

$$\bar{B}(r,z) = B(r,z) - z^T B(r,z) z$$

*Also* $\lambda_A = z^T A(r,z)$, $\lambda_B = z^T B(r,z)$, $r = \|x\|$, $z = \dfrac{x}{\|x\|}$ *and* $\alpha > 0$ *is a constant real number. Moreover, the control (4) stabilises the system (1).*

*Proof:* Since $r = \|x\|$ and $r^2 = x^T x$,

$$2r\dot{r} = 2x^T \dot{x} \tag{5}$$

Therefore

$$r\dot{r} = x^T\big(A(x) + B(x)u\big) \tag{6}$$

and

$$\dot{r} = \frac{x^T(A(x) + B(x)u)}{r} \tag{7}$$

Substituting $x = rz$ into (7) yields

$$\begin{aligned} \dot{r} &= z^T(A(r,z) + B(r,z)u) \\ &= \lambda_A + \lambda_B u \end{aligned} \tag{8}$$

On the other hand, using $z = \dfrac{x}{r}$ and (6) one can obtain

$$\begin{aligned} \dot{z} &= \frac{1}{r}\dot{x} - \frac{\dot{r}}{r^2}x \\ &= \frac{1}{r}\Big(A(r,z) + B(r,z)u\big) - \frac{1}{r^3}\big(x^T A(r,z) + x^T B(r,z)u\Big)x \\ &= \frac{1}{r}\big(\bar{A}(r,z) + \bar{B}(r,z)u\big) \end{aligned} \tag{9}$$

with

$$\bar{A}(r,z) = A(r,z) - z^T A(r,z) z$$

$$\bar{B}(r,z) = B(r,z) - z^T B(r,z) z$$

Select the control

$$u = -\frac{\lambda_A + \alpha r}{\lambda_B} \tag{10}$$

where $\alpha > 0$ is a real number. Then from (6)

$$r\dot{r} = -\alpha r^2 \tag{11}$$

So $\dot{r} = -\alpha r$ which guarantees the stability of the subsystem (2) and therefore, the system (1).

Note that the $z$-subsystem operates on the unit ball and $r$-subsystem is scalar. The real positive number $\alpha$ is a design parameter and only affects the degree of the stability of the system. In other words, for large values of $\alpha$ the state settling time is shorter in comparison with small values of $\alpha$. One way to ensure the accessibility of the control (10) is to consider some specific constrains on $\alpha$.

## III. Multi-input systems

The control design and stabilisation problem using the angular approach, which has been studied in section II is only applicable to single-input nonlinear systems and its extension to a general class of nonlinear system is not straightforward. The degree of nonlinearity in the system is not an important issue for using this method, while the most existing methods are applicable for specific nonlinear classes of nonlinear system in which the structure and the nature of nonlinearities affect the process of the control design. In this section, the angular method is extended to design an appropriate control for multi-input nonlinear systems and the stabilisation criteria are also presented.

Consider the multi-input nonlinear affine system

$$\dot{x} = A(x) + B(x)u \tag{12}$$

where $A(x) \in \mathbb{R}^{n\times n}$, $B(x) \in \mathbb{R}^{n\times m}$, $u \in \mathbb{R}^m$ and $x \in \mathbb{R}^n$. Let $r = \|x\|$ and $z = \frac{x}{\|x\|}$ then

$$\begin{aligned} r\dot{r} &= x^T \dot{x} \\ &= x^T A(x) + x^T B(x) u \end{aligned}$$

Therefore

$$\dot{r} = \frac{1}{r}(\lambda_A(x) + (\lambda_B(x)u) \tag{13}$$

$$\dot{z} = \frac{1}{r}\Big[\big(A - z^T A z\big) + \big(B - z^T B z\big)u\Big] \tag{14}$$

where $\lambda_A = x^T A(x)$, $\lambda_B = x^T B(x) \in \mathbb{R}^{1\times m}$ and $u \in \mathbb{R}^{m\times 1}$. Suppose that there is an $\alpha > 0$ such that $\dot{r} = -\alpha r$. The condition $\dot{r} = -\alpha r$ is a sufficient condition for stability of the system. Therefore the equation (13) implies

$$\lambda_B u = -\lambda_A(x) - \alpha r^2 \tag{15}$$

Since the $r$-subsystem is a one-dimensional system, the vector control input $u$ should be selected such that (15) is satisfied. Select the control

$$u = -\frac{(x^T A(x) + \alpha r^2) R B^T(x) x}{x^T B(x) R B^T(x) x} \tag{16}$$

where the weighting matrix $R$ is nonsingular. Substituting control (16) in (13) yields

$$\begin{aligned}\dot{r} &= \frac{1}{r}(x^TA(x) + x^TB(x)u)\\ &= \frac{1}{r}\left[x^TA(x) + x^TB(x)\frac{(-x^TA(x) - \alpha r^2)RB^T(x)x}{x^TB(x)RB^T(x)x}\right]\\ &= \frac{1}{r}(x^TA(x) + \frac{(-x^TA(x) - \alpha r^2)x^TB(x)RB^T(x)x}{x^TB(x)RB^T(x)x})\\ &= \frac{1}{r}(x^TA(x) - x^TA(x) - \alpha r^2)\\ &= -\alpha r\end{aligned}$$

Therefore, if $x^TBRB^Tx \neq 0$ the control (16) stabilises the system (12). When $x^TBRB^Tx = 0$ the control (16) should be modified. The methods in [16] for removing the singularities, i.e. for points belong to the set

$$\Pi = \{x \in \mathbb{R}^n : x^TBRB^Tx = 0\}$$

may straightforwardly be extended to the nonlinear system (12).

**Remark 1.** *When $m = 1$, the system (12) is a single input system. In this case, $B^T(x)x \in \mathbb{R}$ and $R \in \mathbb{R}$. If $B^T(x)x \neq 0$ the control (16) coincides with the control 10*

$$u = -\frac{x^TA(x) + \alpha r^2}{x^TB(x)}$$

## IV. Selection of the weighting matrix $R$

The control (16) depends on the weighting matrix $R$ and can be selected based on desired system performance and control action. Various selection of the weighting matrix $R$ in (16) yields various alternative controls. However, for any selection of $R$, the control (16) stabilises the system. Consider the following cases:

(*i*) Let $R = \beta I_m \in \mathbb{R}^{m\times m}$ where $\beta > 0$. Then

$$u = -\frac{(\lambda_A(x) + \alpha r^2)B^T(x)x}{x^TB(x)B^T(x)x} \tag{17}$$

or

$$u = -\frac{x^T(A(x) + \alpha x)B^T(x)x}{x^TB(x)B^T(x)x}$$

Therefore, for any $\beta > 0$ the selection of $R = \beta I$ does not yield different control law. In fact, any $\beta$ result in the same control as is given by (17).

(*ii*) Assume that $B$ is full rank. Select $R = (B^TB)^{-1}$. The control (16) is now in the following form

$$u = -\frac{(\lambda_A(x) + \alpha r^2)(B^TB)^{-1}B^Tx}{x^TB(B^TB)^{-1}B^Tx} \tag{18}$$

or

$$u = -\frac{x^T(A + \alpha x)(B^TB)^{-1}B^Tx}{x^TB(B^TB)^{-1}B^Tx}$$

Note that usually $R$ is considered a symmetric positive-definite matrix. However, the weighting matrix $R$ may be considered only as a nonsingular matrix. Control (18) when $R$ is nonsingular matrix guarantees the stability of the multi-input system (12).

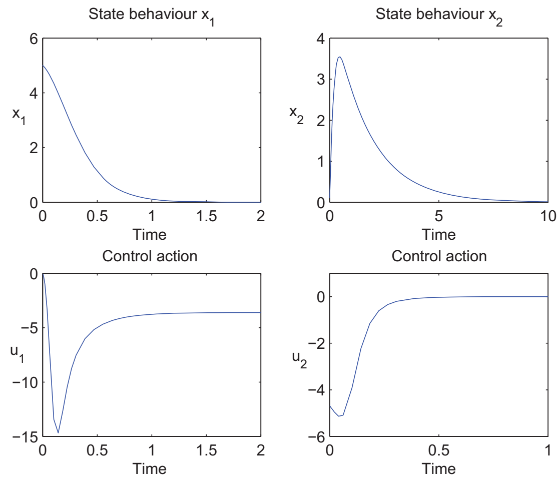


Fig. 1. Responses of the multi-input system using the radial control (16)

## V. Examples

In this section two examples are presented to show the various scenarios of design of an angular controller. In the first example, the control is straightforwardly obtained as there is no singularities (except the origin), if the parameters are appropriately selected. The second example indicates the case when there are singularities. In this case a suitable condition is required.

### A. *Example 1*

Consider the system

$$\begin{aligned}\dot{x}_1 &= (-1 + x_1)x_1 + x_1u_2\\ \dot{x}_2 &= 4x_1 + 3x_2 + x_2u_1\end{aligned}$$

The state space representation of the nonlinear system is

$$\dot{x} = \begin{pmatrix} -1 + x_1 & 0 \\ 4 & 3 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 0 & x_1 \\ x_2 & 0 \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \tag{19}$$

For this system, $x^TAx$ and $B^Tx$ are given by

$$\begin{aligned}x^TAx &= x_1^2(-1 + x_1) + 4x_1x_2 + 3x_2^2\\ B^Tx &= \begin{pmatrix} x_2^2 \\ x_1^2 \end{pmatrix}\end{aligned}$$

The control (17) with $R = I_2$ is

$$\begin{aligned}u &= \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}\\ &= -\frac{x_1^2(-1 + x_1) + 4x_1x_2 + 3x_2^2 + \alpha(x_1^2 + x_2^2)}{x_2^4 + x_1^4}\begin{bmatrix} x_2^2 \\ x_1^2 \end{bmatrix}\end{aligned}$$

The simulation results are shown in Figure 2 for $\alpha = 0.3$.

Now consider

$$R = \begin{pmatrix} \gamma & 0 \\ 0 & \beta \end{pmatrix}$$

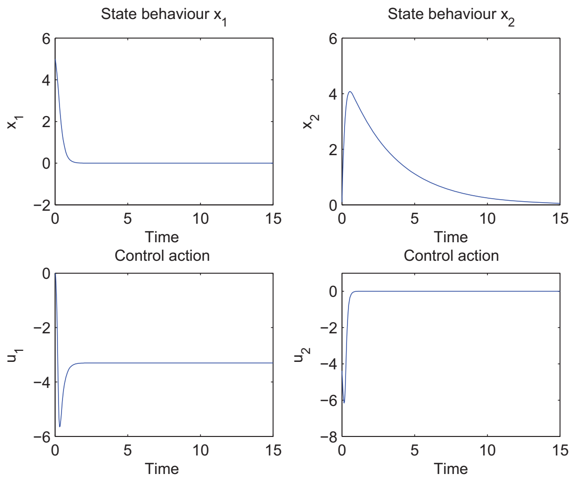


Fig. 2. Responses of the multi-input system using the radial control (17).

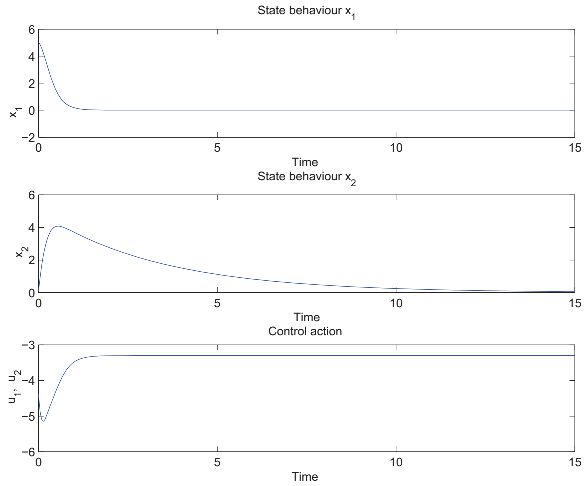


Fig. 3. Responses of the multi-input system using the radial control (18).

where $\gamma \neq \beta$ are positive numbers. Then

$$RB^T x = \begin{pmatrix} \gamma x_2^2 \\ \beta x_1^2 \end{pmatrix}$$

The control (16) is now in the following form

$$\begin{aligned} u &= \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \\ &= -\frac{x_1^2(-1+x_1)+4x_1x_2+3x_2^2+\alpha(x_1^2+x_2^2)}{\gamma x_2^4+\beta x_1^4}\begin{pmatrix} \gamma x_2^2 \\ \beta x_1^2 \end{pmatrix} \end{aligned}$$

If the weighting matrix $R$ is selected such that $\alpha\beta < 0$. Then the control (20) is not defined for all $x$ satisfying $\gamma x_2^4 + \beta x_1^4 = 0$ and therefore the number of singular points are infinite, whilst if $\alpha\beta > 0$ the control (20) is definable for all $x \in \mathbb{R}^n - \{0\}$. Thus, the selection of $R$ is significantly important for designing multi-input nonlinear systems. This example shows that a suitable selection of R is a way for removing the singularities. The simulation results are depicted in Figure 1. The values for $\alpha = 0.3$, $\gamma = 6$, $\beta = 0.3$ and initial conditions $x_0 = [5, 0.1]$ are considered for simulation.

In this example, $B$ is full rank for all $x_1 \neq 0$ and $x_2 \neq 0$. Therefore, $R$ can be selected as $R = (B^TB)^{-1}$. Therefore,

$$B^TB = \begin{pmatrix} x_2^2 & 0 \\ 0 & x_1^2 \end{pmatrix}, \quad (B^TB)^{-1} = \begin{pmatrix} \frac{1}{x_2^2} & 0 \\ 0 & \frac{1}{x_1^2} \end{pmatrix}$$

and

$$x^TAx = x_1^2(-1+x_1)+4x_1x_2+3x_2^2$$
$$x^TB(B^TB)^{-1}B^Tx = x_2^2 + x_1^2$$

Since for $(x_1, x_2) \neq (0,0)$, $x^TB(B^TB)^{-1}B^Tx \neq 0$, the control is defined for all $(x_1, x_2) \neq 0$ and in this case both control laws $u_1$ and $u_2$ are the same

$$u_1 = u_2 = -\frac{x_1^2(-1+x_1)+4x_1x_2+3x_2^2+\alpha(x_1^2+x_2^2)}{x_1^2+x_2^2}$$

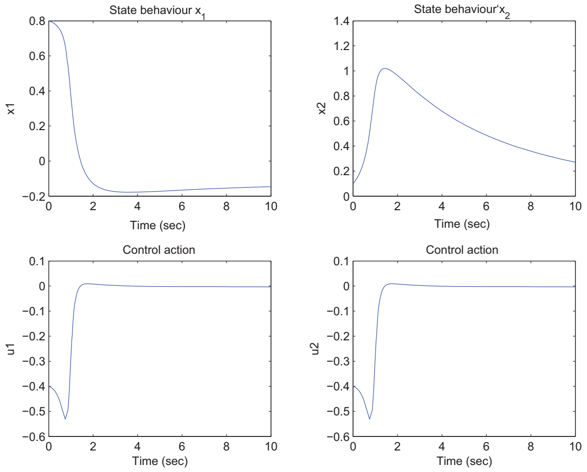


Fig. 4. Responses of the multi-input system using the radial control (20) and $\alpha = 0.4$.

Figure 3 shows the simulation results for $\alpha = 0.3$. This example illustrates that different selections of $R$ yield various controls. When $R = (B^TB)^{-1}$ the two control inputs $u_1$ and $u_2$ are the same, whilst other choice of $R$ presents a control vector with different components. Therefore, based upon the desired system performance and response, the weighting function may be selected.

### B. Example 2

Consider the system

$$\dot{x}_1 = x_2 + x_1u_1 - x_2u_2$$
$$\dot{x}_2 = x_1x_2 - x_2u_1 + x_1u_2$$

The system can be written as

$$\dot{x} = \begin{pmatrix} 0 & 1 \\ x_2 & 0 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} x_1 & -x_2 \\ -x_2 & x_1 \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$$

For this example

$$B^T x = \begin{pmatrix} x_1^2 - x_2^2 \\ 0 \end{pmatrix}$$

$$B^T B = \begin{pmatrix} x_1^2 + x_2^2 & -2x_1x_2 \\ -2x_1x_2 & x_1^2 + x_2^2 \end{pmatrix}$$

and

$$(B^T B)^{-1} = \frac{1}{(x_1^2 - x_2^2)^2} \begin{pmatrix} x_1^2 + x_2^2 & 2x_1x_2 \\ 2x_1x_2 & x_1^2 + x_2^2 \end{pmatrix}$$

Therefore,

$$x^T A x = x_1 x_2^2 + x_1 x_2$$

$$(B^T B)^{-1} B^T x = \frac{1}{x_1^2 - x_2^2} \begin{pmatrix} x_1^2 + x_2^2 \\ 2x_1x_2 \end{pmatrix}$$

$$x^T B (B^T B)^{-1} B^T x = x_1^2 + x_2^2$$

Assume that for all nonzero $x_1$ and $x_2$, $x_1 \neq x_2$. Then the control is

$$u = -\frac{x_1x_2^2 + x_1x_2 + \alpha(x_1^2 + x_2^2)}{x_1^2 - x_2^2} \begin{pmatrix} 1 \\ \dfrac{2x_1x_2}{x_1^2 + x_2^2} \end{pmatrix} \tag{20}$$

The simulation results are shown in Figure 4.

Note that all angular methods proposed for single input systems can straightforwardly be extended to multi-input systems.

## VI. Conclusions

In this paper the control design using the angular approach for multi-input nonlinear systems have been addressed. The radial control law includes a nonsingular weighting matrix which yield various control laws whenever it is not selected as a multiplication of an identity matrix. The weighting matrix does not necessarily have to be a positive-definite, however it only needs to be a nonsingular matrix. A suitable selection of weighting matrix can prevent any singularities in the radial control law. The control (16) has been designed such that when the system is single-input, this control coincides with the control (10). In single-input case, the weighing matrix is not required. In fact, in this case the weighting matrix is only a number.

All methods for single input systems for removing the singularities which have been presented in [15] and [16] are required to be extended to the multi-input nonlinear systems. In particular, similar methods as that presented in [16] should be established for removing the singularities including modifying the control, imposing a sufficient condition on design parameters, using the weighting norm and dynamical radial method.

## References

[1] Isidori, A., Nonlinear control systems II, Springer, London, 1999.

[2] Banks, S. P., On the optimal control of nonlinear systems, Systems and Control Letters, 1986, 6, pp. 337–343

[3] Naidu, D. S., Optimal Control Systems, CRC Press, 2002.

[4] Knobloch, H. W., Isidori, A., and Flokerzi, D., Topics in Control Theory, DMV seminar band 22, Birkhauser, Berlin, 1993.

[5] Green M. and Limebeer, D. J. N., *Linear Robust Control*, Prentice Hall, New Jersey 1995.

[6] Lukyanov, A. G., and Utkin, V. I., Method of reducing equations for dynamic systems to a regular form, Automation and Remote Control, 42, 1981 pp. 413–420.

[7] Liu, J., and Elia, N., Quantized feedback stabilization of non-linear affine systems, Int. J. Control, 77, pp. 239ï£¡249.

[8] Ge,S.S., Wang, Z. and Lee, T.H., Adaptive stabilization of uncertain nonholonomic systems by state and output feedback, Automatica 39, 2003, pp. 1451-1460.

[9] Astolfi, A. and Ortega, R., Invariant manifolds, asymptotic immersion and the (adaptive) stabilization of nonlinear systems. In: Zinober, A. and Owens, D., Editors, 2002. *Nonlinear and adaptive control*, Springer, Berlin.

[10] Utkin, V. I., *Sliding Modes in Control and Optimisation*, Springer-Verlag, Berlin, 1992.

[11] Freeman, R., and Kokotovicï£¡, P., Robust Nonlinear Control Design: State-Space and Lyapunov Techniques. Boston, MA: Birkhï£¡user, 1996.

[12] Byrnes, C. I., and Isidori, I., Asymptotic stabilization of minimum-phase nonlinear systems, IEEE Trans. Automat. Control, vol. AC-36, No. 10, pp. 1122-1137, 1991.

[13] Meyer-Baese, A., Koshkouei, A. J., Emmett, M. R. and Goodall D., Global stability analysis and robust design of multi-time-scale biological networks under parametric uncertainties. Neural Networks 22, pp 658-663, 2009.

[14] Song, B., Hedrick, J. K., and Howell, a., robust stabilization and ultimate boundedness of dynamic surface control systems via convex optimization, int. J. Control, , Vol. 75, No. 12, 870 -881, 2002.

[15] Sangelaji, Z. and Banks, S. P., Radial Control Design for a Class of Nonlinear Systems. Journal of Control and Intelligent Systems, 37(3), 144-151, 2009.

[16] Sangelaji, Z. and Banks, S. P., Stabilisation of Nonlinear Systems Using Weighted Angular Method, IET Control Theory and Applications, 3(4), 445-451, 2009.

[17] Sangelaji, Z. and Banks, S. P., Control design and stabilization analysis of non-linear systems using angular representations. Proc. IMechE, Part I: J. Systems and Control Engineering, 222(I6), 457-463, 2008.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Optimal Output Regulation of Minimum Phase Nonlinear Systems

Attaullah Y. Memon[†]

*Abstract*— **This paper studies the design of an optimal stabilizing controller for output regulation of minimum phase nonlinear systems in the Lyapunov redesign framework of our earlier work [6], and investigates the asymptotic robustness properties of the overall feedback design, given that the optimal stabilizing controller itself possesses strong robustness properties by construction. The motivation comes from the flexibility of incorporating any stabilizing controller within the proposed framework, and we seek for control design methods that yield stabilizing controllers with some additional desirable properties like optimality, disturbance rejection and robustness in the presence of matched uncertainties e.g. static nonlinearities, uncertain parameters and the unmodeled fast dynamics. We exploit the optimal control design methods developed by Kokotovic and his co-researchers [3], [4], [7] for nonlinear systems, where it is shown that in addition to achieving the asymptotic stability of the system and minimizing a cost functional, the optimal feedback control guarantees stability margins which characterize the robustness properties.**

*Index Terms*— **Nonlinear Systems, Inverse Optimal Control, Lyapunov Redesign, Output Regulation**

## I. INTRODUCTION

In this paper, the problem of optimal output regulation of minimum-phase nonlinear systems is considered. The output regulation problem deals with the design of a controller to make the output of a plant asymptotically track reference signals and reject disturbance signals, both produced by an autonomous external system called the *exosystem*. In our earlier work [6], we used the Lyapunov redesign and saturated high-gain feedback approach to design the stabilizing compensator, and included a conditional servocompensator by modifying the original controller that yields asymptotic error regulation without degrading the transient performance. One special feature of the Lyapunov redesign framework of [6] is that it allows us to start with *any* stabilizing controller and then include a conditional servocompensator by modifying the original controller to achieve the desired control objectives. This flexibility of incorporating any stabilizing controller within our framework motivates us to seek for control design methods that yield stabilizing controllers with some additional desirable properties like optimality, disturbance rejection and robustness in the presence of matched uncertainties e.g. static nonlinearities, uncertain parameters and the unmodeled fast dynamics. Herein, we take into consideration the optimal control design methods developed by Kokotovic and his co-researchers [3], [4], [7] for the stabilization of nonlinear systems, where it is shown that in addition to achieving the asymptotic stability of the system and minimizing a cost functional, the optimal feedback control guarantees stability margins which characterize the robustness properties. A major handicap in designing such controller is that it requires the solution of the complicated Hamilton-Jacobi-Bellman (HJB) partial differential equations. Kokotovic and co-researchers introduced an inverse approach [7] to the optimal control design for nonlinear systems, which abrogates the requirement of solving the HJB equations in order to design optimal feedback controllers. We incorporate an optimal stabilizing controller in the Lyapunov redesign framework of [6], to investigate the problem of output regulation of nonlinear systems using conditional servocompensators. We concentrate on the asymptotic robustness properties of the overall Lyapunov-redesign + conditional servocompensator framework, given that the optimal stabilizing controller itself possesses strong robustness properties (e.g. robustness to matched uncertainties and unknown disturbances) by design.

The rest of the paper is organized as follows. We present a brief review of Lyapunov redesign framework of [6] in the next section. Section III introduces the definitions of stability margins for nonlinear systems which are due to [7], and can be considered as the starting point of control design in this paper. Section IV states the problem formulation and assumptions, and is followed by the closed-loop analysis in Section V. A simple example is worked out in Section VI. Finally, Section VII draws the conclusions.

## II. OUTPUT REGULATION USING CONDITIONAL SERVOCOMPENSATORS

In this section we briefly review the Lyapunov redesign approach to output regulation problem using conditional servocompensators [6]. Consider the SISO nonlinear system

$$\begin{aligned} \dot{\xi} &= \tilde{f}(\xi, w) + \tilde{g}(\xi, w)u \\ e &= \tilde{h}(\xi, w) \end{aligned} \tag{1}$$

where $\xi \in R^n$ is the state, $u$ is the control input, $e$ is the regulation error and the functions $\tilde{f}$, $\tilde{g}$ and $\tilde{h}$ are sufficiently smooth. The plant is subjected to a vector of *exogenous* input variables, which are generated by the known exosystem

$$\dot{w} = S_0 w \tag{2}$$

where $S_0$ has distinct eigenvalues on the imaginary axis and $w(t)$ belongs to a compact set $\mathcal{W}$. Suppose that for all $w \in \mathcal{W}$, there exist a continuously differentiable mapping $\xi = \pi(w)$, with $\pi(0) = 0$, and a continuous mapping $\chi(w)$, generated by the internal model

$$\frac{\partial \tau(w)}{\partial w} S_0 w = S\tau(w), \qquad \chi(w) = \Gamma\tau(w)$$

[†] The author is with the Faculty of Electronics & Power Engineering, PN Engineering College, National University of Sciences & Technology, Karachi, Pakistan. (E-mail: attaullah@pnec.edu.pk)

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

where $S$ has distinct eigenvalues on the imaginary axis, such that

$$\begin{aligned} \frac{\partial \pi(w)}{\partial w} S_0 w &= \tilde{f}(\pi, w) + \tilde{g}(\pi, w)\chi(w) \\ 0 &= h(\pi, w) \end{aligned} \tag{3}$$

With the change of variables $x = \xi - \pi$, the system (1) can be represented by

$$\dot{x} = f(x, w) + g(x, w)[u - \chi(w)] \tag{4}$$

The system (4) is in the form where the state feedback regulation problem can be formulated as a state feedback stabilization problem by treating $\chi(w)$ as a matched uncertainty. Suppose there is a locally Lipschitz function $\psi(x, w)$, with $\psi(0, w) = 0$, and a continuously differentiable Lyapunov function $V(x, w)$, possibly unknown, such that

$$\alpha_1(\|x\|) \leq V(x, w) \leq \alpha_2(\|x\|) \tag{5}$$

$$\frac{\partial V}{\partial w} S_0 w + \frac{\partial V}{\partial x}[f(x, w) + g(x, w)\psi(x, w)] \leq -W(x) \tag{6}$$

$\forall x \in \mathcal{X} \subset R^n$, $w \in \mathcal{W}$, where $W(x)$ is a continuous positive definite function and $\alpha_1$ and $\alpha_2$ are class $\mathcal{K}$ functions. The system (4) can be re-written as

$$\begin{aligned} \dot{x} &= f(x, w) + g(x, w)\psi(x, w) \\ &\quad + g(x, w)u - g(x, w)[\chi(w) + \psi(x, w)] \end{aligned} \tag{7}$$

Let $\Omega = \{V(w, x) \leq c_1\} \subset \mathcal{X}$ be a compact set for some $c_1 > 0$ and $\delta(x)$ be a function such that

$$\|\chi(w) + \psi(x, w)\| \leq \delta(x) \qquad \forall x \in \Omega, \quad \forall w \in \mathcal{W} \tag{8}$$

Suppose $(\partial V / \partial x) g(x, w)$ can be expressed as

$$(\partial V / \partial x) g(x, w) = \upsilon(x) H(x, w) \tag{9}$$

where $\upsilon(x)$ is a known, locally Lipschitz function, with $\upsilon(0) = 0$, and $H(x, w)$ is a, possibly unknown, function such that $0 < \theta \leq |H(x, w)| \leq k$, $\forall x \in \Omega$, $\forall w \in \mathcal{W}$.

A *conditional servocompensator* [6] is introduced via the saturated high-gain feedback controller

$$u = -\alpha(x) sat\,(s/\mu) \tag{10}$$

where $s = \upsilon(x) + K_1\sigma$, the saturation function is defined as

$$sat\,(s/\mu) = \begin{cases} \frac{s}{|s|} & \text{if} \quad |s| \geq \mu \\ \frac{s}{\mu} & \text{if} \quad |s| \leq \mu \end{cases} \tag{11}$$

and $\sigma$ is the output of the conditional servocompensator

$$\dot{\sigma} = (S - JK_1)\sigma + \mu J sat\left(\frac{s}{\mu}\right) \tag{12}$$

with $\mu > 0$ being the width of the boundary layer. The pair $(S, J)$ is controllable and $K_1$ is chosen such that $S - JK_1$ is Hurwitz. The function $\alpha(x)$ is chosen to satisfy

$$\alpha(x) \geq \frac{k}{\theta}\delta(x) + \alpha_0, \quad \alpha_0 > 0 \tag{13}$$

It is shown in [6] that if $\sigma(0)$ is $O(\mu)$, the state $\sigma(t)$ of the conditional servocompensator (12) will always be $O(\mu)$.

The analysis in [6] shows that, for sufficiently small $\mu$, every trajectory of the closed-loop system (2), (4), (10) and (12) asymptotically approaches a disturbance-dependant manifold of the form $\{x = 0, \sigma = \bar{\sigma}\}$, on which the regulation error is zero. The state feedback design is extended to output feedback for a class of minimum-phase, input-output linearizable systems. For this class of systems, the state feedback control can be designed as a partial state feedback law that does not use the states of the internal dynamics. A reduced-order high-gain observer is used to estimate the states of the linearizable part of the system, which are derivatives of the output. The output feedback controller, obtained by replacing the states by their estimates, recovers the transient and asymptotic properties of the state feedback controller. The performance recovery is shown using the separation principle of [1] and [2].

## III. Inverse optimal control design [7]

It is well known that the optimal control as a design tool guarantees robustness and stability margins. This design approach deals with the problem of finding a feedback control $u(x)$ for the nonlinear system

$$\dot{x} = f(x) + g(x)u \tag{14}$$

with the objective that the $u(x)$ achieves asymptotic stability of the equilibrium $x = 0$ and minimizes the cost functional

$$J = \int_0^\infty (l(x) + u^T R(x) u) dt \tag{15}$$

where $l(x) \geq 0$ and $R(x) > 0$ for all $x$. A direct determination of the optimal feedback law $u(x)$ for nonlinear optimal control problems requires us to solve the Hamilton-Jacobi-Bellman (HJB) partial differential equations. On the other hand, the robustness properties achieved as a result of the optimality do not depend on a particular choice of functions $l(x) \geq 0$ and $R(x) > 0$. This motivated Freeman and Kokotovic' [3], [4] to pursue the development of the design methods that solve the inverse problem of optimal stabilization. In the inverse approach, a stabilizing feedback is designed first and then shown to be optimal for the cost functional (15). The problem is *inverse* since the functions $l(x) \geq 0$ and $R(x) > 0$ are determined through the stabilizing feedback design process rather than being chosen by the designer.

### *A. Design of the Stabilizing Inverse Optimal Control*

A stabilizing control law $u(x)$ solves an inverse optimal control problem for the system (14) if it can be expressed as

$$u = -k(x) = -\frac{1}{2} R^{-1}(x)(L_g V(x))^T, \; R(x) > 0, \tag{16}$$

where $V(x)$ is a positive semidefinite function (to be called a Control Lyapunov Function (CLF), hereafter) and satisfies the following condition, with $u = -\frac{1}{2}k(x)$,

$$\dot{V} = L_f V(x) - \frac{1}{2} L_g V(x) k(x) \leq 0 \tag{17}$$

With the choice of $l(x) \triangleq -L_f V(x) + \frac{1}{2} L_g V(x) k(x) \geq 0$, $V(x)$ is a solution of the HJB equation

$$l(x) + L_f V(x) - \frac{1}{4}(L_g V(x)) R^{-1}(x)(L_g V(x))^T = 0 \quad (18)$$

Therefore, the control law $u(x)$ is an inverse optimal stabilizing control law for the system (14) if it achieves the asymptotic stability of $x = 0$ for the system (14) and is of the form (16) with $V(x)$ that satisfies the condition (17).

The importance of the existence of a CLF in the framework of inverse optimal control design is that, when a CLF is known, an inverse optimal stabilizing control law can be given by Sontag's formula [8]

$$u_s(x) = \begin{cases} -\Big(c_0 + \frac{a_x + \sqrt{{a_x}^2 + ({b_x}^T b_x)^2}}{{b_x}^T b_x}\Big) b_x & , \ b_x \neq 0 \\ 0 & , \ b_x = 0 \end{cases} \quad (19)$$

where $a_x = L_f V(x)$, $b_x = (L_g V(x))^T$, and $c_0$ is a positive constant. It is shown in [7] that the control law (19) is Lipschitz continuous at $x = 0$, if $V(x)$ is a CLF that satisfies the *small control property*[1] for the nonlinear system (14), and is optimal stabilizing for the cost functional

$$J = \int_0^\infty (\frac{1}{2} p(x) b^T(x) b(x) + \frac{1}{2p(x)} u^T R(x) u) dt \quad (20)$$

where

$$p(x) = \begin{cases} c_0 + \frac{a_x + \sqrt{{a_x}^2 + ({b_x}^T b_x)^2}}{{b_x}^T b_x} & , \ b_x \neq 0 \\ c_0 & , \ b_x = 0 \end{cases} \quad (21)$$

An important consequence of the optimality of the control law (19) is that it has a sector stability margin $(\frac{1}{2}, \infty)$, and, under certain assumptions, it achieves a disk stability margin $D(\frac{1}{2})$. These stability margins, which are defined below, provide guaranteed robustness in the presence of matched uncertainties e.g. static nonlinearities, uncertain parameters and the unmodeled fast dynamics.

### B. Stability Margins for Nonlinear Systems

The basic robustness properties of nonlinear feedback systems can be characterized in terms of stability margins, e.g. *gain*, *sector* and *disk stability margins*. Consider the nonlinear feedback system shown in Figure.1 where $u$ and $y$ are of the same dimension and $\Delta$ represents modeling uncertainty at the input side. Under the nominal conditions, the feedback loop consists of the (nominal) nonlinear plant $H$ with the nominal control $u = -k(x) = -y$, and $\Delta$ is identity. The nominal system is denoted by $(H, k)$ and the perturbed system by $(H, k, \Delta)$. The input uncertainties can be static or dynamic. The two most common static uncertainties are *unknown static nonlinearity* and *unknown parameters*. The dynamic uncertainty arises due to unmodeled fast dynamics of the system. The following definitions are due to [7].

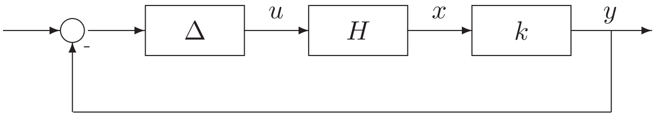


Fig. 1. Nonlinear Feedback Loop with the Control Law $u = k(x)$ and an Input Uncertainty $\Delta$

**Definition 1.** *(Gain Margin) The nonlinear system $(H, k)$ is said to have a gain margin $(\alpha, \beta)$ if the perturbed closed-loop system $(H, k, \Delta)$ is globally asymptotically stable for any $\Delta$ which is of the form $diag\{\kappa_1, \cdots, \kappa_m\}$ with constants $\kappa_i \in (\alpha, \beta), i = 1, \cdots, m$.*

**Definition 2.** *(Sector Margin) The nonlinear system $(H, k)$ is said to have a sector margin $(\alpha, \beta)$ if the perturbed closed-loop system $(H, k, \Delta)$ is globally asymptotically stable for any $\Delta$ which is of the form $diag\{\varphi_1(\cdot), \cdots, \varphi_m(\cdot)\}$ where $\varphi_i(\cdot)$'s are locally Lipschitz static nonlinearities which belong to the sector $(\alpha, \beta)$.*

**Definition 3.** *(Disc Margin) The nonlinear system $(H, k)$ is said to have a disc margin $D(\alpha)$ if the perturbed closed-loop system $(H, k, \Delta)$ is globally asymptotically stable for any $\Delta$ which is globally asymptotically stable and input feedforward passive [7], with a radially unbounded storage function.*

It follows from the definition of disc margin that a nonlinear system having a disc margin $D(\alpha)$ also has gain and sector margins $(\alpha, \infty)$. A disk margin guards against two types of input uncertainties: static nonlinearities and dynamic uncertainties arising from unmodeled fast dynamics of the system.

## IV. PROBLEM FORMULATION AND CONTROL DESIGN

Consider the single-input single-output minimum-phase nonlinear system

$$\begin{aligned} \dot{\zeta} &= \tilde{f}(\zeta, w) + \tilde{g}(\zeta, w)\varphi(u) \\ e &= \tilde{h}(\zeta, w) \end{aligned} \quad (22)$$

where $\zeta \in R^n$ is the state, $u$ is the control input, and $e$ denotes the regulation error. The nonlinear function $\varphi(u)$ belongs to a sector $[\theta_1, \theta_2]$ and satisfies the inequality

$$\theta_1 u^2 \leq u\varphi(u) \leq \theta_2 u^2, \quad 0 \leq \theta_1 \leq \theta_2 \quad (23)$$

The plant is subjected to a set of *exogenous* input variables $w$ that belong to a compact set $\mathcal{W} \in R^w$, which include unknown disturbances to be rejected and references to be tracked. The functions $\tilde{f}$, $\tilde{g}$ and $\tilde{h}$ are sufficiently smooth in $\zeta$ on a domain $\Xi \subset R^n$ and are continuous in $w$ for $w \in \mathcal{W}$. Our goal is to design a controller to asymptotically regulate $e$ to zero.

[1] A nonlinear system $\dot{x} = f(x, u)$, with a known Lyapunov function $V(x)$, is said to satisfy the *small control property* if for every $\epsilon > 0$ there exists a $\delta > 0$ such that for all $\|x\| < \delta$ there exists $u$ with $\|u\| < \epsilon$ so that $\dot{V}(x)$ is negative definite.

We now cast the given output regulation problem in the Lyapunov redesign framework of Section II. As described there, $\zeta = \pi(w)$ is a zero-error invariant manifold and $\chi(w)$ is the steady-state control that maintains the motion on this manifold, in the presence of any exogenous input $w$, which is generated by the exosystem (2). With the change of variables $x = \zeta - \pi(w)$, the system (22) can be represented by

$$\dot{x} = f(x,w) + g(x,w)[\varphi(u) - \chi(w)] \tag{24}$$

where $f(x,w) = \tilde{f}(x+\pi,w) - \tilde{f}(\pi,w) + [\tilde{g}(x+\pi,w) - \tilde{g}(\pi,w)]\chi(w)$ and $g(x,w) = \tilde{g}(x+\pi,w)$.

In what follows, we assume that an optimal stabilizing state feedback controller $\psi(x)$ is available for the nominal system

$$\dot{x} = f(x,w) + g(x,w)u \tag{25}$$

such that the origin of the nominal closed-loop system

$$\dot{x} = f(x,w) + g(x,w)\psi(x) \tag{26}$$

is uniformly asymptotically stable. It is shown in [7] that if $\psi(x)$ is an optimal stabilizing state feedback controller for the system (25) for a cost functional

$$J = \int_0^\infty (l(x) + u^T R(x)u)dt \tag{27}$$

with $l(x) \geq 0$ and $R(x) > 0$ for all $x$, then it achieves a sector margin $(\frac{1}{2},\infty)$. The optimal stabilizing feedback control $\psi(x)$ takes the form

$$\psi(x) = -\frac{1}{2}R^{-1}(x)(L_gV(x))^T,\ R(x) > 0, \tag{28}$$

where the *optimal value function* $V(x)$ is radially unbounded, and is such that the time-derivative of $V$ along the solutions of the closed-loop system (26) is

$$\dot{V} = L_fV(x) - \frac{1}{4}L_gV(x)R^{-1}(x)(L_gV(x))^T \leq 0 \tag{29}$$

As reviewed in Section III, we use the inverse design approach [7], in which a stabilizing feedback controller is designed first and then shown to be optimal for the cost functional (27). When $V(x)$ (called *Control Lyapunov Function*, hereafter) is known, an inverse optimal stabilizing feedback control $\psi(x)$ for the nominal system (25) can be given by Sontag's formula (19), that yields in a sector stability margin $(\frac{1}{2},\infty)$, and, if $R(x) = I$, it achieves a disk stability margin $D(\frac{1}{2})$.

**Assumption 1.** *There exists a smooth positive-definite, and radially unbounded function $V(x,w)$ for the system (25) that satisfies*

$$L_gV(x,w) = 0 \quad \Rightarrow L_fV(x,w) < 0, \quad \forall x \neq 0 \tag{30}$$

**Remark 1.** *Any Lyapunov function whose time-derivative can be rendered negative definite is a CLF. The importance of CLF concept in the framework of inverse optimality is that, when a CLF is known, an inverse optimal stabilizing control such as (19) can be designed, and the CLF becomes an optimal value function.*

It follows that, with a known $V$, the optimal feedback control $\psi(x)$ can robustly stabilize the system (25) in the presence of any sector-bounded nonlinearity $\varphi$ that belongs to a sector $[\theta_1,\theta_2]$. Our goal is to show that with the optimal feedback control $\psi(x)$ we can solve the problem of robust output regulation in the presence of sector nonlinearity $\varphi$. The results of [6] can not be directly applied since the nature of problem differs in that, the Equation (4) depends linearly on control, whereas in the current problem the control depends on a sector-bounded nonlinear function $\varphi$. Furthermore, with Assumption 1, Equations (5)-(6) can be written as

$$\alpha_1(\|x\|) \leq V(x,w) \leq \alpha_2(\|x\|) \tag{31}$$

$$\frac{\partial V}{\partial w}S_0w + \frac{\partial V}{\partial x}[f(x,w) + g(x,w)\varphi(\psi(x))] \leq -W(x) \tag{32}$$

$\forall x \in X \subset R^n$ and $\forall w \in \mathcal{W}$, where $\alpha_1$ and $\alpha_2$ are some class $\mathcal{K}$ functions, $W(x)$ is a continuous positive definite function, and $X$ is a given domain that contains the origin.

The system (24) can also be written as

$$\begin{aligned}\dot{x} \quad &= \quad f(x,w) + g(x,w)\varphi(\psi(x)) + g(x,w)\varphi(u)\\ &\quad\ - g(x,w)[\chi(w) + \varphi(\psi(x))]\end{aligned} \tag{33}$$

We use Lyapunov redesign to construct the saturated high-gain feedback controller to deal with the uncertain term $[\chi(w) + \varphi(\psi(x))]$. Let $\Omega = \{sup_{w\in\mathcal{W}}V(x,w) \leq c_1\} \subset X$, for some $c_1 > 0$, and $\delta(x)$ be a continuous function, independent of the sector-bounded nonlinearity $\varphi(\cdot)$, such that

$$\|\chi(w) + \varphi(\psi(x))\| \leq \delta(x), \quad \forall x \in \Omega,\ \forall w \in \mathcal{W} \tag{34}$$

For simplicity, with $H = 1$, $(\partial V/\partial x)g(x,w)$ as given in (9) can be expressed as

$$(\partial V/\partial x)g(x,w) = \upsilon(x), \quad \forall x \in \Omega,\ \forall w \in \mathcal{W} \tag{35}$$

where $\upsilon(x)$ is a known, locally Lipschitz function, with $\upsilon(0) = 0$. We introduce the conditional servocompensator via the saturated high-gain feedback controller

$$u = -\alpha(x)sat\,(s/\mu) \tag{36}$$

where $s = \upsilon(x) + K_1\sigma$, the continuous function $\alpha(x)$ is chosen such that

$$\alpha(x) \geq \delta(x) + \alpha_0, \quad \alpha_0 > 0 \tag{37}$$

the saturation function is defined as in (11) and $\sigma$ is output of the conditional servocompensator (12).

## V. CLOSED-LOOP ANALYSIS

We will now show that, for sufficiently small $\mu$, the set $\Phi = \Omega \times \left\{V_0(\sigma) \leq \mu^2c_2\right\}$ is a subset of the region of attraction, and for all initial conditions in $\Phi$, every trajectory of the closed-loop system

$$\begin{aligned}\dot{w} \quad &= \quad S_0w\\ \dot{x} \quad &= \quad f(x,w) + g(x,w)\varphi(\psi(x))\\ &\quad\ + g(x,w)\varphi\left(-\alpha(x)sat\,(s/\mu)\right)\\ &\quad\ - g(x,w)[\chi(w) + \varphi(\psi(x))]\\ \dot{\sigma} \quad &= \quad (S - JK_1)\sigma + \mu J sat\,(s/\mu)\end{aligned} \tag{38}$$

asymptotically approaches an invariant manifold on which the error is zero. The forthcoming analysis follows the outline of the analysis in [6], with various technical differences due to the nature of the problem under consideration. We start by showing that the set $\Phi$ is positively invariant and there is a class $\mathcal{K}$ function $\rho$ such that every trajectory in $\Phi$ enters the set $\Phi_\mu = \{\|x\| \leq \rho(\mu)\} \times \left\{V_0(\sigma) \leq \mu^2 c_2\right\}$ in finite time and stays thereafter. The derivative of $V(x, w)$ along the trajectories of the closed-loop system (38) satisfies

$$\begin{aligned}
\dot{V} &= \frac{\partial V}{\partial w} S_0 w + \frac{\partial V}{\partial x}[f(x,w) + g(x,w)\varphi(\psi(x))] \\
&\quad + \frac{\partial V}{\partial x} g(x,w)\varphi\left(-\alpha(x) sat\left(s/\mu\right)\right) \\
&\quad - \frac{\partial V}{\partial x} g(x,w)[\chi(w) + \varphi(\psi(x))] \\
&= \frac{\partial V}{\partial w} S_0 w + \frac{\partial V}{\partial x}[f(x,w) + g(x,w)\varphi(\psi(x))] \\
&\quad + \upsilon(x)\varphi\left(-\alpha(x) sat\left(s/\mu\right)\right) \\
&\quad - \upsilon(x)[\chi(w) + \varphi(\psi(x))] \\
&\leq -W(x) + [s - K_1\sigma]\varphi\left(-\alpha(x) sat\left(s/\mu\right)\right) \\
&\quad - [s - K_1\sigma][\chi(w) + \varphi(\psi(x))] \\
&= -W(x) + s\varphi\left(-\alpha(x) sat\left(s/\mu\right)\right) \\
&\quad - (K_1\sigma)\varphi\left(-\alpha(x) sat\left(s/\mu\right)\right) \\
&\quad - s[\chi(w) + \varphi(\psi(x))] + (K_1\sigma)[\chi(w) + \varphi(\psi(x))]
\end{aligned}$$

Inside $\Phi$, $\|\sigma\| \leq \mu\sqrt{c_2/\lambda_{min}(P_0)}$. Using this along with (23), (37) and (11), it can be shown that when $|s| \geq \mu$, we have [2]

$$\begin{aligned}
\dot{V} &\leq -W(x) - \theta_1\alpha(x)\left|s\right| + \theta_1\alpha(x)\left\|K_1\right\|\left\|\sigma\right\| \\
&\quad + \delta(x)\left|s\right| + \delta(x)\left\|K_1\right\|\left\|\sigma\right\|
\end{aligned}$$

$$\begin{aligned}
\dot{V} &\leq -W(x) - [\theta_1\alpha(x) - \delta(x)]\left|s\right| + [\theta_1\alpha(x) \\
&\quad + \delta(x)]\left\|K_1\right\|\left\|\sigma\right\| \\
&\leq -W(x) + \mu\gamma_1 \qquad (39)
\end{aligned}$$

where $\gamma_1 = \max_{x\in\Omega} k_0[\theta_1\alpha(x) + \delta(x)]$, in which the constant $k_0 = \|K_1\|\sqrt{c_2/\lambda_{min}(P_0)}$.

Similarly, when $|s| \leq \mu$, we have

$$\begin{aligned}
\dot{V} &\leq -W(x) - \theta_1\alpha(x)\frac{s^2}{\mu} + \theta_1\alpha(x)\left\|K_1\right\|\left\|\sigma\right\|\frac{s}{\mu} \\
&\quad + \delta(x)\left|s\right| + \delta(x)\left\|K_1\right\|\left\|\sigma\right\| \\
&\leq -W(x) + \mu\gamma_2 \qquad (40)
\end{aligned}$$

where $\gamma_2 = \max_{x\in\Omega}\ k_0\left[\theta_1\alpha(x) + \delta(x)(1 + 1/k_0)\right] \ \geq \gamma_1$. From (39) and (40),

$$\dot{V} \leq -W(x) + \mu\gamma_2, \quad \forall (x,\sigma) \in \Phi$$

Hence, from [5, Theorem 4.18], for sufficiently small $\mu$, $\Phi$ is positively invariant and all trajectories starting in $\Phi$ enter the positively invariant $\Phi_\mu$ in finite time and stay thereafter.

In the next step, we use $V_s = \frac{1}{2}s^2$, and Assumption 2, below, to show that the trajectories reach the boundary layer $\{|s| \leq \mu\}$ in finite time.

**Assumption 2.** $(\partial\upsilon/\partial x)g(x,w)$ *can be expressed as*

$$(\partial\upsilon/\partial x)g(x,w) = \beta(x), \quad k_p \leq |\beta(x)| \leq k_q, \quad k_q > k_p > 0$$

*for all* $x \in \{\|x\| \leq \rho(\mu)\}$ *and for all* $w \in \mathcal{W}$*. Furthermore,* $\alpha(0) \geq \left(\frac{k_q}{\theta_1 k_p}\right)\delta(0) + \alpha_0,\ \alpha_0 > 0$.

For $(x,\sigma) \in \Phi_\mu$

$$\begin{aligned}
s\dot{s} &= s\frac{\partial\upsilon}{\partial x}[f(x,w) + g(x,w)\varphi(\psi(x))] \\
&\quad + s\beta(x)\varphi\left(-\alpha(x) sat\left(s/\mu\right)\right) \\
&\quad - s\beta(x)[\chi(w) + \varphi(\psi(x))] + sK_1(S - JK_1)\sigma \\
&\quad + \mu s K_1 J sat\left(s/\mu\right)
\end{aligned}$$

When $|s| \geq \mu$, we have

$$\begin{aligned}
s\dot{s} &\leq -k_p\theta_1\alpha(x)\left|s\right| + k_q\left\|[\chi(w) + \varphi(\psi(x))]\right\|\left|s\right| \\
&\quad + \left\|\frac{\partial\upsilon}{\partial x}[f(x,w) + g(x,w)\varphi(\psi(x))]\right\|\left|s\right| \\
&\quad + \left\|\sigma\right\|\left\|K_1\right\|\left\|(S - JK_1)\right\| \\
&\quad + \mu\left\|K_1\right\|\left\|J\right\|\left|s\right|
\end{aligned}$$

Inside $\Phi_\mu$, $\|\sigma\| \leq \mu\sqrt{c_2/\lambda_{min}(P_0)}$. Also, the function $\frac{\partial\upsilon}{\partial x}[f(x,w) + g(x,w)\varphi(\psi(x))]$ is continuous such that $\frac{\partial\upsilon}{\partial x}[f(0,w) + g(0,w)\varphi(\psi(x)(0,w))] = 0$. Therefore, the norm $\left\|\frac{\partial\upsilon}{\partial x}[f(x,w) + g(x,w)\varphi(\psi(x))]\right\|$ together with the norms $\|\sigma\|\,\|K_1\|\,\|(S_J K_1)\|$, $\mu\|K_1\|\,\|J\|$, $\|\alpha(x) - \alpha(0)\|$, and $\|\delta(x) - \delta(0)\|$ can be bounded by a class $\mathcal{K}$ function $\rho_1(\mu)$. Hence,

$$\begin{aligned}
s\dot{s} &\leq -\theta_1 k_p\alpha(0)\left|s\right| + k_q\delta(0)\left|s\right| + \rho_1(\mu)\left|s\right| \\
\Rightarrow \quad \dot{V}_s &\leq -k_p\left[\theta_1\alpha_0 - \frac{\rho_1(\mu)}{k_p}\right]\left|s\right| \qquad (41)
\end{aligned}$$

Thus, for sufficiently small $\mu$, all trajectories inside $\Phi_\mu$ would reach the boundary layer $\{|s| \leq \mu\}$ in finite time. Inside the boundary layer, the closed-loop system (38) is given by

$$\begin{aligned}
\dot{w} &= S_0 w \\
\dot{x} &= f(x,w) + g(x,w)\varphi(\psi(x)) \\
&\quad + g(x,w)\varphi\left(-\alpha(x)s/\mu\right) \\
&\quad - g(x,w)[\chi(w) + \varphi(\psi(x))] \qquad (42) \\
\dot{\sigma} &= S\sigma + J\upsilon(x)
\end{aligned}$$

By following the analysis in [6], it can be shown that inside the boundary layer, the trajectories of the closed-loop system (42) will asymptotically approach an invariant manifold on

[2] When $|s| \geq \mu$, from (23) and (11), we have:

$$\begin{aligned}
\theta_1 u^2 &\leq u\varphi(u) \\
\theta_1\left(-\alpha(x)\frac{s}{|s|}\right)^2 &\leq -\alpha(x) sat\left(s/\mu\right)\varphi\left(-\alpha(x) sat\left(s/\mu\right)\right) \\
-\alpha(x)\theta_1\frac{s^2}{|s|} &\geq s\varphi\left(-\alpha(x) sat\left(s/\mu\right)\right) \\
\Rightarrow -\alpha(x)\theta_1\left|s\right| &\geq s\varphi\left(-\alpha(x) sat\left(s/\mu\right)\right)
\end{aligned}$$

Similarly, when $|s| \leq \mu$, we have:

$$-\alpha(x)\theta_1(s^2/\mu) \geq s\varphi\left(-\alpha(x) sat\left(s/\mu\right)\right)$$

which the regulation error is zero. These conclusions are formally summarized in the following theorem.

**Theorem 1.** *Under stated assumptions, consider the closed-loop system (38). Suppose $w(0) \in \mathcal{W}$. Then, there exists $\mu^* > 0$ such that $\forall \mu \in (0, \mu^*]$, the set $\Psi = \Omega \times \left\{V_0(\sigma) \leq \mu^2 c_2\right\}$ is a subset of the region of attraction, and for all initial conditions in $\Psi$, the state variables are bounded and* $\lim_{t\to\infty} \ e(t) = 0$.

## VI. ILLUSTRATIVE EXAMPLE

Consider the nonlinear system

$$\begin{aligned} \dot{\zeta}_1 &= \zeta_2 \\ \dot{\zeta}_2 &= 2\zeta_1\zeta_2 + u + d(t) \\ y &= \zeta_1 \end{aligned} \tag{43}$$

It is desired to achieve optimal regulation of the system's output $y$ to a constant reference signal $r_0$ in the presence of a disturbance signal, $d(t) = d_0 sin(\omega t)$. Both these signals are generated by the exosystem

$$\dot{w} = \begin{bmatrix} 0 & \omega & 0 \\ -\omega & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} w, w(0) = \begin{bmatrix} d_0 \\ 0 \\ r_0 \end{bmatrix}, \begin{bmatrix} d(t) \\ r_0 \end{bmatrix} = \begin{bmatrix} w_1 \\ w_3 \end{bmatrix}$$

With change of variables $x_1 = \zeta_1 - w_3, x_2 = \zeta_2$, we have

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= 2(x_1 + w_3)x_2 + u + w_1 \\ e &= x_1 \end{aligned} \tag{44}$$

To achieve a sector margin for the nominal system, we use a CLF to design an optimal stabilizing control [7]. With

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, P = \begin{pmatrix} 1 & c \\ c & 1 \end{pmatrix},$$

the Ricatti inequality $A^T P + PA - PBB^T P < 0$ is satisfied for any $c \in (0, 1)$. Then, $V = x^T P x$ is a CLF for the nominal nonlinear system. Assumption 1 is satisfied with this CLF which is also an optimal value function. Furthermore, using Sontag's Formula (19), we get the optimal stabilizing control law for the nominal system as

$$\begin{aligned} \psi(x) &= -2x_1x_2 \\ &\quad - \frac{\alpha_1 x_1 + \sqrt{(2x_1x_2\alpha_2 + x_2\alpha_1)^2 + \alpha_2{}^4}}{\alpha_2} \end{aligned} \tag{45}$$

in which $\alpha_1 = x_1 + cx_2$, and $\alpha_2 = cx_1 + x_2$. This optimal stabilizing control law has two desirable properties, namely, it has a sector margin $(\frac{1}{2}, \infty)$, and it can achieve a disk margin $(\frac{1}{2})$ for the nominal system. The system (44) can also be written as

$$\begin{aligned} \dot{x} &= Ax + B[f(x, w) + g(x, w)(u - \chi(w))] \\ e &= Cx \end{aligned} \tag{46}$$

where

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 0 \end{pmatrix},$$

$$f(x, w) = 2(x_1 + w_3)x_2 + w_1 + g(x, w)\chi(w)$$

and $\chi(w) = -\left(\frac{w_1 + 2w_3}{g(w)}\right)$. With this formulation, the problem fits into the Lyapunov redesign framework as given by (33) with an optimal stabilizing controller (45), and by using the control $u$ of the form (36), it can be shown that the system yields zero steady-state regulation error in the presence of disturbance signal generated by the exosystem.

## VII. CONCLUSIONS

This paper incorporates an optimal stabilizing controller in the Lyapunov redesign framework of [6], and investigates the asymptotic robustness properties of the overall feedback design, given that the optimal stabilizing controller itself possesses strong robustness properties by design. The motivation comes from the flexibility of incorporating any stabilizing controller within the framework as proposed in [6], and we seek for control design methods that yield stabilizing controllers with some additional desirable properties like optimality, disturbance rejection and robustness in the presence of matched uncertainties e.g. static nonlinearities, uncertain parameters and the unmodeled fast dynamics. The synthesis of such controller requires the solution of the complicated Hamilton-Jacobi-Bellman (HJB) partial differential equations. Here, we exploit the inverse optimal control design methods presented in [7] for nonlinear systems, where it is shown that in addition to achieving the asymptotic stability of the system and minimizing a cost functional, the optimal feedback control law which is designed based on the existence of a Control Lyapunov Function (CLF), guarantees stability margins which characterize the system's robustness properties. A simple illustrative example is also presented to delineate the overall design process.

## VIII. ACKNOWLEDGEMENTS

The author gratefully acknowledges the valuable guidance provided by Professor Hassan K. Khalil.

## REFERENCES

[1] Atassi A. N., Khalil H. K., *A Separation Principle for the Control of a Class of Nonlinear Systems*, IEEE Transactions on Automatic Control; **46**: 742 − 746; 2001.

[2] Atassi A. N., *A Separation Principle for the Control of a Class of Nonlinear Systems*, PhD Thesis; College of Engineering, Michigan State University, East Lansing, MI; 1999.

[3] Freeman, R. A., and Kokotovic', P., *Optimality of robust nonlinear feedback controls*, SIAM Journal on Control and Optimization; Volume 34 , Issue 4, pp. 1365 − 1391, 1996.

[4] Freeman, R. A., and Kokotovic', P., *Robust control of nonlinear systems*, Birkhauser, Boston, 1996.

[5] Khalil, H. K., *Nonlinear Control Systems, 3rd edition*, Prentice-Hall: Upper Saddle River, New Jersey, 2002.

[6] Memon A. Y., Khalil H. K., *Output Regulation of Nonlinear Systems Using Conditional Servocompensators*, Automatica, 46 (7), pp. 1119-1128, 2010.

[7] Sepulchre, R., Jankovic', M., and Kokotovic', P., *Constructive nonlinear control*, Springer, London, 1997.

[8] Sontag, E. D., *Remarks on stabilization and input-to-state stability*, In proceedings of 28th Conference on Decision & Control, Tampa, FL, pp. 1376 − 1378, 1989.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# *Research on a novel PID based controller for non-magnetic hydraulic navigation simulator with AMESim simulation*

Zhou Kaibo, Wang Xuyong, Tao Jianfeng, Guo Xiaofeng, Xu Chuanhui
School of Mechanical Engineering
Shanghai Jiaotong University
Shanghai, China
zhoukaibo87@gmail.com

*Abstract*—**In this paper we research on the run time characteristics of non-magnetic hydraulic navigation simulator and approach a corresponding control strategy. Since the simulator is designed to be non-magnetic, it consists of long hydraulic pipes separating electric part and mechanical part to avoid the magnetic interference from electric part. Meanwhile the mechanical part is made of non-magnetic materials. Hydraulic motor, instead of electric motor, is used to drive the simulator and it stands varied load at run time. Because of the non-magnetic design, normal control strategy doesn't satisfy the system. We analyze the characteristics of the simulator system both in frequency-domain and time-domain with the help of transfer function and AMESim, which is a modeling environment for simulation of engineering, to make the characteristics clear and finally propose a PID based control strategy (denoted as $D_{ff}$ - $PIDD^2$ in this paper) combining feedforward differential controller ($D_{ff}$) and 2nd order differential controller ($D^2$) for the system. At last we verify the control strategy with position tracking simulation and obtain a satisfied result.**

***Keywords-non-magnetic; long-hydraulic pipe; navigation simulator; $D_{ff}$ - $PIDD^2$; AMESim simulation***

## I. Introduction

Non-magnetic navigation simulator is a platform which is used to test and to calibrate magnetic navigation devices [1]. Magnetic navigation devices are widely used to obtain accurate direction data in various fields, such as military, aviation, aerospace, resource exploration, and etc. When being tested and calibrated, magnetic navigation device is installed in the middle of inner ring of the simulator (Fig. 1) and the direction data from magnetic navigation device will be compared with the precisely controlled rotation data of the simulator. The simulator we work on is designed for a magnetic navigation device which navigates depending on the geomagnetic field. The geomagnetic field is very small (6e-5T averagely) compared to common magnetic fields in our daily life and is easy to be deformed in space by some paramagnetic or diamagnetic media. Because of the special characteristics of the geomagnetic field, the simulator has to be designed without influencing the magnitude and direction of the geomagnetic field nearby.

To achieve the non-magnetic requirement, generally there are three driven schemes for navigation simulator: motor driven, pneumatic driven and hydraulic driven. Motor driven simulator requires separating the motor and the tested device to avoid the magnetic interference from electric motor. The non-magnetic motion simulator used by U.S. Coastal System Station is based on motor driven scheme [2]. Pneumatic driven avoids the magnetic interference in principle. The motion platform designed by University of Newcastle, UK, which is used to test high-precision magnetic device, is based on pneumatic driven scheme [3]. Hydraulic driven also in principle avoids the magnetic interference and the compressibility of hydraulic oil is much less than gas so that it is more likely to achieve high-precision positioning and fast motion tracking. Besides driven scheme, non-magnetic materials are necessity to construct the simulator. Some commonly used non-magnetic materials are austenitic stainless steel, copper alloy, polymers, ceramic material, and etc. [4].

The non-magnetic navigation simulator we designed is based on the hydraulic driven scheme and many other design features, which will be included in the next section, to achieve non-magnetic requirement. Because of the non-magnetic requirement, the run time characteristics of the system are quite different from normal navigation simulator and normal PID control strategy doesn't satisfy the system. In this paper we focus on the derivation of control strategy for the non-magnetic navigation simulator. And we focus more on the theoretical control scheme rather than real application details so that only the most important and crucial features of the system will be counted and some necessary supplementary control strategies for real application, e.g. using low-pass filter to eliminate high frequency noise signal from sensor, may be neglected in this paper. The performance of the derived control strategy will be verified by a simulated simulator system tracking a sine wave position singal with AMESim software.

National Natural Science Foundation of China [grant numbers 50975172, 50805090]

Research Fund of State Key Lab of MSV, China [grant number MSV-2010-05]

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

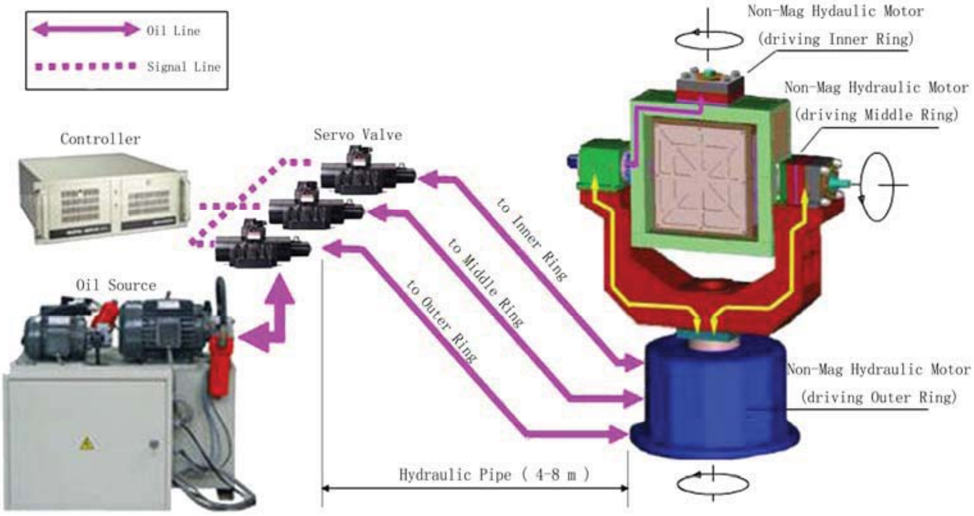


Figure 1. System composition of non-magnetic hydraulic navigation simulator

## I. Principle of the Navigation Simulator

The basic structure of the non-magnetic hydraulic navigation simulator system is shown in Fig. 1.

Generally we divide the system into the electronic control part and the mechanical part. They are connected by signal lines mainly including current signals from controller to servo valves and feedback angular displacement signals from angle sensors to controller. Angle sensor is mounted at the end of each non-magnetic hydraulic motor. The angle sensor on the simulator is a kind of optical encoder, which is a passive device and has an acceptable influence on geomagnetic field [5].

The fundamental principle of the whole system is close to a classic valve controlled hydraulic motor system. But there are several different features in this navigation simulator, which are mainly designed to achieve the non-magnetic requirement. These features are:

- We use beryllium bronze (Cu-Be-Co) [6] and other non-magnetic materials for hydraulic pipe and mechanical components of the navigation simulator;
- We use hydraulic motor made of non-magnetic materials instead of electric motor and optical angle sensor instead of electromagnetic sensor;
- In order to eliminate the electromagnetic interference from the servo valve and oil source side, we separate the mechanical part and the electric part by connecting them with long hydraulic pipes.

When the simulator runs, three servo valves independently control the rotation of three rings, which are known as pitch (outer ring), course (middle ring), and roller (inner ring). The controller outputs the current signals, which are computed based on the feedback angular displacement and our control strategy, to control the spool position of servo valve so that each ring can rotate at a set speed or to a set position.

## II. Modeling with AMESim

In order to know the run time characteristics of navigation simulator especially the influence of non-magnetic designs on the system and approach a suitable control strategy, we model and simulate the system with AMESim (Fig. 2). We set up two models for different reasons. The model with no controller is used to simulate the open-loop characteristic of the system and to check the characteristic of each component. The model with $D_{ff}$ - PIDD$^2$ controller is to test the effectiveness of our control strategy and to simulate the open-loop characteristic of the controller-controlled system as well. During modeling we focus on the most important parts of the system to make the model neat and effective. Here are some instructions for the AMESim model of non-magnetic navigation simulator system:

- We set the oil source pressure to follow the equation of $p_s$=95+5sin(200πt) (bar) to simulate the pressure ripple caused by periodic rotation of pump at oil source side;
- To smooth the ripple and to keep a steady oil pressure inside the pipe we set up an accumulator right after oil source and the filter;
- We select a three-position-four-port servo valve, the spool dynamics of which is modeled as a 2nd order system with a specified natural frequency and damping ratio;
- The hydraulic pipe is modeled as distributive hydraulic line with lumped elements [7];
- Without losing the effectiveness of the model, we simply model the hydraulic motor as an ideal fixed displacement hydraulic motor and we only simulate one ring of the simulator because the three rings have similar characteristics and work independently;
- The rotary load is set with specified moment of inertia, coefficient of viscous friction, Coulomb friction torque, and stiction torque [8].

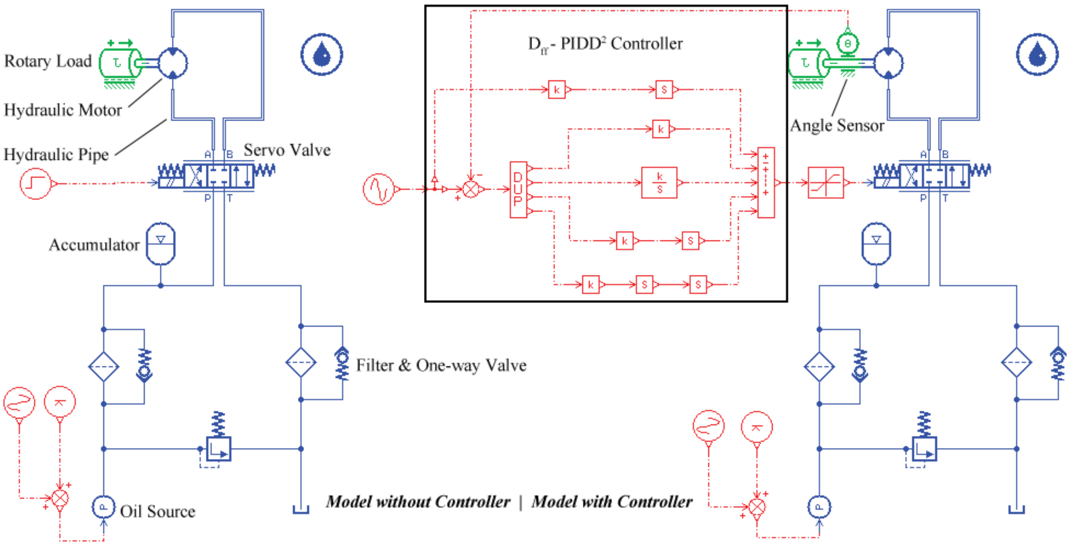


Figure 2. Model with no controller (left) and model with $D_{ff}$ - PIDD$^2$ controller (right)

The distributive hydraulic line model describes the hydraulic pipe as line with several nodes (Fig. 3). The compressibility of the fluid and expansion of the pipe/hose wall with pressure are considered in the model by using an effective bulk modulus. Pipe friction is taken into account using a friction factor based on the Reynolds number and relative roughness. And inertia of the fluid is also taken into account. The parameters of the whole model are set in Table I.

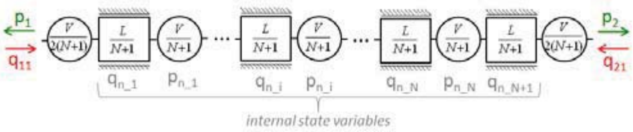


Figure 3. The principle of distributive hydraulic line model (Source: AMEHelp HL042)

TABLE I. PARAMETERS OF NON-MAGNETIC HYDRAULIC NAVIGATION SIMULATOR AMESIM MODEL

| No. | Parameters | Value or Expression |
|---|---|---|
| 1 | Oil source pressure (bar) | 95+5sin(200πt) |
| 2 | Accumulator volume (L) | 6.3 |
| 3 | Accumulator pre-charge pressure (bar) | 80 |
| 4 | Valve rated current (mA) | 40 |
| 5 | Valve nature frequency (Hz) | 80 |
| 6 | Valve damping ratio (null) | 0.8 |
| 7 | Valve flow rate at maximum opening (L/min) | 10 |
| 8 | Valve pressure drop at maximum opening (bar) | 35 |
| 9 | Pipe inner diameter (mm) | 8 |
| 10 | Pipe wall thickness (mm) | 2 |
| 11 | Pipe length (m) | 5 |
| 12 | Pipe relative roughness (null) | 1.5e-4 |
| 13 | Pipe Young's modulus (bar) | 1.3e+6 |
| 14 | Pipe internal node number (null) | 5 |
| 15 | Motor displacement (cm$^3$/rev) | 80 |
| 16 | Load moment of inertia (kgm$^2$) | 5 |
| 17 | Load viscous friction coefficient (Nm/[rev/min]) | 0.1 |
| 18 | Load stiction torque (Nm) | 4 |
| 19 | Load coulomb friction torque (Nm) | 2 |
| 20 | Position tracking signal (degree) | 40sin(πt) |

## III. Simulation and Analysis

### A. *Fundamental analysis of the system*

As we mentioned in the principle section, the fundamental part of the system is a valve controlled hydraulic motor system. So firstly we study on the characteristics of this fundamental part in frequency-domain. The linearized flow rate incremental equation of hydraulic valve is defined as:

$$\Delta q_{Lv} = K_q \Delta x_v - K_c \Delta p_{Lv} \tag{1}$$

where $q_{Lv}$ is flow rate, $x_v$ is spool position, $p_{Lv}$ is load pressure, $K_q$ is flow rate gain, $K_c$ is flow-pressure coefficient. Then we give out the expression of angular displacement of hydraulic motor directly [9]:

$$\theta_m(s) = \frac{\frac{K_q}{D_m}\frac{K_{sv}}{\frac{s^2}{\omega_{sv}^2}+\frac{2\zeta_{sv}}{\omega_{sv}}s+1} i_c(s) - \frac{K_{ce}}{D_m^2}(1+\frac{V_t}{4\beta_e K_{ce}}s)T_{ml}(s)}{s(\frac{s^2}{\omega_h^2}+\frac{2\zeta_h}{\omega_h}s+1)} \tag{2}$$

where $i_c$ is current signal output by controller, $T_{ml}$ is torque load reflected on hydraulic motor shaft, $D_m$ is displacement of motor, $K_{ce}$ is total flow-pressure coefficient, $\beta_e$ is effective bulk modulus of motor system, $\omega_{sv}$ is servo valve natural frequency, $\zeta_{sv}$ is servo valve damping ratio, $K_{sv}$ is servo valve gain, $\omega_h$ is hydraulic natural frequency, $\zeta_h$ is hydraulic damping ratio.

In order to reduce the complexity of the model, we temporarily make the assumption $T_{ml}(s) = 0$, which means motor load is zero for the moment. Then (2) reduces to

$$\frac{\theta_m(s)}{i_c(s)} = \frac{K_q K_{sv}/D_m}{s(\frac{s^2}{\omega_h^2}+\frac{2\zeta_h}{\omega_h}s+1)(\frac{s^2}{\omega_{sv}^2}+\frac{2\zeta_{sv}}{\omega_{sv}}s+1)} \tag{3}$$

It can be seen from (3), if we neglect the long hydraulic pipe and assume zero motor load, the model with no controller is a first order astatic system with two oscillation elements, which can be theoretically regulated by a PID controller [10].

### B. *Analysis of hydraulic motor load*

The second step we take the hydraulic motor load into account. The motor load is defined as:

$$T_{ml} = C_1\dot{\omega}_m + C_2\omega_m + C_3(1-|\mathrm{sgn}(\omega_m)|) + C_4 \tag{4}$$

where $\omega_m$ is the angular velocity of motor, $C_1$, $C_2$, $C_3$, and $C_4$ are constants defined in Table I (No. 16-19). The transfer function of the system will become very complicated when combining (2) and (4), and it will be difficult to draw a proper control strategy. So we analyze the system in time-domain based on AMESim simulation result.

The motor load influences the system by changing the load pressure, i.e. the pressure difference between the inlet and outlet of motor. Recalling (1), we see if the opening amount of valve spool keeps unchanged while the load pressure increases/decreases, the flow through the motor will reversely decrease/increase. In order to eliminate the varied motor load influence on the flow rate, we introduce compensation current to simultaneously act on the opening amount of spool. According to (4) we need to involve angular velocity and acceleration of the motor and a constant into control signals. Because the velocity signal is already able to be reflected in the PID controller and the constant is relatively small compared to other terms in (4), we only introduce an extra 2nd order differential controller ($D^2$) to get the acceleration of the motor. Through the simulation of model with $D_{ff}$ - PIDD$^2$ controller tracking a sine wave position signal in AMESim (detailed simulation parameters will be included in later section), it can be seen from Fig. 4 that the simulated output of $D^2$ controller is always in the same pace with the motor load and reflects the acceleration of motor.

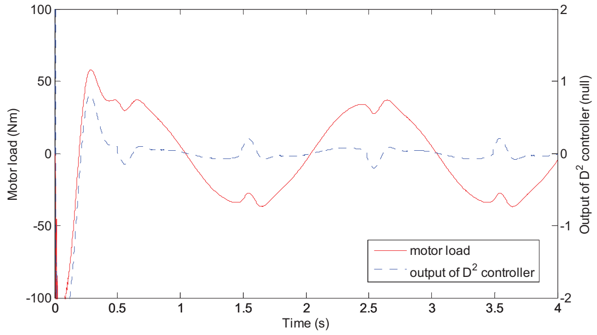


Figure 4. Comparison of motor load and $D^2$ controller output in model with $D_{ff}$ - PIDD$^2$ controller

### C. *Analysis of long hydraulic pipe*

At last we research on control strategy with long hydraulic pipe taken into consideration. The distributive hydraulic line model describes the pipe as a line with several nodes and calculates the pressure and flow rate in each node with the following equations (Source: AMEHelp HL042):

$$\begin{cases} \frac{\partial p_p}{\partial t} = -\frac{\beta_{ep}}{A_p}\cdot\frac{\partial q_p}{\partial x} \\ \frac{\partial q_p}{\partial t} = \frac{A_p}{\rho}\cdot\frac{\partial p_p}{\partial x} - gA_p\sin(\theta_p) - v\frac{\partial q_p}{\partial x} - \frac{f_f q_p^2 \mathrm{sgn}(q_p)}{2D_pA_p} \end{cases} \tag{5}$$

where $p_p$ is pressure in node, $q_p$ is flow rate in node, $A_p$ is the cross-sectional area of pipe, $\beta_{ep}$ is the effective bulk modulus of pipe/fluid combination, $\rho$ is the oil density, $D_p$ is the diameter of the section of pipe, $\theta_p$ is the inclination of pipe, $v$ is the mean velocity of flow in node, $f_f$ is the friction factor of node.

It can be seen that, the derivative of the flow rate is determined by applying conservation of momentum principles to a control volume cell. Forces due to pressure, gravity and pipe friction are taken into account as well as convection of momentum. And we again analyze the pipe effect in time-domain. As shown in Fig. 5, through the simulation of model with no controller tracking a step position signal, the simulated flow rate in hydraulic motor delays about 0.2 second compared to flow rate at port A of the servo valve.

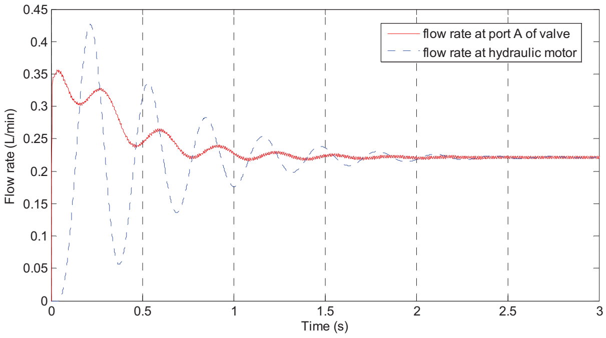


Figure 5. Flow rate in model with no controller

Because pipe is the only component connecting motor and servo valve, it is obviously that the long hydraulic pipe causes the time delay. It means if we could give out the tracking signal a certain time period in advance, we are able to cancel the flow rate delay in pipe. We now assume the tracking signal can be described as a continuous differentiable function f(t), then for $t_B > t_A$ and $t_B - t_A << 1$ we have

$$\begin{aligned} f(t_B) &= f(t_A + t_B - t_A) \\ &= f(t_A) + f'(t_A)(t_B - t_A) + f''(t_A)(t_B - t_A)^2 + \cdots \end{aligned} \quad (6)$$

Since $t_B - t_A << 1$, we neglect the terms with order equal to or higher than 2. Equation (6) reduces to

$$f(t_B) - f(t_A) \approx f'(t_A)(t_B - t_A) \quad (7)$$

Equation (7) shows that we can approach the incremental of signal at $t_B$ to signal at $t_A$ with the derivative of signal at $t_A$. Inspired by this equation we introduce the feedforward differential controller ($D_{ff}$) into control strategy:

$$D_{ff} = Ks \quad (8)$$

Compared to (7), we roughly have gain $K = t_B - t_A$.

### *D. Overall control strategy and simulation result*

Combining all above analysis, we draw the overall control strategy for the system. We use a PID controller as fundamental control strategy together with a $D^2$ controller regulating the hydraulic motor load fluctuation and a $D_{ff}$ controller eliminating time delay effect caused by long hydraulic pipe. In order to check the effectiveness of the controllers, we simulate a position tracking test in AMESim. The tracking signal is set to be a sine wave position signal described in Table I (No. 20), which meets the non-magnetic simulator specification. The result of the simulation test is shown in Fig. 7 with controller parameters in Table II.

Through the comparison between $D_{ff}$ - PID control and PID control, we see the $D_{ff}$ controller effectively eliminates the time delay of the system. Through the comparison of $PIDD^2$ control and PID control, we see the $D^2$ controller reduces the tracking error fluctuation of the system at start stage, during which the motor load changes rapidly and greatly (Fig. 4). Because our system stands varied motor load and the feedback signal is angular displacement of motor, the $D^2$ controller precisely catches the load fluctuation and compensates it to the control current signal. $D^2$ controller is rarely seen in traditional control strategy because it usually introduces instability to system. In real applications a low-pass filter might be used to eliminate high frequency noise signal from sensor.

Besides the time-domain analysis with AMESim, we also do a frequency-domain test (Bode diagram) of the two models in open-loop state. We analyze the stability of the two models based on Bode diagram stability criterion for open-loop system (Fig. 6) [11]. This criterion defines the magnitude crossover frequency ($\omega_{mc}$) as the frequency where the magnitude is equal to 0 and the phase crossover frequency ($\omega_{pc}$) as the frequency where phase shift is equal to -180 degree. If at the phase crossover frequency, the corresponding magnitude is less than 0 dB, then the feedback system is stable. According to Bode diagram criterion (using the vertical lines in Fig. 8 as reference), the model with no controller is not stable while the model with controller is stabilized. The diagram also shows that model with $D_{ff}$ - $PIDD^2$ controller has faster response in low frequency region and broader frequency band without phase delay. Although the attenuation of model with controller in high frequency region becomes slower, it is within an acceptable amount.

TABLE II. CONTROLLER PARAMETERS FOR SINE POSITION TRACKING SIMULATION

| **Controller** | **PID** | **PIDD$^2$** | **D$_{ff}$ - PID** | **D$_{ff}$ - PID** |
|---|---|---|---|---|
| $D_{ff}$ | null | null | 0.05 | 0.05 |
| P | 0.6 | 0.6 | 0.6 | 0.6 |
| I | 0.5 | 0.5 | 0.5 | 0.5 |
| D | 0.05 | 0.05 | 0.05 | 0.05 |
| $D^2$ | null | 0.002 | null | 0.002 |

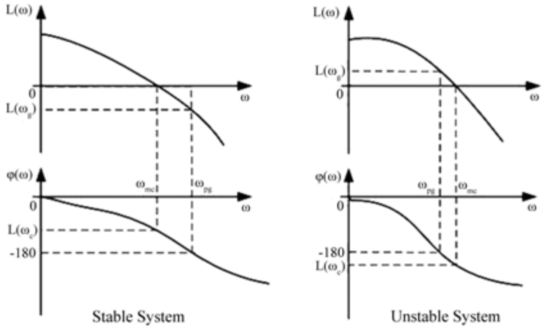


Figure 6. Bode diagram stability criterion for open-loop system

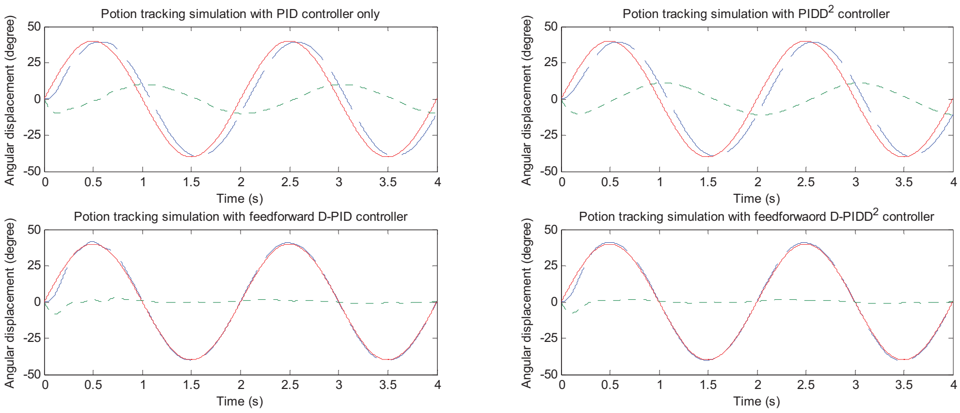


Figure 7. Simulation result of sine position signal tracking (solid line: target signal, dashed line: tracking result, dense dashed line: tracking error)

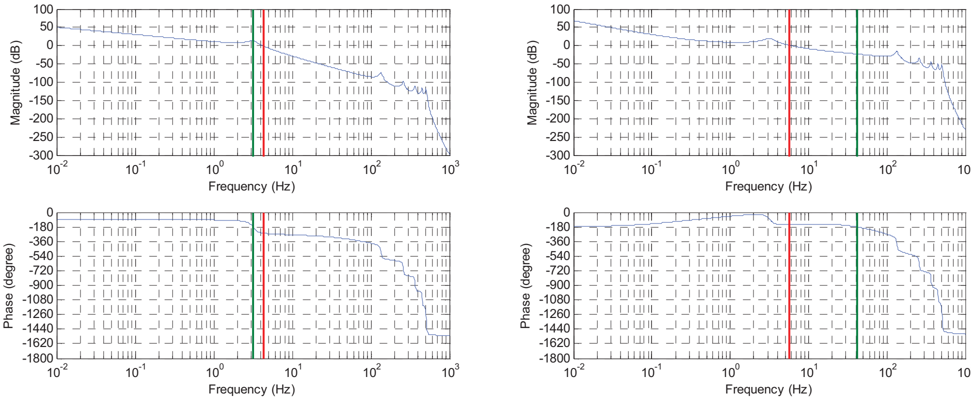


Figure 8. Bode diagram of model with no controller (left) and model with $D_{ff}$-PIDD$^2$ controller (right)

## IV. Conclusion

Through modeling, simulation and analysis of non-magnetic hydraulic navigation simulator, we conclude that the simulator is a time-delay valve controlled hydraulic motor system with varied motor load. We firstly focus on the most fundamental parts of the system, i.e. valve controlled hydraulic motor, which is a common hydraulic system and on which mature control strategy can be applied. Secondly we analyze the special parts of the system in time-domain with simulation results. During this process we isolate each component and come up with different control strategies. Finally we propose the $D_{ff}$ - PIDD$^2$ controller for the system, which is proven more effective than the traditional PID controller theoretically and with simulation as well. The $D_{ff}$ - PIDD$^2$ controller may also be applied to real applications of similar system with time delay or varied force/torque load.

## References

[1] L. A. DeMore, R. A. Peterson, L. B. Conley, and et al., “Design study for a high-accuracy three-axis test table,” Journal of Guidance, 1987, 10(1): 104-114.

[2] George I. Allen, John Purpura, David Overway, “Measurement of magnetic noise characteristics on select AUVS with some potential mitigation techniques,” OCEANS '02 MTS/IEEE. 2002, 9782-9978.

[3] Kevin P. Humphrey, Thomas J. Horton, Mark N. Keene, “Detection of mobile targets from a moving platform using an actively shielded, adaptively balanced SQUID gradiometer,” IEEE Transaction on applied superconductivity. 2005, 15(2): 753-756.

[4] K.H.J. Buschow. Handbook of Magnetic Materials, vol 18. North Holland Press, 2009.

[5] Kenneth Schofield, Desmond J. O'Farrell, Kenneth L. Schierbeek, Magnetic compass with optical encoder. US4937945, Jul 3 1990.

[6] Ana Galet, Ana Belén Gaspar, M. Carmen Muñoz, and et al., “Tunable bistability in a three-dimensional spin-crossover sensory- and memory-functional material,” Advanced Materials, Volume 17, Issue 24, pages 2949–2953, December, 2005.

[7] Kong Xiaowu, Qiu Minxiu, “A study of the influences of pipe on valve control hydraulic system,” Proceedings of the 5th International Conference on Fluid Power Transmission and Control. Hangzhou: Zhejiang University Press, 2001:23-27.

[8] Valentin L. Popov, Contact Mechanics and Friction: Physical Principles and Applications. Springer, 2010.

[9] Yang Zhengrui, Hua Keqin, Xu Yi, Electro-hydraulic Propotional and Servo Control. Beijing: Metallurgical Industry Press, 2009.8.

[10] Wang Xianzheng, Chen Zhenghang, Wang Xuyong, Control Theory. Beijing: Science Press, 2000.

[11] Benjamin C. Kuo, Farid Golnaraghi, Automatic Control Systems. John Wilev & Sons, 2003.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Trajectory tracking control of a car-like mobile robot in presence of sliding

Faiza Hamerlain
Division Robotique et Productique, CDTA,
Cité du 20 Août 1956, BP N°17, Baba Hassen, Alger, Algérie.
hamerlainf@yahoo.fr

Abstract- This paper investigates the trajectory tracking control problem of a nonholonomic car-like mobile robot in the presence of sliding effects. Sliding (slipping and skidding) effects are treated as disturbances and introduced into the kinematic model of the car-like using the singular perturbation approach. In order to compensate for the effects of tire slipping and skidding, a robust second order sliding mode controller is developed based on the super twisting algorithm. It is theoretically proven that for car-like vehicle subjected to sliding, the lateral-longitudinal deviations and the orientation errors can be stabilized near the origin. Simulations results show the effectiveness and the robustness of the proposed controller with respect to the sliding effects.

Index Terms- Car-like Mobile Robot, Second order sliding mode, Tracking control, Sliding effects.

## I. Introduction

The control problem of nonholonomic systems with parameter uncertainties has been extensively studied in past decades. Regarding the trajectory tracking problem of the Wheeled Mobile Robots WMR, many works address violation of the nonholonomic constraints. A robust control law against decoupled skidding and slipping effects has been proposed for solving a velocity tracking problem for a unicycle-type robot [1]. For a dynamic model of the unicycle-type robot including sliding effects, simulation results were reported in [2], showing the robustness of a second order sliding tracking control with respect to the non ideal constraints and to the nonlinearities. In [3], a path tracking control of an autonomous robot was considered under slipping and unknown dynamics. A robust adaptive neural network controller was proposed for WMR with the aid of backsteeping techniques and the learning ability. Based on the backsteeping methods, a robust adaptive tracking controller was designed for farm agriculture vehicles in presence of sliding [4]. The sliding effects are introduced as time-varying parameters to the ideal kinematic model. Both simulation and experimental results confirm the high longitudinal-lateral tracking controller accuracy. In [5], it was derived kinematic models of WMR that explicitly relate the perturbations to the vehicle skidding and slipping that are geometrically well defined. This description allows the analyse of these perturbations from a control perspective. The reference [6] copes with the control of WMR not satisfying the ideal kinematic constraints by using slow manifolds methods, but the parameters characterizing the sliding effects are assumed to be exactly known. In [7], a time varying stabilizing control law based on the linear quadratic theory was proposed and the necessity for the trajectory to satisfy the dynamics of the skidding effects is pointed out. This feedback control law ensures the local asymptotic convergence of the error dynamics of the unicycle but only under some conditions on the reference trajectory (accelerations should be sufficiently small). In [8], a singular perturbation formulation was derived which leads to robust linearizing feedback laws ensuring trajectory tracking in presence of sufficiently small sliding effects. In [9], a discrete-time sliding mode control was proposed for trajectory tracking of the kinematic model of a unicycle type mobile robot in the presence of skidding effects. The sliding effect is modeled taking into account its most important component which is skidding for no straight trajectories. In [10], a robust adaptive controller based on the tunable dynamic oscillator was developed for a skid steer mobile robot in the presence of disturbances violating nonholonomic nonslipping constraint. Simulation results were performed to solve both the tracking and regulation problems.

In this paper, the robust trajectory tracking problem for a kinematic model of a car-like WMR in the presence of sliding effects is solved by means of a higher order sliding mode control. Using the singular perturbation approach [8], the kinematic tracking model of the car-like WMR is derived. Then, a second order sliding mode controller of the super twisting algorithm is used. The proposed control law is based on two nonlinear sliding manifolds ensuring the asymptotic tracking of the output variables in spite of the transgression of the nonholonomic constraints during the motion.

This paper is organized as follows, in Section 2 a kinematic tracking model considering sliding effects is derived and the problem statement is presented. The second order sliding mode control law is presented in Section 3, while Section 4 shows simulation results.

## II. Model of the car-like and problem statement

The Robucar in a single drive mode is an example of car-like WMR. For model simplicity, we assume that each axle of the physical wheels (front and rear) is represented by a virtual wheel located in the middle axle. Then, the schematic representation of the car-like WMR can be given by Figure 1. It is fully described by a four-dimensional vector of generalized coordinates $q = (x_P, y_P, \theta, \phi)^t$ where $x_P, y_P$ are the coordinates of point $P$, $\theta$ is the orientation angle of the vehicle w.r.t to a fixed frame and $\phi$ represents the front steering angle relative to the car body. $L$ denotes the WMR's wheelbase.

Note that mobile robot of type $(\delta_m, \delta_s)$ with a degree of mobility $\delta_m = 1$ and a degree of steerability $\delta_s = 1$ present four independent constraints (the non skidding constraint being the same for both driving wheels) of the form:

$$A^t(q)\dot{q} = 0 \tag{1}$$

with

$$\begin{aligned} A &= \begin{pmatrix} s(\theta) & s(\theta+\phi) & c(\theta) & c(\theta+\phi) \\ -c(\theta) & -c(\theta+\phi) & s(\theta) & s(\theta+\phi) \\ 0 & -s(\theta) & 0 & s(\theta) \\ 0 & 0 & 0 & 0 \end{pmatrix}, \\ c(\theta) &= \cos(\theta), s(\theta) = \sin(\theta). \end{aligned}$$

which correspond to the ideal hypothesis of a " pure rolling and non slipping " condition.

Defining a full matix $S(q)$ such that $A^t(q)S(q) = 0$, a vector $v = (v_1, v_2)^t$ exists satisfying

$$\dot{q} = S(q)v \tag{2}$$

with

$$S = \begin{pmatrix} c(\theta) & 0 \\ s(\theta) & 0 \\ \tan(\phi) & 0 \\ 0 & 1 \end{pmatrix},$$

where $v_1, v_2$ represent respectively the linear driven (steering) velocity of the car-like WMR

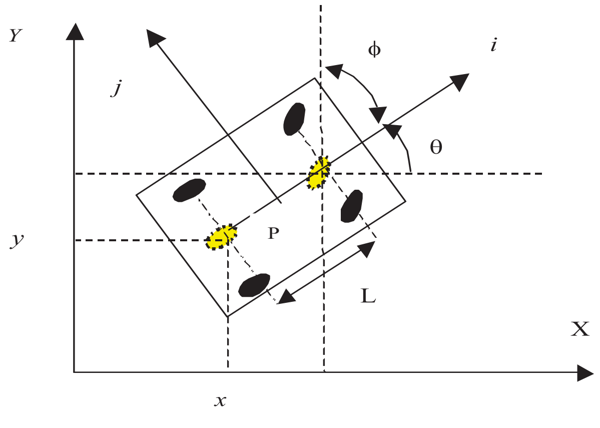


Figure 1. The car-like WMR.

In practise, it is well known that the nonholonomic constraints are never strictly satisfied during the motion of robot, due to various factors such as sliding, deformability or flexibility of the wheels. Hence, the kinematic model given by (2) is no longer valid. So, the interaction forces and slipping effects have to be modelled. Based on the singular perturbation formalism [8], the kinematic model of the car-like WMR with sliding can be expressed as:

$$\dot{q} = S(q)v + A_1\epsilon\mu_1 + A_2\epsilon\mu_2 + A_3\epsilon\mu_3 + A_4\epsilon\mu_4,$$

where $A_1, A_2, A_3, A_4$ are the colomns of the matrix $A, \mu$ is a four dimension vector reflecting the violation of the constraints and $\epsilon$ is a positive scalar, which is the inverse of the largest stiffness. $\epsilon\mu_1$ and $\epsilon\mu_2$ corresponds to skidding effects while $\epsilon\mu_3$ et $\epsilon\mu_4$ represents slipping effects.

Let $v_x$ and $v_y$ be the longitudinal and the lateral velocities of point $P$ (see Figure 1), respectively, and $v_\theta$ the vehicle angular velocity:

$$\begin{cases} v_x = v_1 + s(\phi)\varepsilon\mu_2 + \varepsilon\mu_3 + c(\phi)\varepsilon\mu_4, \\ v_y = \varepsilon\mu_1 + c(\phi)\varepsilon\mu_2 - s(\phi)\varepsilon\mu_4, \\ v_\theta = \tan(\phi)v_1 - c(\phi)\varepsilon\mu_2 + s(\phi)\varepsilon\mu_4, \end{cases} \tag{3}$$

Under the hypothesis of a quite small front steering angle, the kinematic model of the car-like WMR is given in the base frame by time evolution of point $P$ of coordinates:

$$\begin{pmatrix} \dot{x}_P \\ \dot{y}_P \\ L\dot{\theta} \end{pmatrix} = M(\theta) \begin{pmatrix} v_x \\ v_y \\ v_\theta \end{pmatrix}$$

with

$$M(\theta) = \begin{pmatrix} c(\theta) & s(\theta) & 0 \\ s(\theta) & -c(\theta) & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

The objective is to make the car-like robot follow a given reference trajectory that satisfies the kinematic equations of a virtual WMR:

$$\begin{pmatrix} \dot{x}_{Pr} \\ \dot{y}_{Pr} \\ L\dot{\theta}_r \end{pmatrix} = M(\theta_r) \begin{pmatrix} v_{xr} \\ v_{yr} \\ v_{\theta r} \end{pmatrix}, \tag{4}$$

where $(x_{Pr}, y_{Pr})$ is the position of a virtual reference vehicle at the point $(P_r)$ in the base frame, $\theta_r$ is the orientation. $v_{xr}$, $v_{yr}$ are the longitudinal lateral reference velocities at the point $(P_r)$ and $v_{\theta r}$ is the angular velocity. The reference velocities, as well as their first and second time derivatives, are assumed to be bounded for all $t$.

Denote by $e = \begin{pmatrix} \widetilde{x} & \widetilde{y} & L\widetilde{\theta} \end{pmatrix}^t$ the position tracking error vector expressed in the frame linked to the virtual vehicle, i.e.:

$$\begin{pmatrix} \widetilde{x}(t) \\ \widetilde{y}(t) \\ L\widetilde{\theta}(t) \end{pmatrix} = M^t(\theta_r) \begin{pmatrix} x_P(t) - x_{Pr}(t) \\ y_P(t) - y_{Pr}(t) \\ L\theta(t) - L\theta_r(t) \end{pmatrix} \tag{5}$$

By time derivation of (4), the full kinematic tracking model for the car-like can be easily developed (see these for detail):

$$\begin{cases} \dot{\widetilde{x}} = \frac{v_{\theta r}}{L}\widetilde{y} + c(\widetilde{\theta})\widetilde{v}_x + s(\widetilde{\theta})\widetilde{v}_y + p_{\widetilde{x}}(t,\widetilde{\theta})\widetilde{\theta}, \\ \dot{\widetilde{y}} = -\frac{v_{\theta r}}{L}\widetilde{x} + s(\widetilde{\theta})\widetilde{v}_x - c(\widetilde{\theta})\widetilde{v}_y + p_{\widetilde{y}}(t,\widetilde{\theta})\widetilde{\theta}, \\ L\dot{\widetilde{\theta}} = \widetilde{v}_\theta. \end{cases} \tag{6}$$

where

$$p_{\widetilde{x}}(t,\widetilde{\theta}) = \frac{(\cos(\widetilde{\theta})-1)}{\widetilde{\theta}}v_{xr} + \frac{\sin(\widetilde{\theta})}{\widetilde{\theta}}v_{yr},$$
$$p_{\widetilde{y}}(t,\widetilde{\theta}) = \frac{\sin(\widetilde{\theta})}{\widetilde{\theta}}v_{xr} + \frac{(1-\cos(\widetilde{\theta}))}{\widetilde{\theta}}v_{yr}.$$

Note that $p_{\widetilde{x}}(t,\widetilde{\theta})$ and $p_{\widetilde{y}}(t,\widetilde{\theta})$ are bounded functions. Sliding effects are treated as kinematic perturbations. As we can see, the model rely on the measurement of the parameter perturbations. Hence, its accurate measurement is difficult to obtain. To introduce the perturbations which can be present during the motion of the robot (Robucar), we have negligated the sliding of the rear wheels ($\mu_2 = \mu_4 = 0$) and take into account only the sliding of the front wheels ( $\epsilon\mu_1 \neq 0$, $\epsilon\mu_3 \neq 0$). Then, the system (3) becomes:

$$\begin{cases} v_x = v_1 + \varepsilon\mu_3 \\ v_y = \varepsilon\mu_1 \\ v_\theta = \tan(\phi)v_1 \end{cases} \tag{7}$$

In this situation, two solution are proposed here. The first one, by adding a term which represent the perturbations on the longitudinal velocity control $v_x$ (longitudinal sliding or slipping), and the second one, by modifying the reference trajectory that the vehicle must follow (lateral sliding or skidding). We introduce the longitudinal (lateral) perturbation in the expression of the velocity control (the reference trajectory) respectively, as follow:

$$\begin{cases} v_x = v_1 + \varepsilon\mu_3 \\ v_{yr} = -c(\widetilde{\theta}).\varepsilon\mu_1 \end{cases} \tag{8}$$

In the case when no sliding occurs, it is obvious that $v_x = v_1$ and $v_y = 0$.

The problem addressed in this paper is to find a state feedback controller which can guarantee the asymptotic stabilization of system (6) about the origin, i.e:

$$\lim_{t\to\infty}\widetilde{x}(t) = 0, \lim_{t\to\infty}\widetilde{y}(t) = 0, \lim_{t\to\infty}\widetilde{\theta}(t) = 0.$$

For this, a robust nonlinear control law based on second order sliding mode is derived in this paper. It will be assumed that only $x$, $y$, $\theta$ are available for measurement. it is shown that both the position and the angular tracking errors of the robot are asymptotically stabilized in an arbitrarily small neighborhood of the origin.

## III. Second order sliding mode controller

Sliding mode control approach exhibits relative simplicity of design and some robustness properties with respect to matching perturbations [11]. To overcome the well know chattering phenomenon while guaranteeing better convergence acuracy, second order sliding mode controllers have been proposed [12]. They are characterized by a discontinuous control acting on the second time derivatives of the sliding constraint s (instead of the first time derivative in classical sliding mode), whose vanishing defines the sliding manifold. The main principle of the second order sliding mode is to obtain a finite time convergence onto the second order sliding set $\{s = \dot{s} = 0\}$ (see, [14][13]).

Different kind of algorithms (twisting, drift , super twisting, sub-optimal...) able to ensure the finite time convergence have been proposed [13]. Here, the super twisting algorithm that only requires the knowlegde of the sliding surface will be adopted. In what follow, it is shown that both the position and the angular tracking errors of the robot are asymptotically stabilized in an arbitrarily small neighborhood of the origin.

### A. Design and attractivity of the sliding manifold

Let us define the sliding constraint $s = [s_1, s_2]^T$ as:

$$s_1 = \lambda_1\widetilde{x} + p_{\widetilde{x}}(t,\widetilde{\theta})\widetilde{\theta}, \tag{9}$$
$$s_2 = p_{\widetilde{y}}(t,\widetilde{\theta})\widetilde{\theta} + \lambda_2\widetilde{y}. \tag{10}$$

where $\lambda_1$, $\lambda_2$ are positive parameters. Note that the system has relative degree one with respect to both $s_1$ and $s_2$ and that the second time derivatives $s_1$and $s_2$ of are given by:

$$\ddot{s}_1 = \psi_1(\widetilde{x},\widetilde{y},\widetilde{\theta},t) + \dot{v}_1$$
$$\ddot{s}_2 = \psi_2(\widetilde{x},\widetilde{y},\widetilde{\theta},t) + \dot{v}_2$$

For sake of place, the expression of $\psi_1(\widetilde{x},\widetilde{y},\widetilde{\theta},t)$ and $\psi_2(\widetilde{x},\widetilde{y},\widetilde{\theta},t)$ are not reported here. The task is to generate a second order sliding mode on the second order sliding set given by the equalities: $s = \dot{s} = 0$. For this, assume that the reference velocities $(v_{1r}, v_{2r})$ and their first and second time derivatives are bounded and that the functions $\psi_i(\widetilde{x},\widetilde{y},\widetilde{\theta},t)$ are bounded such that $\left|\psi_i(\widetilde{x},\widetilde{y},\widetilde{\theta},t)\right| \leq K_i,\ i = 1,2$ where $K_i$ are positive constants. Then, it is known that one can apply the super twisting algorithm defined by the following control law [13]:

$$\begin{array}{l} v_i = -\lambda_{mi}\left|s_i\right|^{\frac{1}{2}} sign(s_i) + v_{1i}, \\ \dot{v}_{1i} = -W_i sign(s_i), \qquad i = 1,2 \end{array} \tag{11}$$

where $W_i$ and $\lambda_{mi}$ are positive constants that satisfy the following conditions:

$$\begin{cases} W_i > K_i \\ \lambda_{mi}^2 \geq 4K_i\frac{W_i+K_i}{W_i-K_i} \end{cases}$$

It can be shown that the control laws (11) generate a second order sliding mode on the second order sliding set $\{s = \dot{s} = 0\}$ (see [14], [13]). In particular this implies that:

$$\lambda_1\widetilde{x} = -p_{\widetilde{x}}(t,\widetilde{\theta})\widetilde{\theta} \tag{12}$$
$$\lambda_2\widetilde{y} = -p_{\widetilde{y}}(t,\widetilde{\theta})\widetilde{\theta} \tag{13}$$

### B. Asymptotic stability of the sliding motion

In order to show that, once in sliding mode, the posture errors of the robot are vanishing asymptotically, let us introduce the following candidate Lyapunov function:

$$V = \frac{1}{2}\left(\widetilde{x}^2 + \widetilde{y}^2\right).$$

The time derivative of $V$ along the trajectories of the system is given by:

$$\begin{aligned}\dot{V} &= \widetilde{x}\left(\frac{v_{\theta r}}{L}\widetilde{y} + c(\widetilde{\theta})\widetilde{v}_x + s(\widetilde{\theta})\widetilde{v}_y + p_{\widetilde{x}}(t,\widetilde{\theta})\widetilde{\theta}\right) \qquad (14)\\ &\quad +\widetilde{y}\left(-\frac{v_{\theta r}}{L}\widetilde{x} + s(\widetilde{\theta})\widetilde{v}_x - c(\widetilde{\theta})\widetilde{v}_y + p_{\widetilde{y}}(t,\widetilde{\theta})\widetilde{\theta}\right).\end{aligned}$$

Replacing the expressions (12) and (13) in (14), one gets:

$$\begin{aligned}\dot{V} &= -\lambda_1\widetilde{x}^2 - \lambda_2\widetilde{y}^2 + \widetilde{x}\left[c(\widetilde{\theta})\widetilde{v}_x + s(\widetilde{\theta})\widetilde{v}_y\right]\\ &\quad +\widetilde{y}\left[s(\widetilde{\theta})\widetilde{v}_x - c(\widetilde{\theta})\widetilde{v}_y\right]\end{aligned}$$

From the equations (12) and (13), we have:

$$\widetilde{y} = k(t,\widetilde{\theta})\widetilde{x}$$

with

$$k(t,\widetilde{\theta}) = \frac{\lambda_1 p_{\widetilde{y}}(t,\widetilde{\theta})}{\lambda_2 p_{\widetilde{x}}(t,\widetilde{\theta})}$$

Then one gets:

$$\begin{aligned}\dot{V} &= -\lambda_1\widetilde{x}^2 - \lambda_2\widetilde{y}^2 + \widetilde{x}\widetilde{v}_x\left[c(\widetilde{\theta}) + k(t,\widetilde{\theta})s(\widetilde{\theta})\right]\\ &\quad +\widetilde{x}\widetilde{v}_y\left[s(\widetilde{\theta}) - k(t,\widetilde{\theta})c(\widetilde{\theta})\right].\end{aligned}$$

As $p_{\widetilde{x}}(t,\widetilde{\theta})$ and $p_{\widetilde{y}}(t,\widetilde{\theta})$ are bounded for all $\widetilde{\theta}$, one can write:

$$\begin{aligned}\left|(c(\widetilde{\theta}) + k(t,\widetilde{\theta})s(\widetilde{\theta}))\right| &\leq \Gamma_1,\\ \left|(s(\widetilde{\theta}) - k(t,\widetilde{\theta})c(\widetilde{\theta}))\right| &\leq \Gamma_2.\end{aligned}$$

and suppose that

$$\begin{aligned}\|\widetilde{v}_x\| &\leq \Gamma_3 + \Gamma_4\|X\|,\\ \|\widetilde{v}_y\| &\leq \Gamma_5 + \Gamma_6\|X\|.\end{aligned}$$

where $\Gamma_1, \Gamma_2\ \Gamma_3, \Gamma_4, \Gamma_5, \Gamma_6$ are positive constants and $X = \begin{pmatrix}\widetilde{x} & \widetilde{y}\end{pmatrix}^t$.

Let us define $\lambda^\star = \min_{i=1,2}(\lambda_i)$, then one can write:

$$\dot{V} \leq -\lambda^\star\|X\|^2 + \|X\|^2\left(\frac{\Gamma_1\Gamma_3 + \Gamma_2\Gamma_5}{\|X\|} + \Gamma_1\Gamma_4 + \Gamma_2\Gamma_6\right)$$

Taking $\|X\| \geq \frac{\varepsilon}{2}$, this implies that

$$\dot{V} \leq -\lambda^* \|X\|^2 + \|X\|^2\left(\frac{2(\Gamma_1\Gamma_3 + \Gamma_2\Gamma_5)}{\varepsilon} + \Gamma_1\Gamma_4 + \Gamma_2\Gamma_6\right).$$

Define the ball $\mathcal{B}_{\varepsilon/2} = \left\{X : \|X\| \leq \frac{\varepsilon}{2}\right\}$. It results that outside the ball $\mathcal{B}_{\varepsilon/2}$, one has

$$\dot{V} \leq -\widetilde{\lambda}\|X\|^2 = -2\widetilde{\lambda}V$$

with

$$\widetilde{\lambda} = \lambda^\star - \left(\frac{2(\Gamma_1\Gamma_3 + \Gamma_2\Gamma_5)}{\varepsilon} + \Gamma_1\Gamma_4 + \Gamma_2\Gamma_6\right).$$

Thus $\dot{V}$ will be negative definite if the the parameter $\lambda^\star$ is chosen as:

$$\lambda^\star > \left(\frac{2(\Gamma_1\Gamma_3 + \Gamma_2\Gamma_5)}{\varepsilon} + \Gamma_1\Gamma_4 + \Gamma_2\Gamma_6\right)$$

Hence, with this choice of $\lambda^\star$ the solution of the system is given by

$$\|X(t)\| \leq \|X(0)\|\exp\left(-\widetilde{\lambda}t\right)$$

and there exists a finite time $t_1$ such that $\forall\, t > t_1 : X(t) \in \mathcal{B}_\varepsilon$. Thus, $\widetilde{x}$, $\widetilde{y}$ are stabilized in an arbitrarily small neighborhood of the origin. From the equations (12) and (13), it can be seen that the orientation angle error $\widetilde{\theta}$ is also stabilized in an arbitrarily small neighborhood of the origin.

## IV. Simulation results

In this section, simulation results on the kinematic model (6) using the proposed controller are presented. The sampling time is chosen to be $0.01s$ with the physical paramter $L = 1.2m$. The car-like WMR must follow a reference trajectory of a circular path in a time interval $T = 40s$. Simulation was performed testing the developed control both without and with sliding. The design of parameters of sliding surfaces and control gain are: $\lambda_1 = 1.5, \lambda_2 = 0.4, \lambda_{mi} = 1, W_i = 10^{-5}(i = 1, 2)$. Simulated tracking responses of the car-like WMR given in the base frame are reported relative to the two classes of simulations. Simulation results of the trajectory tracking of the car-like considering the ideal case (without sliding) are given in Figure 2 with initial condition: $\widetilde{x}(0) = 0.2m, \widetilde{y} = 0.1m, \widetilde{\theta}(0) = -0.1rd, v_{xr}(t) = 1m.s^{-1}, v_{\theta r}(t) = 0.03rd.s^{-1}$. The position tracking errors are given in the first line and the second line provides the orientation and steering angles errors, while the last line gives the input time error response (driven velocity) and the trajectories in the phase plane $(x, y)$. We remark that all the errors converge to the origin after a time equal to $10s$. Despite the initial errors of the car-like, this one joins the path made by the virtual vehicle and the control activity is accpetable and exhibits no chattering.

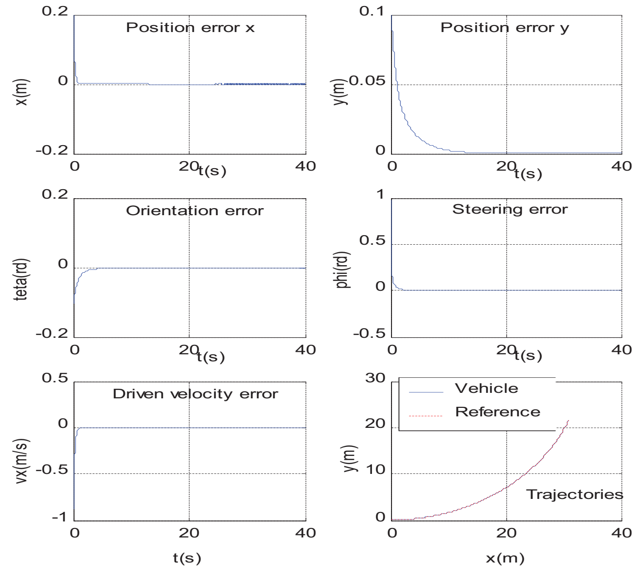


Figure 2. Tracking of the car-like without sliding.

As noted in the second section, two types of pertubations were introduced and applied to the car-like WMR at time $5s$. The first one is reported in Figure $3$, representing the vehicle slipping and the second one the vehicle skidding while curving, see Figure $4$. Simulation results for the trajectory tracking of the car-like considering only the slipping and both slipping and skidding are shown respectively in Figures $5, 6$ with null initial posture condition and $v_{xr}(t) = 0.6m.s^{-1}, v_{\theta r}(t) = 0.06rd.s^{-1}$. The Figures $5, 6$ shown that at the introduction of pertubations, the control inputs (driven velocity and steering angle) changes the values for converging to the desired reference one after the annulation of perturbations at time $t = 10s$. It can be seen that, the tracking of the car-like diverge quietly from the reference trajectory at time $t = 5s$ and its remains close to the reference trajectory after $t = 10s$. From these results, the robustness with respect to the sliding effects of the proposed sliding mode controller is confirmed.

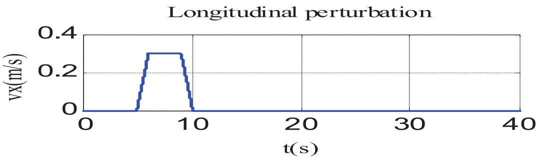


Figure 3. Perturbation: case of slipping.

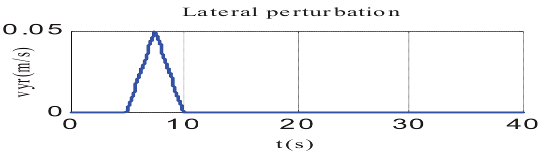


Figure 4. Perturbation: case of skidding.

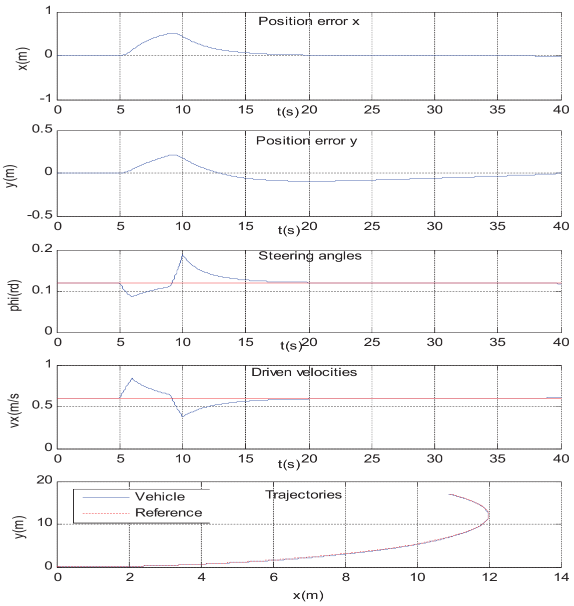


Figure 5. Tracking of the car-like with slipping.

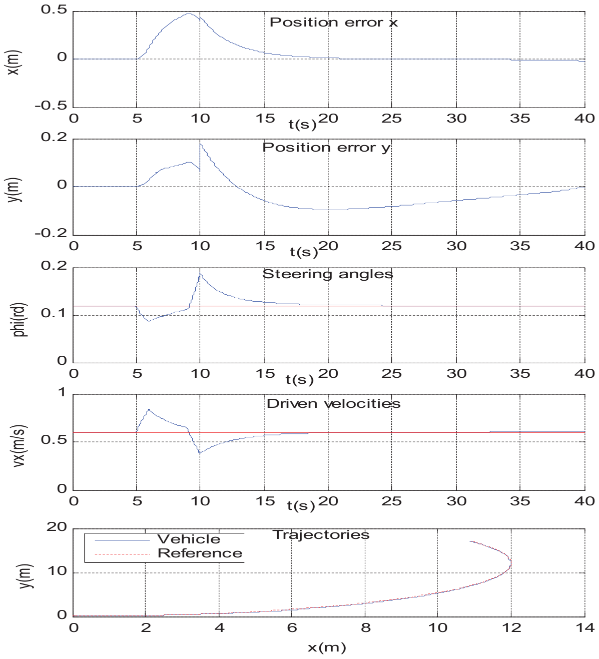


Figure 6. Tracking of the car-like with sliding.

## V. Conclusion

In this paper, we suggest higher order sliding mode control approach for solving the trajectory tracking problem for the car-like WMR considering sliding effects. Based on the tracking kinematic model of the car-like WMR, a second order sliding mode controller has been developed using the super twisting algorithm. The asymptotic convergence of the tracking errors has been proven by means of the Lyapunov function method. However, due to the nonlinearities, those errors are only stabilized in an arbitrarily small neighborhood of the origin. Simulation results are reported, showing the robustness of the proposed control law against nonlinearities and sliding effects. As a future work, it will be interesting to test the proposed tracking controller in experimentation and show how this approach works in real problems.

## References

[1] M. Ellouze and B. D'Andréa-Novel. Modeling and control of unicycle type-robot in presence of decoupled slipping and skidding effects. European Control Conference 1999.

[2] F. Hamerlain, K. Achour, T. Floquet, W. Perruquetti, Higher Order Sliding Mode Control of wheeled mobile robots in the presence of sliding effects, in Proceeding of the $44^{th}$ Conference on Decision and Control and European Control Conference, December 12-15, Spain 2005.

[3] W. Dong. Control of wheeled mobile robots with slipping. Proceeding of the $49^{th}$ Conference on Decision and Control and European Control Conference, December, 2010.

[4] H. Fang, R. Fan, B. Thuilot, P. Martinet. Trajectory tracking control of farm vehicles in presence of sliding. Journal of Robotics and Autonomous Systems, 2006.

[5] D. Wang, C. B. Low. Modeling and analysis of skidding and slipping in wheeled mobile robots: control design perpective. IEEE Transactions on Robotics, 2008, 24 (3): 676–687.

[6] I. Motte and G. Campion. A slow manifold approach  for the control of mobile robots not satisfying the kinematic constraints. IEEE Transactions on Robotics and Automation 2000, 16 (6): 875-880.

[7] W. Leroquais and B. D'Andréa-Novel. Modeling and control of wheeled mobile robots not satisfying ideal velocity constraints. Conference on Decision and Control 1996, 437–1442.

[8] B. D'Andréa-Novel, G. Campion, and G. Bastin. Control of wheeled mobile robots not satisfying ideal constraints: a singular perturbation approach. International Journal of Robust and Nonlinear Control 1995, (5): 243–267.

[9] M. Corradini, T. Leo and G. Orlando. Experimental testing of a discrete-time sliding mode control for trajectory tracking of a wheeled mobile robot in the presence of skidding effects. Journal of Robotic Systems 2002, 19 (4): 177–188.

[10] E. Mohammadpour, M. Naraghi.Tracking and regulation of wheeled mobile robot violating kinematic constraints. International Journal of Robotics and Automation 2010, 25 (4).

[11] V.I. Utkin. Sliding Modes in Control and Optimization. Springer Verlag, 1992.

[12] S.V. Emel'yanov, S.V. Korovin and L.V. Levantovsky. Higher Order Sliding Modes in the Binary Control System. Soviet Physics 1986, 31 (4): 291–293.

[13] A. Levant. Sliding order and sliding accuracy in sliding mode control. International Journal of Control 1993, 58 (6): 1247-1263.

[14] L. Fridman and A. Levant. Higher order sliding modes. In Sliding Mode Control in Engineering. W. Perruquetti and J. P. Barbot (Eds), Marcel Dekker, 2002, 53-101–618.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Predictive Control with Trajectory Planning in the Presence of Obstacles

Rubens Junqueira Magalhães Afonso, Roberto Kawakami Harrop Galvão and Karl Heinz Kienitz
Instituto Tecnológico de Aeronáutica, Divisão de Engenharia Eletrônica,
12228-900 São José dos Campos, SP, Brasil
Emails: rubensjm@ita.br, kawakami@ita.br, kienitz@ita.br

***Abstract*—In this work, a trajectory planning technique for an autonomous vehicle is proposed. A Predictive Control formulation is used both to plan a trajectory and control the vehicle in the presence of obstacles and dynamic constraints. However, some particularities of this sort of missions may make the time required for solution of the associated optimization problem prohibitive for a given sampling period. In this context, the possibility of using smaller prediction and control horizons is important to obtain a suitable control sequence within each sampling time. For this purpose, a trajectory planner which distributes waypoints along a previously established path is employed in the present paper. Each waypoint is determined so that it can be reached in a horizon which is smaller than the one necessary to reach the target set from the initial position, thus reducing the computational burden during the control phase. Moreover, during the planning phase the waypoints are chosen under the restriction that the target set should be reached within finite time so that the mission can be accomplished.**

***Index Terms*—Predictive control, trajectory planning, waypoint.**

## I. Introduction

Model-based Predictive Control (MPC) techniques involve the solution of an optimal control problem within a moving horizon, which is repeated (usually at every sampling time) on the basis of feedback from the sensors of the plant [1], [2]. One of the main advantages of MPC is the explicit treatment of constraints over the outputs and the controls of the plant. In aeronautical applications, it allows the various aircraft actuators to work closer to the limits of saturation, providing an increase in the flight envelope without compromising the safety of operation.

In typical problems of guidance of vehicles the state must reach a given set in finite time for the mission to be completed successfully [3]. Another issue is the existence of constraints that result in a non-convex optimization problem, such as the presence of obstacles which the vehicle must avoid.

In [3] the problem of reaching a terminal set in finite time is addressed by using a variable horizon MPC formulation and minimizing a weighted-time-fuel cost function. The resulting optimization problem is a *Mixed Integer Linear Programming* (MILP) one, because it involves both continuous and logical variables. A kinematic model of the vehicle is used to determine the trajectory, but vehicle guidance and control can be carried out by enhancing the model with rigid body and/or actuator dynamics. The obstacle avoidance constraints can also be encoded in the MILP formulation by using logical variables in conjunction with a "big-M" method, thus circumventing the difficulties brought about by the loss of convexity of the optimization problem.

However, the resulting MILP problem may not be feasible with small horizons, causing the need of larger ones in order to reach the terminal set from the initial state. In turn, larger horizons are associated with a higher number of optimization variables and may render the computational treatment of the optimization problem impracticable.

In this scenario, it may be convenient to split the mission into a series of intermediate goals that can be achieved within smaller horizons. Nevertheless, this division should be done judiciously, considering information about the limits of the actuators and constraints on the states of the plant, as well as the obstacles and the final goal of the mission. One way to accomplish the division is to introduce a sequence of waypoints, i. e., intermediate points through which the vehicle must pass to reach the terminal set. Such waypoints must be followed in a predetermined sequence obeying constraints and leading to the terminal set. Their determination must also take into account the capacity of achieving the waypoints from the current position of the vehicle within a horizon of acceptable size. Thus, the problem is divided into two steps: 1) off-line trajectory planning - involving the calculation of waypoints; 2) online execution of the planned trajectory. The first step is performed before the start of the maneuver.

In this work a technique for trajectory planning via waypoints in the presence of obstacles is proposed. The employment of such a technique along with an MPC-MILP formulation is evaluated regarding the computational burden involved in the control task.

The remainder of this paper is organized as follows. Section II reviews the MPC-MILP formulation adopted in the present work, which involves minimizing the weighted-time-fuel cost function to reach a given terminal set in the presence of obstacles [3]. Next, in Section III, trajectory planning for vehicles in the presence of obstacles is briefly discussed. The main contribution of this work is introduced in Section IV, in which the proposed approach for trajectory planning is presented. The scenarios adopted in the simulations are described in Section V. Section VI presents the simulation results of the proposed approach, which are compared to the direct application of the original MPC-MILP formulation. Finally, conclusions are drawn and suggestions for future work

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

are given in section VII.

### A. Notation

- $x \in \mathbb{R}^n$: plant state;
- $x_0 \in \mathbb{R}^n$: initial plant state;
- $u \in \mathbb{R}^p$: control signal;
- $r \in \mathbb{R}^2$: vehicle position;
- $b \in \{0,1\}$: binary variable associated to the horizon minimization;
- $b_{i,m}^{obs} \in \{0,1\}$: binary variables associated to the obstacle avoidance constraints;
- $k$: current time;
- $\hat{\Diamond}(k+i|k)$: predicted value of the variable $\Diamond$ at time $k+i$ based on the information available up to time $k$;
- $\Diamond^*$: optimal value of the variable $\Diamond$;
- $N(k) \in \mathbb{N}$: MPC control and prediction horizon;
- $C_r \in \mathbb{R}^{2\times n}$: matrix that extracts position information from the state vector;
- $\mathbb{U}(j) \subset \mathbb{R}^p$: set of admissible control values at time $j$;
- $\mathbb{X}(j) \subset \mathbb{R}^n$: set of admissible state values at time $j$;
- $\mathbb{Q}(N(k)+1) \subset \mathbb{R}^n$: set of terminal state values at the end of the horizon;
- $\mathcal{Z}_m \subset \mathbb{R}^2$: polygon defining the $m$-th obstacle;
- $V_i \in \mathbb{R}^2$: $i$-th vertex in the planned path;
- $\bar{N} \in \mathbb{N}$: maximal horizon in the one-step formulation;
- $\bar{N}_P \in \mathbb{N}$: maximal horizon between waypoints;
- $N_{WP} \in \mathbb{N}$: number of waypoints;
- $N_{obs} \in \mathbb{N}$: number of obstacles;
- $N_f \in \mathbb{N}$: number of sides in each obstacle;
- $N_V \in \mathbb{N}$: number of vertices in the planned path;
- $\alpha_i \in \mathbb{R}$: variable that determines the position of the $i$-th waypoint along the planned path;
- $M \in \mathbb{R}_+$: constant large enough to make terminal constraints inactive;
- $M_x \in \mathbb{R}_+$: constant large enough to make state constraints inactive;
- $M_u \in \mathbb{R}_+$: constant large enough to make control constraints inactive;
- $M_{obs} \in \mathbb{R}_+$: constant large enough to make obstacle avoidance constraints inactive;
- $M_{WP} \in \mathbb{R}_+$: constant large enough to make waypoint location constraints inactive;
- $r_x \in \mathbb{R}$: position along a coordinate axis in a horizontal plane regarding an arbitrary origin;
- $r_y \in \mathbb{R}$: position along a coordinate axis (perpendicular to the first) in a horizontal plane regarding an arbitrary origin;
- $v_x \in \mathbb{R}$: velocity regarding the $r_x$ position;
- $v_y \in \mathbb{R}$: velocity regarding the $r_y$ position;
- $a_x \in \mathbb{R}$: acceleration regarding the $v_x$ velocity;
- $a_y \in \mathbb{R}$: acceleration regarding the $v_y$ velocity;
- $\gamma \in \mathbb{R}$: weight of the term associated to the fuel consumption in the cost function;
- $\mathbf{1}_\Diamond \in \mathbb{R}^\Diamond$: column vector of $\Diamond$ elements equal to 1;
- $\|\Diamond\|_1$: 1-norm of the vector $\Diamond$.

## II. Predictive Control

As depicted in Fig. 1, the basic elements of a predictive controller operating in discrete time are:

- A model used to predict the state of the plant over a horizon of $N$ steps in the future, based on the current state $x(k)$ and the control sequence $\{\hat{u}(k+j|k)\},\ j = 0,\ldots,N-1$ to be applied.
- An algorithm to optimize the control sequence regarding the cost function specified for the problem and the existing constraints on inputs and states of the plant.

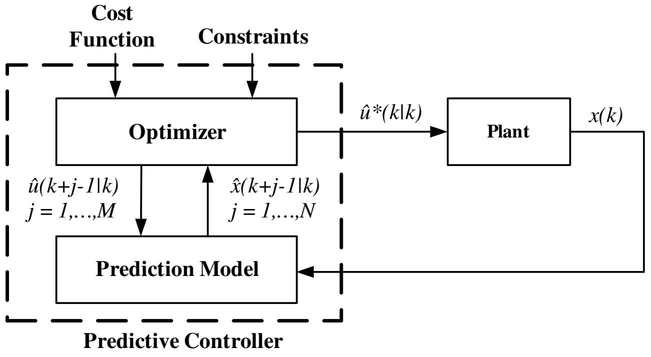


Fig. 1. Predictive control loop using state feedback.

In [3] the cost function is of the form:

$$J\left[\hat{x}(\cdot|k),\hat{u}(\cdot|k),N(k)\right] = \sum_{j=0}^{N(k)} \left(1+\gamma\left\|\hat{u}(k+j|k)\right\|_1\right),\ \ \gamma > 0 \tag{1}$$

subject to

$$\hat{x}(k+j|k) = \left\{ \begin{array}{l} x(k),\ j=0 \\ A\hat{x}(k+j-1|k) + B\hat{u}(k+j-1|k),\ j>0 \end{array} \right. \tag{2a}$$

$$\hat{x}(k+j|k) \in \mathbb{X}(j),\ \ j = 1,\ldots,N(k) \tag{2b}$$

$$\hat{u}(k+j|k) \in \mathbb{U}(j),\ \ j = 0,\ldots,N(k) \tag{2c}$$

$$\hat{x}(k+N(k)+1|k) \in \mathbb{Q}(N(k)+1) \tag{2d}$$

In the present work, robustness to unknown disturbances is not addressed in order to simplify the presentation of the main contribution, which will be stated in Section III. Therefore, the dependence of the sets $\mathbb{X}$, $\mathbb{U}$ and $\mathbb{Q}$ on $j$ and $N(k)$ is disregarded.

It can be seen from Eq. (1) that a compromise between the time to reach the terminal set and the fuel spent during the task is achieved by penalizing the time in the first term of the cost function and the fuel expense in the second. By manipulating the weight $\gamma$, the planner can be adjusted to put more emphasis in time minimization (small values of $\gamma$) or fuel expense minimization (large values of $\gamma$).

This cost is denoted simply by $J(k)$ to indicate that it is a function to be optimized at the sampling time $k$.

The optimal control sequence $\{\hat{u}^*(k+j|k), j = 0,\ldots,N(k)\}$ that minimizes the cost given by Eq. (1) subject to the constraints in Eqs. (2a), (2b), (2c) and (2d) usually cannot be analytically determined. Therefore, an optimization algorithm has to be used to obtain the control sequence subject to constraints. Customarily, an strategy known as "receding horizon" [2] is applied, i. e., only the first element of the control sequence is applied to the plant ($u(k) = \hat{u}^*(k|k)$) and the optimization is repeated at the next sampling time, making $u(k+1) = \hat{u}^*(k+1|k+1)$.

### *A. Horizon minimization*

If the terminal set is given in terms of linear constraints:

$$\begin{gathered} \mathbb{Q} = \{x : p_i^T x \le q_i,\ i = 1, \ldots, N_Q\}, \\ p_i \in \mathbb{R}^n,\ q_i \in \mathbb{R},\ i = 1, \ldots, N_Q \end{gathered} \tag{3}$$

then the terminal constraints can be rewritten as:

$$p_i^T \hat{x}(k+j+1|k) \le q_i + M[1 - b(j)],\ i = 1, \ldots, N_Q \tag{4}$$

with $b(j)$ defined as

$$b(j) = \begin{cases} 1, \text{ if } j = N(k), \\ 0, \text{ if } j \ne N(k) \end{cases} \tag{5}$$

The scalar $M$ must be taken so that $M > p_i^T x - q_i$, $\forall i$ for all admissible $x$ [4].

Thus, the cost can be recast in terms of a maximum preset value $\bar{N}$ for the horizon, that is

$$J(k) = \sum_{j=0}^{\bar{N}} \left(jb(j) + \gamma \left\|\hat{u}(k+j|k)\right\|_1\right) \tag{6}$$

subject to (4) with the following additional constraints:

$$\sum_{j=0}^{\bar{N}} b(j) = 1 \tag{7}$$

The cost expressed in Eq. (6) coincides with the one in Eq. (1) if the optimal value $N^*(k)$ for the horizon is less than or equal to $\bar{N}$ and the optimal control is null after $N^*(k)$, i. e., $\hat{u}^*(k+j|k) = 0,\ j > N^*(k)$. This last condition is guaranteed as the constraints in Eqs. (2b) and (2c) are imposed only up to the horizon $N^*(k)$. After this horizon, there is no constraint to be satisfied and thus the minimization of $\left\|\hat{u}(k+j|k)\right\|_1$ for $j > N^*(k)$ results in a zero control.

The state and control constraints up to the horizon $\bar{N}$ are rewritten in [3] using scalars large enough so that they become inactive after $N(k)$. Indeed, let the sets of admissible states and controls be

$$\begin{gathered} \mathbb{X} = \{x : r_{i,x}^T x \le q_i^x,\ i = 1, \ldots, N_x\}, \\ \mathbb{U} = \{u : r_{l,u}^T u \le q_l^u,\ l = 1, \ldots, N_u\}, \\ r_{i,x} \in \mathbb{R}^n,\ r_{l,u} \in \mathbb{R}^p,\ q_i^x,\ q_l^u \in \mathbb{R}, \\ i = 1, \ldots, N_x,\ l = 1, \ldots, N_u \end{gathered} \tag{8}$$

The constraints on the states and controls can then be rewritten as

$$\begin{aligned} & r_{i,x}^T \hat{x}(k+j|k) \le q_i^x + M_x \sum_{m=1}^{j-1} b(m),\ i = 1, \ldots, N_x \\ & r_{l,u}^T \hat{u}(k+j-1|k) \le q_l^u + M_u \sum_{m=1}^{j-1} b(m),\ l = 1, \ldots, N_u \end{aligned} \tag{9}$$

which makes the constraints inactive for $j > N(k)$ as $b(N(k)) = 1$. For this purpose, $M_x$ must be such that $M_x > r_{x,i}^T x - q_i^x$, $\forall i$, for all $x$ reachable in up to $\bar{N}$ steps from the terminal set with null control. $M_u$ is a scalar that renders the inequalities inactive for all admissible values of $u$.

Therefore the problem is defined with a fixed horizon $\bar{N}$ and a linear cost involving real and integer variables subject to linear constraints. Thus, algorithms for MILP can be used to obtain the optimal control sequence.

### *B. Obstacle avoidance*

Obstacles such as buildings, hills and dangerous areas to be avoided are commonly present in problems of vehicle guidance. The obstacle avoidance constraints lead to the loss of convexity of the optimization problem that has to be solved in order to calculate the control sequence. The present work adopts the formulation for the avoidance of obstacles presented in [5] and [3], which introduces a set of binary variables for each obstacle.

The constraint that the trajectory in space does not cross an obstacle can be written as $r = C_r x \notin \mathcal{Z}_m$, in which $\mathcal{Z}_m = \{r | P_m^{obs} r \le q_m^{obs}\}$. Without loss of generality, all obstacles will be assumed to have the same number of sides $N_f$. It is therefore required that the position $r$ is not in the sets $\mathcal{Z}_m$, $1 \le m \le N_{obs}$ at each sampling time, which is equivalent to imposing that the sets $\mathcal{I}_m = \{i \in \{1, \ldots, N_f\} : P_{i,m}^{obs} r > q_{i,m}^{obs}\} \ne \varnothing$, where $P_{i,m}$ is the $i$-th row of $P_m^{obs}$ and $q_{i,m}$, the $i$-th element of $q_m^{obs}$. To this end, binary variables can be used as follows:

$$\begin{aligned} & -P_{i,m}^{obs} r(k+j) \le -q_{i,m}^{obs} + M_{obs}[1 - b_{i,m}^{obs}(k+j)] - \epsilon \\ & \sum_{i=1}^{N_f} b_{i,m}^{obs}(k+j) \ge 1,\ b_{i,m}^{obs} \in \{0, 1\}, \\ & 1 \le j \le \bar{N},\ 1 \le m \le N_{obs} \end{aligned}$$

Thus, with a large enough scalar $M_{obs}$, when $b_{i,m}^{obs}(k+j) = 0$, the constraint becomes inactive. If $b_{i,m}^{obs}(k+j) = 1$, the constraint is effectively enforced. The condition $\sum_{i=1}^{N_f} b_{i,m}^{obs}(k+j) \ge 1$ requires that at least one of the constraints is active at every sampling time, ensuring that the position $r$ is "outside" the $m$-th obstacle. $\epsilon > 0$ is chosen arbitrarily small so that the inequality "$\le$" becomes "$<$", thus removing the border of the obstacle from the set of allowed positions.

## III. Trajectory planning Architecture

Path planning refers to the search for a curve in two or three dimensions connecting the starting point to the goal point or terminal set and avoiding obstacles. If the planning is successful, a set of positions that the vehicle must occupy to reach the destination is produced. However, the dynamic constraints of the vehicle are not taken into account. In contrast, the problem of trajectory planning includes dynamic constraints. So the result must also include a sequence of velocity vectors associated to the position of the vehicle.

In the context of aircraft guidance and control, the path planning involves a number of issues in addition to the avoidance of obstacles. Such issues include the presence of dynamic constraints, usually in the form of velocity and acceleration limits, the need for a feedback control strategy in real time in order to make the system robust to atmospheric disturbances, and constraints on the amount of fuel available to execute the maneuver. These factors contribute to increase the complexity of trajectory planning and control of aircraft in the presence of obstacles, often making the problem computationally intractable. This limits the application of established algorithms that have been developed in robotics and path planning for land vehicles [6].

Among the possible solutions for this problem, one that enjoys relative success divides the planning and control in hierarchical levels, counting often with layers that employ heuristics to reduce the computational load [6]. Thus, on the upper level, planning is carried out, in which dynamic constraints may be included and which may rely on some optimization criterion. Then, if the dynamic constraints have not yet been considered, it proceeds to a smoothing of the path to adapt to these constraints when possible and discarding it otherwise. The next step is the addition of time tags to the path, obtaining a trajectory. At this stage, one can employ some kind of optimality criterion to define the trajectory. Finally, this trajectory is used to generate references to the feedback controller. Also, criteria for an optimal control solution can be adopted. In this context, [6] presents a thorough review of the literature.

The technique proposed in the present paper involves three layers, namely:

1) A path planner which produces a path composed of the connection of successive straight-line segments connecting the initial position to the target set, while avoiding obstacles. Dynamics constraints are not considered in this layer.
2) A trajectory planner which determines waypoints along the planned path obtained from the first layer. The determination of the waypoints is done considering the dynamic constraints of the vehicle, the existence of obstacles, the arrival at the target set in finite time and the capacity to reach each waypoint from the previous one within a fixed small horizon.
3) A Predictive Control layer which employs the waypoints determined during the second phase as targets of predictive control problems with small horizon, until the last one is reached and the target set can be reached within the small horizon. In this phase, the dynamic and obstacle avoidance constraints are again enforced.

In the present work, the path planner is not addressed and a path which satisfies the conditions described above is assumed available. For the purpose of obtaining such a path, many techniques may be used, such as Voronoi graphs, probabilistic roadmaps, $A^*$ search [7], and RRTs (Rapidly-exploring Random Trees) [8]. It is further assumed that the path is provided in the form of a sequence of vertices connected by straight-line segments.

The trajectory planning and control architecture adopted herein is depicted in Fig. 2. The dashed lines mark the blocks addressed in the present work. The MPC controller was discussed in the previous section. The trajectory planner provides a list of target waypoints in the order that should be followed to reach the final target set. The "Active target selection logic" simply checks whether the position of the vehicle is equal to the waypoint (up to a certain numerical tolerance); if true, the active target is the next waypoint in the sequence; otherwise, the active target remains the same. After the last waypoint is reached, the logic commutes to the final target set. Since the "Trajectory planner" passes only the waypoints to the control loop, and not every position, velocity and control signal used to reach them, and since the controller cost function and the planner one can be different, the planned trajectory and the one that is actually followed may in general present differences.

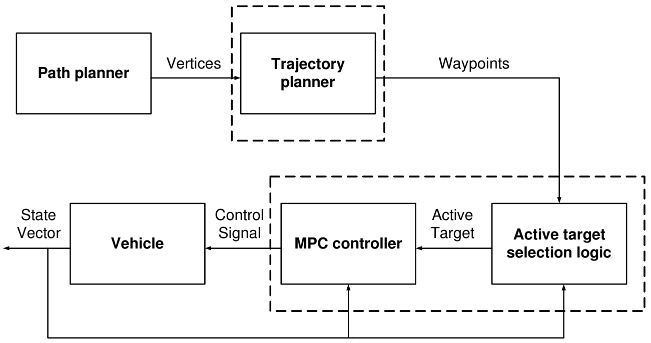


Fig. 2. Trajectory planning and control architecture used in this work.

## IV. Proposed trajectory planning technique

The technique proposed in this paper for trajectory planning involves the determination of a preset number of waypoints. These are scattered between the initial position of the vehicle and the terminal set. Their determination considers a horizon shorter than the one necessary to reach the terminal set from the initial state. Therefore, computational burden is expected to be lighter.

In order to limit the search space of solutions, the waypoints are constrained to a previously planned path, given in terms of the vertices of a sequence of straight-line segments which constitute a collision-free path from the initial position of the vehicle to the target set. This turns the problem of searching a two-dimensional space for a solution to a one-dimensional search over a piecewise linear curve.

In addition, the obstacle avoidance constraints are enforced in the optimal determination of the positions of the waypoints along the planned path in order to ensure the existence of collision-free trajectories between the waypoints. The problem of determining waypoint positions along the planned path composed of straight-line segments, avoiding obstacles and leading to the terminal set can be posed as follows:

*Problem 4.1:* Let $N_{WP}$, $\bar{N}_P$, $N_{obs}$, $N_f$, and $\{V_i\}$, $i = 1, \ldots, N_V + 1$ be the preset number of waypoints, the maximal horizon to reach an waypoint from the previous one, the number of obstacles, the number of sides of each obstacle, and the ordered sequence of vertices whose connection via straight-line segments produces the collision-free path ($V_1$ is the initial position of the vehicle), respectively. The waypoint determination problem is stated as

$$\min_{\hat{u}(k+j|k),\, \alpha_i,\, b_l^{obs},\, b_{i,j}^{WP}} \sum_{i=1}^{N_{WP}} \alpha_i + \mu \sum_{i=1}^{N_{WP}} \sum_{j=1}^{N_V} j b_{i,j}^{WP} \tag{10}$$

s.t.

$$\begin{aligned} &\hat{r}(k+i\bar{N}_P|k) \leq (\alpha_i - N_V + j)V_{j+1} + \\ &+ \left[1 - (\alpha_i - N_V + j)\right] V_j + M_{WP}(1 - b_{i,j}^{WP}), \\ &1 \leq i \leq N_{WP},\ 1 \leq j \leq N_V \end{aligned} \tag{11a}$$

$$\begin{aligned} &-\hat{r}(k+i\bar{N}_P|k) \leq -\left\{(\alpha_i - N_V + j)V_{j+1} + \right. \\ &\left. + \left[1 - (\alpha_i - N_V + j)\right] V_j + M_{WP}(1 - b_{i,j}^{WP})\right\}, \\ &1 \leq i \leq N_{WP},\ 1 \leq j \leq N_V \end{aligned} \tag{11b}$$

$$\sum_{j=1}^{N_V} (N_V - j) b_{i,j}^{WP} \leq \alpha_i,\ 1 \leq i \leq N_{WP} \tag{11c}$$

$$\begin{aligned} &\sum_{j=1}^{N_V} b_{i,j}^{WP} = 1,\ 1 \leq i \leq N_{WP} \\ &b_{i,j}^{WP} \in \{0,1\},\ 1 \leq i \leq N_{WP},\ 1 \leq j \leq N_V \end{aligned} \tag{11d}$$

$$0 \leq \alpha_{N_{WP}} \leq \alpha_{N_{WP}-1} \leq \cdots \leq \alpha_1 \leq N_V \tag{11e}$$

$$\hat{x}(k + (N_{WP}+1)\bar{N}_P|k) \in \mathbb{Q} \tag{11f}$$

$$\hat{u}(k+j|k) \in \mathbb{U},\ 0 \leq j \leq (N_{WP}+1)\bar{N}_P - 1 \tag{11g}$$

$$\hat{x}(k+j|k) \in \mathbb{X},\ 1 \leq j \leq (N_{WP}+1)\bar{N}_P - 1 \tag{11h}$$

$$P_m^{obs}\hat{x}(k+j|k) \leq -q_m^{obs} + M_{obs}(\mathbf{1}_{N_f} - b_m^{obs}(k+j|k)) \tag{11i}$$

$$\begin{aligned} &\sum_{l=1}^{N_f} b_{l,m}^{obs}(k+j|k) \geq 1,\ b_{l,m}^{obs}(k+j|k) \in \{0,1\}, \\ &1 \leq j \leq (N_{WP}+1)\bar{N}_P,\ 1 \leq m \leq N_{obs} \end{aligned} \tag{11j}$$

where $\mu > 0$ is a scalar, $\hat{r}(k+i\bar{N}_P|k) = C_r\hat{x}(k+i\bar{N}_P|k)$ is the predicted position at the sampling time $(k+i\bar{N}_P)$, and $P_m^{obs}$, $q_m^{obs}$, and $M_{obs}$ are defined as in section II-B.

The binary variables $b_{i,j}^{WP}$ are used to make the constraints in Eqs. (11a) and (11b) active or inactive. If $b_{i,j}^{WP} = 1$, then the inequalities (11a) and (11b) are active, which imposes an equality constraint restricting the position of the $i$-th waypoint to the straight-line segment between the $j$-th and $(j+1)$-th vertices. Otherwise, if $b_{i,j}^{WP} = 0$, the inequalities (11a) and (11b) are inactive for the particular values of $i$ and $j$, meaning that the $i$-th waypoint is not located in the straight-line segment between the $j$-th and $(j+1)$-th vertices. For this purpose, the scalar $M_{WP}$ is chosen large enough to render the constraints in Eqs. (11a) and (11b) inactive.

The inequalities in Eqs. (11a), (11b), (11c) and (11e) along with the equality in Eq. (11d) impose that the positions of the waypoints remain in one of the straight-line segments that compose the planned path. If $N_V - j \leq \alpha_i \leq N_V - j + 1$, then the $i$-th waypoint is located in the straight-line segment between vertices $V_j$ and $V_{j+1}$. For instance, if $\alpha_i = N_V - j$, then the $i$-th waypoint is exactly at the $j$-th vertex of the planned path.

The first term of the cost function in Eq. (10) aims at minimizing the values of $\alpha_i$, $1 \leq i \leq N_{WP}$, which prioritizes solutions that locate the waypoints farther from the initial position and closer to the terminal set, in order to avoid low initial speeds. The second term is introduced in order to obtain the maximal possible value to the term $(N_V - j)b_{i,j}^{WP}$, $1 \leq j \leq N_V$. As a consequence, the value of $(N_V - j)b_{i,j}^{WP}$ resulting from the minimization of this term subject to the constraint in Eq. (11c) is the greatest integer which is smaller or equal to $\alpha_i$ for any positive value of the scalar $\mu$. This, in turn, means that the term $(\alpha_i - N_V + j)$ in the constraints (11a) and (11b) is restricted to the set $[0, 1)$, thus resulting in a position between $V_j$ and $V_{j+1}$ for the $i$-th waypoint.

The resulting values of $\alpha_i$, $1 \leq i \leq N_{WP}$ correspond to the positions of the waypoints between the initial position and the last vertex, each farther from the initial position than the one before. The last waypoint is chosen so that it is possible to reach the terminal set $\mathbb{Q}$ from this position within the horizon $\bar{N}_P$.

An example is presented in Fig. 3, in which two waypoints were used with a horizon $\bar{N}_P = 10$ for the system dynamics that will be described in Section V. The planned path contains three vertices ($N_V = 2$, since the initial position is an additional vertex): $V_1 = [0\ \ 0]^T$ (initial position), $V_2 = [1.2\ \ 0.4]^T$ (intermediate vertex) and $V_3 = [1.6\ \ 1.5]^T$ (vertex in the border of the target set). The values for $\alpha_1$ and $\alpha_2$ were 1.583 and 0.385, respectively. This means that the first waypoint should be between the vertices $V_1$ and $V_2$ and the second, between $V_2$ and $V_3$, which can be corroborated by the resulting positions of the waypoints depicted in Fig. 3.

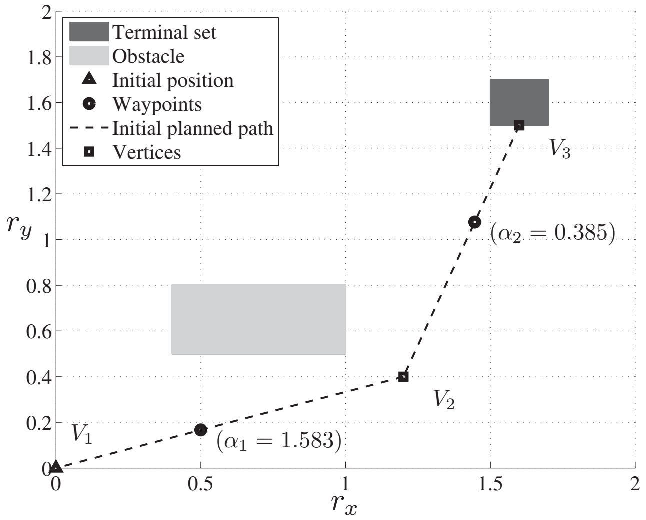


Fig. 3. Example of determination of the waypoints.

## V. Simulation scenarios

A kinematic model describing the movement of a vehicle in two dimensions was employed for simulation. The continuous-time model equations are:

$$\dot{r}_x = v_x,\ \dot{v}_x = a_x,\ \dot{r}_y = v_y,\ \dot{v}_y = a_y \tag{12}$$

where $r_x$ and $r_y$ define the position of the vehicle in a horizontal plane with respect to an arbitrary origin. This equation can be recast in state-space form ($\dot{x} = A_c x + B_c u$) by defining the state and control vectors as $x = [r_x \quad v_x \quad r_y \quad v_y]^T$, $u = [a_x \quad a_y]^T$. For use in the proposed MPC approach with trajectory planning, a discrete-time model of the form $x(k+1) = Ax(k) + Bu(k)$ was obtained with

$$A = \begin{bmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{bmatrix}, B = \begin{bmatrix} 0.5T^2 & 0 \\ T & 0 \\ 0 & 0.5T^2 \\ 0 & T \end{bmatrix} \tag{13}$$

in which $T$ is the sampling period. For the simulations in this paper $T$ was normalized to one time unit.

The dynamical constraints imposed on the velocities are $-1 \leq x_2,\, x_4 \leq 1$. As for the accelerations, $-1 \leq u_1,\, u_2 \leq 1$. Constraints $0 \leq x_1,\, x_3 \leq 2$ were also imposed on the position in order to limit it to the known terrain, over which information was assumed to be available.

The initial state of the vehicle was arbitrarily set to $x_0^T = [0\ 0\ 0\ 0]^T$, i. e., it started at rest. The goal was to reach a terminal set in the form of a rectangle described by the following inequalities on the positions $1.5 \leq x_1,\, x_3 \leq 1.7$.

As for the obstacles, they were also represented as rectangles with $0.6 \leq x_1^1, x_1^2 \leq 1$, $0.1 \leq x_3^1 \leq 0.8$, and $1.2 \leq x_3^1 \leq 1.6$, where the superscript refers to each of the two obstacles present in the simulations. It is worth noting that, since only the discrete-time predictions of the position are considered in the inequalities, this does not avoid stretches of the continuous-time trajectory crossing the obstacle. One alternative to handle this issue is proposed in [9], which involves incorporating restrictions on the transition of the vehicle to each region of the space defined by obstacle inequalities. However, it involves the introduction of more binary variables, increasing the complexity of the MILP problem. In this work, the length and width of the obstacle were expanded. To this end, an amount determined through the maximal admissible absolute value of the velocity in each axis was used to expand the borders of the obstacles. Therefore, the adopted avoidance constraints were constructed based on the following expanded obstacles: $0.5 \leq x_1^1, x_1^2 \leq 1.1$, $0 \leq x_3^1 \leq 0.9$, and $1.1 \leq x_3^1 \leq 1.7$.

The weight $\gamma$ of the fuel in the cost function was set to $0.1$. For the one-step solution, the maximal horizon was set to $\bar{N}_{OS} = 35$. Meanwhile, for the planner solution $\bar{N}_P = 8$ was adopted and the number of waypoints was set to $\bar{N}_{WP} = 3$. The computation times were taken as an average of 10 runs of each simulation, in order to eliminate fluctuations due to external factors. All simulations were carried out in a personal computer equipped with a Pentium$^{®}$ Dual-Core E5400 processor with $2.7GHz$ clock. For solution of the MILP, the CPLEX toolbox from IBM ILOG was used in Matlab environment, under an academic license.

## VI. Results and discussion

Initially, the simulation was carried out with the controller employing the one-step solution, i. e., trying to reach the terminal set from the beginning. The resulting path is presented in Fig. 4. The terminal set (dark gray rectangle) was reached successfully and the obstacles (light gray rectangles) were avoided. Moreover, as shown in Fig. 5, the accelerations $a_x$ and $a_y$ (which correspond to the controls $u_1$ and $u_2$, respectively) remained within the ±1 bounds. It took 24 sample periods to reach the terminal set from the starting position. The fuel cost was 28.4 and the average computation time was $17.72s$. The highest computational time was $4.22s$ and the mean computational time was $0.74s$. It can also be noted that a stretch of the continuous-time path crosses the prohibited region (black rectangle), but not the original obstacle. This justifies the choice to expand the original obstacle as means to avoid collisions.

The second simulation introduces the waypoint guidance using a previously planned path. An arbitrary path that connects the initial position to the terminal set while avoiding obstacles was employed for illustration, as shown in Fig. 6. The terminal set was reached successfully and the obstacles were avoided. Again, as shown in Fig. 7, the accelerations $a_x$ and $a_y$ remained within the ±1 bounds. It took 32 sample periods from the initial position to the terminal set. The planning phase lasted $0.33s$. The maximal computational time in the control phase was $0.18s$ and the mean was $0.08s$. Therefore, the total time to plan and execute the trajectory was about $3.03s$, which is much smaller than the time required by the one-step planner. The fuel cost was 52.57, which is larger than the one obtained with the one-step solution. In fact, since the waypoints are restricted to the previously planned

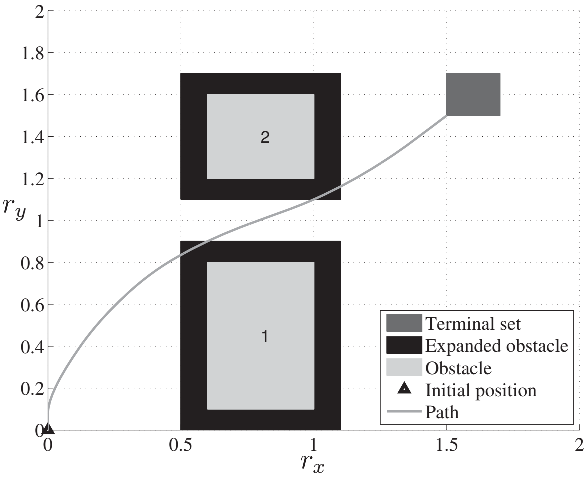


Fig. 4. Path obtained with one-step solution.

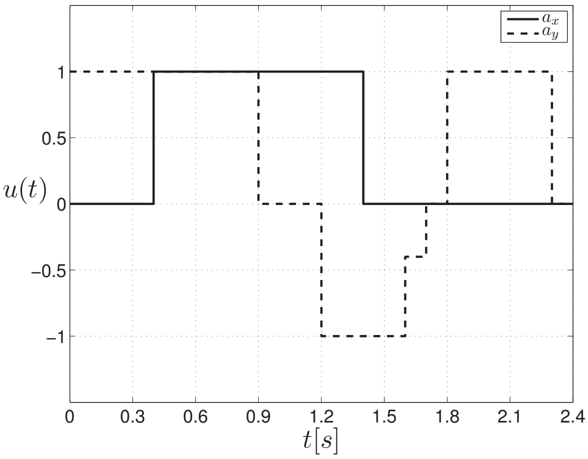


Fig. 5. Control signal obtained with one-step solution.

path that avoids the obstacles, and the horizon is shorter, the minimization of the fuel expense is compromised. It is interesting to note that the path does not cross the prohibited region, due to the fact that the waypoints steer the trajectory to the planned path, which is distant from the expanded obstacles. If the path planner emphasizes a safe path instead of the one that demands less fuel or the shortest one, it is likely to obtain a safer trajectory at the cost of a larger fuel consumption.

## VII. Conclusions

The proposed trajectory planning approach conduced to a feasible trajectory and reduced considerably the computational burden. The results showed also that a compromise between optimality and computation time can be inferred. This is suggested by the greater fuel cost and maneuver time obtained with the proposed approach as compared to the one-step MILP approach.

Future works could include a formulation which provides robustness to an unknown but limited disturbance in the trajectory planning phase, as was done in [3] for the MPC MILP formulation.

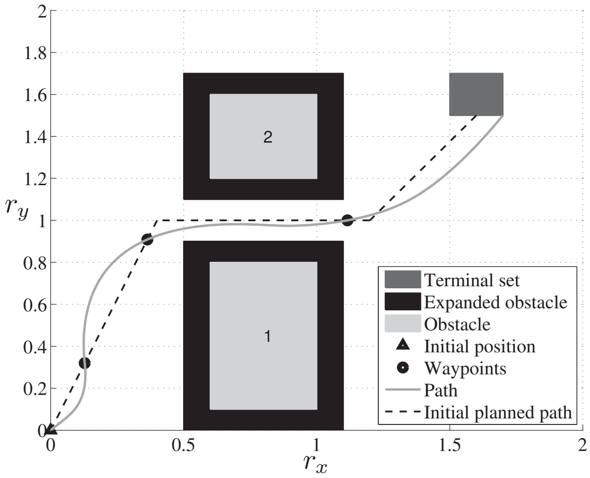


Fig. 6. Path obtained with trajectory planning.

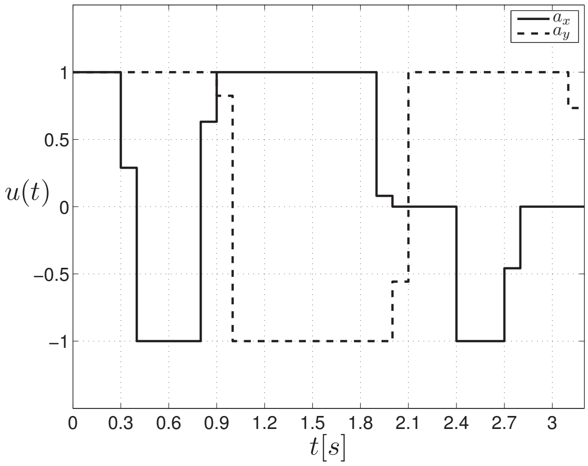


Fig. 7. Control signal obtained with trajectory planning.

## Acknowledgments

The authors acknowledge the support of FAPESP (2011/18632-8, 2011/17610-0) and CNPq (research fellowships).

## References

[1] E. F. Camacho and C. Bordons, *Model Predictive Control*. London: Springer-Verlag, 1999.

[2] J. A. Rossiter, *Model-based Predictive Control: a practical approach*. Boca Raton: CRC Press, 2003.

[3] A. Richards and J. P. How, "Robust variable horizon model predictive control for vehicle maneuvering," *Int. J. Robust and Non-linear Control*, vol. 16, no. 7, pp. 333 – 351, 2006.

[4] A. Bemporad and M. Morari, "Control of systems intregrating logic, dynamics, and constraints," *Automatica*, vol. 35, no. 3, pp. 407 – 427, 1999.

[5] A. Richards, T. Schouwenaars, J. P. How, and E. Feron, "Spacecraft trajectory planning with avoidance constraints using mixed-integer linear programming," *J. Guid. Control Dyn.*, vol. 25, no. 4, pp. 755 – 764, 2002.

[6] C. Goerzen, Z. Kong, and B. Mettler, "A survey of motion planning algorithms from the perspective of autonomous UAV guidance," *J. Intell. and Robotic Systems*, vol. 57, no. 1, pp. 65 – 100, 2010.

[7] S. Russel and P. Norvig, *Artificial Intelligence: a modern approach*, 3rd ed. Prentice Hall, 2009.

[8] S. LaValle, "Motion planning: the essentials," *IEEE Robotics and Automation Society Magazine*, vol. 18, no. 1, pp. 79 – 89, 2011.

[9] M. H. Maia and R. K. H. Galvão, "On the use of mixed-integer linear programming for predictive control with avoidance constraints," *Int. J. Robust and Non-linear Control*, vol. 19, pp. 822 – 828, 2009.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Multitask Trajectory Planning based on Predictive Control

Rubens Junqueira Magalhães Afonso, Roberto Kawakami Harrop Galvão and Karl Heinz Kienitz
Instituto Tecnológico de Aeronáutica, Divisão de Engenharia Eletrônica,
12228-900 São José dos Campos, SP, Brasil
Emails: rubensjm@ita.br, kawakami@ita.br, kienitz@ita.br

***Abstract*—In this work, a Predictive Control formulation for trajectory planning with multiple target sets is proposed, which solves the problem of performing all tasks in finite time via minimization of a weighted-time-fuel cost function, generating a feasible trajectory. An approach involving a procedure to order the list of the target sets to be visited in terms of the distance between them is used for comparison and it is shown that the proposed technique outperforms this approach in terms of time and fuel spent to accomplish the mission.**

***Index Terms*—Predictive control, trajectory planning, target set.**

## I. Introduction

In the context of aircraft guidance and control, the path planning problem becomes more complex than the simple search for a curve that connects the starting point to the goal while avoiding obstacles. Some of the reasons for that include the presence of dynamic constraints, usually in the form of velocity and acceleration limits, the need for a feedback control strategy in real time in order to make the system robust to atmospheric disturbances, and constraints on the amount of fuel available to execute the maneuver [1].

Model-based Predictive Control (MPC) techniques have been increasingly employed in the aeronautical industry [2] due to their ability to handle constraints on inputs and states of the plant [3]. More recently, MPC formulations as the one proposed in [4] have been used to perform trajectory planning for autonomous vehicles. Many particularities of the trajectory planning problem have been addressed in [4], namely the task of reaching a terminal set in finite time, while avoiding obstacles. Through the introduction of a variable horizon, it is possible to calculate the smallest horizon needed to reach the terminal set by solving a *Mixed Integer Linear Programming* (MILP) problem. Thus, a minimum time trajectory between a source point and a target set can be determined by using a kinematic model of the vehicle. Moreover, by enhancing the model with rigid body dynamics and characteristics of actuators, this problem can be extended to a more elaborate guidance and vehicle control framework. The MILP formulation also circumvents the difficulties brought about by the introduction of obstacles, particularly the loss of convexity of the set of admissible solutions.

However, a mission may require that the vehicle visits a number of sets. In this scenario, it may be convenient to consider all sets in the trajectory planning in order to minimize the total time to carry out all the tasks (visit all the target sets), instead of setting a single target set to be reached at every step.

In the present work a formulation to solve the trajectory planning problem with multiple target sets (termed Multitask Trajectory Planning) in the presence of obstacles is proposed. Simulation results with a fictional vehicle are presented in order to illustrate the success of the proposed technique. For comparison, the formulation presented in [4] with a single terminal target set was used with a list of sets to be visited, which was updated upon reaching each one of them. This list was ordered according to a criterion based on minimal distance between the initial position of the vehicle and the first target set as well as between the current target set and the next one.

The remainder of this paper is organized as follows. In Section II, the Predictive Control formulation adopted in the present work and proposed in [4] is presented, which involves minimizing a weighted-time-fuel cost function that penalizes the time to reach a given terminal set in the presence of obstacles. The approach for trajectory planning with multiple target sets is proposed later in Section III. Section IV presents the set ordering technique used for comparison. Simulation results of the application of both approaches are presented and discussed in Section V. Finally, conclusions are drawn and suggestions for future work are given in Section VI.

## II. MPC Formulation

The formulation employed in this work is similar to the one adopted in [4], with exception of the contribution related to the inclusion of multiple target sets, which will be introduced in Section III. The problem is recast in a *Mixed-Integer Linear Programming* (MILP) form much in the same way as in [4], again with exception of the inclusion of multiple target sets in Section III.

Figure 1 presents the basic elements of a predictive controller operating in discrete time, namely:

- A model used to predict the state of the plant over a horizon of $N$ steps in the future, based on the current state $x(k)$ and the control sequence $\{\hat{u}(k+j|k)\},\ j = 0, \ldots, N-1$ to be applied.
- An algorithm to optimize the control sequence regarding the cost function specified for the problem and the existing constraints on inputs and states of the plant.

The notation used is as follows: $u \in \mathbb{R}^p$ and $x \in \mathbb{R}^n$ denote the input and state variables of the plant, respectively. $\hat{\Diamond}(k +$

$j|k)$ denotes the predicted value of variable $\diamondsuit$ at time $k+j$ ($j \geq 1$) based on information available up to time $k$. The optimal control to be applied to the plant at time $k$ is denoted by $\hat{u}^*(k|k)$.

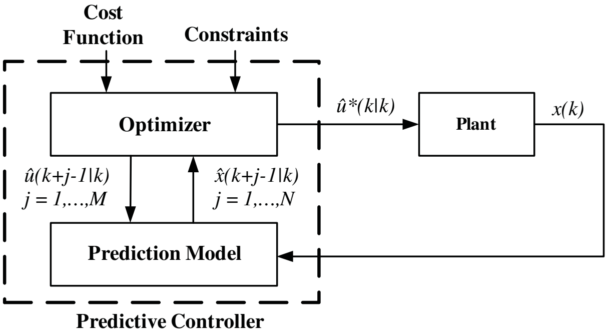


Fig. 1. Predictive control loop using state feedback.

In [4] the cost function is of the form:

$$J\left[\hat{x}(\cdot|k),\hat{u}(\cdot|k),N(k)\right] = \sum_{j=0}^{N(k)} \left(1 + \gamma \left\|\hat{u}(k+j|k)\right\|_1\right), \ \ \gamma > 0 \tag{1}$$

subject to

$$\hat{x}(k+j|k) = \begin{cases} x(k), \ j = 0 \\ A\hat{x}(k+j-1|k) + B\hat{u}(k+j-1|k), \ j > 0 \end{cases} \tag{2a}$$

$$\hat{x}(k+j|k) \in \mathbb{X}(j), \ \ j = 1, \ldots, N(k) \tag{2b}$$

$$\hat{u}(k+j|k) \in \mathbb{U}(j), \ \ j = 0, \ldots, N(k) \tag{2c}$$

$$\hat{x}(k+N(k)+1|k) \in \mathbb{Q}(N(k)+1) \tag{2d}$$

where $\mathbb{U}(j)$ and $\mathbb{X}(j)$ are the sets of admissible controls and states, respectively, and $\mathbb{Q}(N(k)+1)$ is the terminal set. In [4] the dependence of the sets $\mathbb{X}$, $\mathbb{U}$ and $\mathbb{Q}$ in terms of $j$ and $N(k)$ is inserted in order to provide robustness to an unknown but limited disturbance input. If this disturbance is disregarded, the sets are independent of time.

The first term of the cost function in Eq. (1) penalizes the time necessary to reach the target set. The second term, involving the norm of the control vector at each sampling time, penalizes the fuel spent. Thus, such a formulation provides a compromise between minimizing the time to achieve the goal and the fuel spent, which can be adjusted by the weight $\gamma$.

With a small abuse of notation, this cost is usually denoted simply by $J(k)$ in order to clarify that it is a function to be optimized at the $k$-th sampling time.

Optimization algorithms should be employed to obtain the optimal control sequence $\{\hat{u}^*(k+j|k), j = 0, \ldots, N(k)\}$ that minimizes the cost given by Eq. (1) subject to the constraints of Eqs. (2a), (2b), (2c) and (2d). The first element of such a sequence is applied to the plant (i. e., $u(k) = \hat{u}^*(k|k)$) and the optimization is repeated at the next sampling time, making $u(k{+}1) = \hat{u}^*(k{+}1|k{+}1)$. This strategy is known as "receding horizon" [5].

### A. Horizon minimization

If the terminal set is given in terms of linear inequalities

$$\begin{array}{c} \mathbb{Q} = \{x : p_i^T x \leq q_i, \, i = 1, \ldots, N_Q\}, \\ p_i \in \mathbb{R}^n, \, q_i \in \mathbb{R}, \, i = 1, \ldots, N_Q \end{array} \tag{3}$$

then the terminal constraints can be rewritten as:

$$p_i^T \hat{x}(k+j+1|k) \leq q_i + M[1 - b(j)], \, i = 1, \ldots, N_Q \tag{4}$$

where $M > 0$ is a constant and $b(j)$ is a binary decision variable defined as

$$b(j) = \begin{cases} 1, \text{ if } j = N(k), \\ 0, \text{ if } j \neq N(k) \end{cases} \tag{5}$$

The scalar $M$ must be taken so that $M > p_i^T x - q_i$, $\forall i$ for all admissible $x$ [6].

Thus, the cost can be recast in terms of a maximum preset value $\bar{N}$ for the horizon, that is

$$J(k) = \sum_{j=0}^{\bar{N}} \left(jb(j) + \gamma \left\|\hat{u}(k+j|k)\right\|_1\right) \tag{6}$$

subject to additional constraints

$$\sum_{j=0}^{\bar{N}} b(j) = 1 \tag{7}$$

The cost expressed in Eq. (6) coincides with the one in Eq. (1) if the optimal value $N^*(k)$ for the horizon is less than or equal to $\bar{N}$ and the optimal control is null after $N^*(k)$, i. e., $\hat{u}^*(k+j|k) = 0, \, j > N^*(k)$. This last condition is guaranteed as the constraints of equations (2b) and (2c) are imposed only up to the horizon $N^*(k)$. After this horizon, there is no constraint to be satisfied and thus the minimization of $\left\|\hat{u}(k+j|k)\right\|_1$ for $j > N^*(k)$ results in a zero control.

In [4] the state and control constraints are also rewritten up to the horizon $\bar{N}$ using scalars large enough so that they become inactive after $N(k)$. Indeed, let the sets of admissible states and controls be

$$\begin{array}{c} \mathbb{X} = \{x : r_{i,x}^T x \leq q_i^x, \, i = 1, \ldots, N_x\}, \\ \mathbb{U} = \{u : r_{l,u}^T u \leq q_l^u, \, l = 1, \ldots, N_u\}, \\ r_{i,x} \in \mathbb{R}^n, \, r_{l,u} \in \mathbb{R}^p, \, q_i^x, \, q_l^u \in \mathbb{R}, \\ i = 1, \ldots, N_x, \, l = 1, \ldots, N_u \end{array} \tag{8}$$

The constraints on the states and controls can then be rewritten as

$$\begin{array}{c} r_{i,x}^T \hat{x}(k+j|k) \leq q_i^x + M_x \displaystyle\sum_{m=1}^{j-1} b(m), \, i = 1, \ldots, N_x \\ r_{l,u}^T \hat{u}(k+j-1|k) \leq q_l^u + M_u \displaystyle\sum_{m=1}^{j-1} b(m), \, l = 1, \ldots, N_u \end{array} \tag{9}$$

which makes the constraints inactive for $j > N(k)$ as $b(N(k)) = 1$. $M_x \in \mathbb{R}$ must be such that $M_x > r_{x,i}^T x - q_i^x$, $\forall i$, for all $x$ reachable in up to $\bar{N}$ steps from the terminal set with null control. $M_u > 0$ is a scalar large enough to render the control constraints inactive for all admissible values of $u$.

Therefore the problem is defined with a fixed horizon $\bar{N}$ and a linear cost involving real and integer variables subject to linear constraints. Thus, algorithms for MILP can be used to obtain the optimal control sequence.

### *B. Obstacle avoidance*

In problems involving the guidance of vehicles, the presence of obstacles is usual, such as buildings, hills, dangerous areas to avoid, among others. In the presence of such obstacles the set of admissible states will no longer be convex. In [7], a form of dealing with polygonal obstacles through the use of MILP was proposed.

The constraint that the trajectory in space does not cross the obstacle can be written as $r = C_r x \notin \mathcal{Z}$, in which matrix $C_r$ extracts the position information from the state vector and $\mathcal{Z} = \{r | P^o r \le q^o\}$ defines an obstacle in the form of a polygon with $N_f$ sides. It is therefore required that the position $r$ is not in the set $\mathcal{Z}$ at each sampling time, which is equivalent to imposing that the set $\mathcal{I} = \{i \in \{1, \ldots, N_f\} : P_i^o r > q_i^o\} \neq \varnothing$ where $P_i$ is the $i$-th row of $P^o$ and $q_i$, the $i$-th element of $q^o$. To this end, binary variables can be used as follows:

$$\begin{gathered} -P_i^o r(k+j) \le -q_i^o + M_o[1 - b_i^o(k+j)] - \epsilon \\ \sum_{i=1}^{N_f} b_i^o(k+j) \ge 1,\ b_i^o \in \{0,1\},\ 1 \le j \le N \end{gathered}$$

Thus, when $b_i^o(k+j) = 1$, the constraint is effectively enforced. If $b_i^o(k+j) = 0$, with a large enough scalar $M_o > 0$, the constraint becomes inactive. The condition $\sum_{i=1}^{N_f} b_i^o(k+j) \ge 1$ requires that at least one of the constraints is active at every sampling time, ensuring that the position $r$ is "outside" the obstacle. $\epsilon > 0$ is chosen arbitrarily small so that the inequality "$\le$" becomes "$<$", thus removing the border of the obstacle from the set of allowed positions.

## III. Proposed Multitask Trajectory Planning Technique

The contribution of the present paper is the proposition of a novel method to enhance the capability of the trajectory planning technique proposed in [4] to visit $N_{ts} \ge 1$ target sets $\mathbb{Q}_1, \mathbb{Q}_2, \ldots \mathbb{Q}_{N_{ts}}$. This inclusion is in accordance with real-world mission demands, which usually require that the autonomous vehicle visits more than one target. In this context, one alternative is to arrange all target sets in a sequence to be visited. This arrangement may be performed by considering some optimization criteria, such as minimal distance from the starting position of the vehicle to define the first set and then, minimal distance between sets, choosing the next set to be visited as the closest to the last visited one, until all sets have been included in the sequence. With this sequence at hand, one can divide the problem in $N_{ts}$ single-target trajectory planning problems and solve each of them using the framework proposed in [4]. Upon reaching the target set for the current problem, it is replaced with the subsequent one in the pre-established order. This procedure is repeated until the final target set is reached. However, this may not yield the minimum time or minimum fuel solution, since the criteria employed to order the list of target sets do not consider the dynamics of the vehicle and constraints over the variables. Therefore, a framework which is capable of considering multiple target sets within the solution of the Predictive Control trajectory planning problem may bring about interesting results regarding the optimal solution to the trajectory planning problem.

It is assumed that the $N_{ts}$ target sets are defined as in Eq. (3), in terms of linear inequalities:

$$\begin{gathered} \mathbb{Q}_h = \{x : p_{h,i}^T x \le q_{h,i},\ i = 1, \ldots, N_{Q_h}\}, \\ p_{h,i} \in \mathbb{R}^n,\ q_{h,i} \in \mathbb{R},\ i = 1, \ldots, N_{Q_h}, \\ h = 1, \ldots, N_{ts} \end{gathered} \tag{10}$$

in which $\mathbb{Q}_h$ is the $h$-th target set and $N_{Q_h}$ is the number of inequalities used to describe it.

Since the vehicle only needs to visit each target set at one sample time, the constraints can be rewritten as:

$$\begin{gathered} p_{h,i}^T \hat{x}(k+j+1|k) \le q_{h,i} + M[1 - b_h(j)], \\ i = 1, \ldots, N_{Q_h},\ h = 1, \ldots, N_{ts}. \end{gathered} \tag{11}$$

where $M > 0$ is a constant and $b_h(j)$ is a binary decision variable defined as

$$b_h(j) = \begin{cases} 1, \text{ if } j = N_h, \\ 0, \text{ if } j \neq N_h \end{cases} \tag{12}$$

in which $N_h$ is the number of sample times that the vehicle takes to reach the $h$-th target set from the current position. For example, if the vehicle takes 2 sample times to reach the first target set from the starting position and 3 more to go from the first target set to the second, then $N_1 = 2$ and $N_2 = 5$.

An auxiliary variable has to be introduced in order to consider the minimization of the horizon to visit all of the target sets. This variable in be defined as:

$$N_f = \max(N_h),\ h = 1, \ldots, N_{ts} \tag{13}$$

Through the introduction of a new sequence of binary variables $\{b_f(j),\ j = 1, \ldots, \bar{N}\}$, it is possible to penalize this horizon $N_f$ to visit all of the target sets by rewriting the cost function in Eq. (6) as:

$$J(k) = \sum_{j=0}^{\bar{N}} \left( j b_f(j) + \gamma \left\| \hat{u}(k+j|k) \right\|_1 \right) \tag{14}$$

with

$$b_f(j) = \begin{cases} 1, \text{ if } j = N_f, \\ 0, \text{ if } j \neq N_f \end{cases} \tag{15}$$

subject to additional constraints

$$\sum_{j=0}^{\bar{N}} b_h(j) = 1,\ h = 1, \ldots, N_{ts} \tag{16}$$

$$\sum_{j=0}^{\bar{N}} b_f(j) = 1, \tag{17}$$

$$b_f(j) \le \frac{\sum_{h=1}^{N_{ts}} \sum_{i=0}^{j} b_h(i)}{N_{ts}} \tag{18}$$

The constraint in Eq. (16) ensures that each target set is visited at least once, whereas the ones in Eqs. (17) and (18) make sure that the definitions of Eqs. (13) and (15) are applied.

The only step left is to remove the state and control constraints after $N_f$. This is performed by replacing $b(m)$ in Eq. (9) by $b_f(m)$:

$$\begin{aligned} r_{i,x}^T \hat{x}(k+j|k) &\leq q_i^x + M_x \sum_{m=1}^{j-1} b_f(m),\ i = 1,\ldots,N_x \\ r_{l,u}^T \hat{u}(k+j-1|k) &\leq q_l^u + M_u \sum_{m=1}^{j-1} b_f(m),\ l = 1,\ldots,N_u \end{aligned} \tag{19}$$

The formulation presented in this section allows for planning a trajectory which visits all the sets while minimizing the cost function penalizing the overall time taken to perform the mission and the amount of fuel spent. However, when receding horizon feedback control has to be implemented, it is important to have a logic which removes a set that has already been visited, in order to avoid visiting the same sets over and over. The algorithm described bellow is proposed to circumvent this issue. This algorithm has to be run at every sampling time before calculating the control sequence:

**Algorithm 1:** Remove a set from the list after visiting it

```
for $i = 1 \rightarrow N_{ts}$ do
    if $x(k) \in SetList(i)$ then
        $SetList \leftarrow SetList \backslash SetList(i)$
        $N_{ts} \leftarrow N_{ts} - 1$
    end if
end for
```

□

in which the list of sets to be visited is given by $SetList$. The $\diamondsuit\backslash\spadesuit$ operator removes a set $\spadesuit$ from a set list $\diamondsuit$. After removing a set from the list, the number of binary variables can be reduced, since there are less sets to visit, which in turn means less constraints to be imposed.

## IV. Minimum-distance Trajectory Planning Algorithm

In this Section a trajectory planning algorithm which divides the mission in several tasks is presented. This algorithm will be used for comparison with the technique proposed in Section III. It is basically composed of two parts:

1) Order the list of sets to visit according to the distance, i. e., the first set to be visited is the closest to the initial position, the second is the closest to the first, and so on.
2) Apply the formulation presented in [4] with the terminal set as the next to be visited in the ordered list, starting by the first and changing to the next subsequently after the current target is reached.

In the following, the first part will be detailed, as it is the most cumbersome one, since the second involves only a test of pertinence of a point to a set and an update to the list of target sets.

The first part can be divided into two main algorithms, requiring only a list of the $N_{ts}$ target sets ($SetList$) in arbitrary order. In the first algorithm (Algorithm 2), $x_0$ is the initial state and $C_r$, a matrix that extracts position information from the state vector, thus $C_r x_0$ is the initial position of the vehicle. $dist(\alpha,\Omega)$ is a function that returns the minimal distance between a point $\alpha$ and a set $\Omega$ (employing 2-norm):

$$\begin{aligned} dist(\alpha,\Omega) = \quad & \min\nolimits_{\alpha,\beta} \|\alpha - \beta\|_2 \\ & s.t.\ \beta \in \Omega \end{aligned} \tag{20}$$

If $\Omega$ is a convex polygon, $dist(\alpha,\Omega)$ can be evaluated by using a quadratic programming solver.

**Algorithm 2:** Determine the closest set to the initial position

```
$d \leftarrow \infty$
for $i = 1 \rightarrow N_{ts}$ do
    if $d \geq dist(C_r x_0, SetList(i))$ then
        $d \leftarrow dist(C_r x_0, SetList(i))$
        $ClosestSet \leftarrow i$
    end if
end for
$OrdList(1) \leftarrow SetList(ClosestSet)$
$SetList \leftarrow SetList \backslash SetList(ClosestSet)$
```

□

In the second algorithm (Algorithm 3), $N_{elem}(\diamondsuit)$ returns the number of target sets in the list $\diamondsuit$ and $dist(\Gamma,\Omega)$ is a function that returns the minimal distance (employing the 2-norm) between two sets $\Gamma$ and $\Omega$ (notice that the use of $dist(\cdot,\cdot)$ in the previous algorithm is a particular case in which the point $\alpha$ is the only element of the set $\Gamma$).

$$\begin{aligned} dist(\Gamma,\Omega) = \quad & \min\nolimits_{\alpha,\beta} \|\alpha - \beta\|_2 \\ & s.t.\ \alpha \in \Gamma,\ \beta \in \Omega \end{aligned} \tag{21}$$

If $\Gamma$ and $\Omega$ are convex polygons, $dist(\Gamma,\Omega)$ can be again evaluated by using a quadratic programming solver.

Figure 2 shows two examples of the use of $dist(\cdot,\cdot)$ to calculate the minimal distance between a point and a set (a) and between two sets (b).

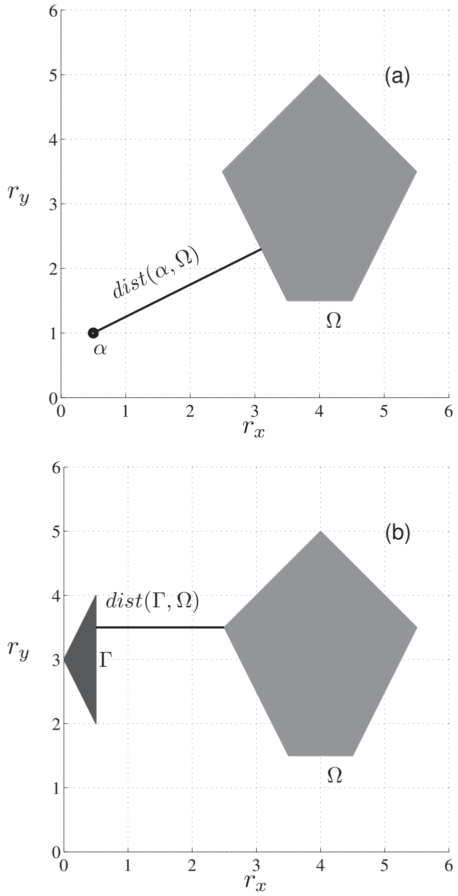


Fig. 2. Illustration of function $dist(\cdot,\cdot)$ involving (a) a point $\alpha$ and a set $\Omega$ and (b) two sets $\Gamma$ and $\Omega$.

**Algorithm 3:** Order the sets from the first to the last, based on smallest distance between them

```
$N_o \leftarrow 1$
while $N_{elem}(SetList) \neq 0$ do
    $d \leftarrow \infty$
    for $i = 1 \rightarrow N_{elem}(SetList)$ do
        if $d \geq dist(OrdList(N_o),SetList(i))$ then
            $d \leftarrow dist(OrdList(N_o),SetList(i))$
            $ClosestSet \leftarrow i$
        end if
    end for
    $OrdList(N_o + 1) \leftarrow SetList(ClosestSet)$
    $SetList \leftarrow SetList \backslash SetList(ClosestSet)$
    $N_o \leftarrow N_o + 1$
end while
```

□

By employing Algorithms 2 and 3, the result is an ordered list of the target sets $OrdList$ containing the set which is the closest to the initial position of the vehicle as its first element and the set that is the closest to the $i$-th set as the $(i+1)$-th element, for $i = 1,\ldots,N_{ts}-1$. This list may be used to provide the terminal target set for the trajectory planner formulation presented in [4]. An algorithm similar to Algorithm 1 may then be used to update the ordered list of sets to be visited ($OrdList$).

## V. RESULTS

This Section contains two subsections: the simulation scenario is described in Subsection V-A and the simulation results are presented and discussed in Subsection V-B.

### A. Simulation Scenario

A kinematic model describing the movement of a vehicle in two dimensions was employed for simulation. The continuous-time model equations are:

$$\dot{r}_x = v_x,\, \dot{v}_x = a_x,\, \dot{r}_y = v_y,\, \dot{v}_y = a_y \tag{22}$$

where $r_x$ and $r_y$ define the position of the vehicle in a horizontal plane with respect to an arbitrary origin. This equation can be recast in state-space form ($\dot{x} = A_c x + B_c u$) by defining the state and control vectors as

$$x = [r_x \quad v_x \quad r_y \quad v_y]^T, \quad u = [a_x \quad a_y]^T \tag{23}$$

For use in the proposed MPC approach with trajectory planning, a discrete-time model of the form $x(k+1) = Ax(k) + Bu(k)$ was obtained with

$$A = \begin{bmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{bmatrix}, B = \begin{bmatrix} 0.5T^2 & 0 \\ T & 0 \\ 0 & 0.5T^2 \\ 0 & T \end{bmatrix} \tag{24}$$

in which $T$ is the sampling period. For the simulations in this paper $T$ was normalized to one time unit.

The dynamical constraints imposed on the velocities were $-1 \leq x_2,\, x_4 \leq 1$. As for the accelerations, the imposed limits were $-5 \leq u_1,\, u_2 \leq 5$.

Constraints $0 \leq x_1,\, x_3 \leq 2$ were also imposed on the position in order to limit it to the known terrain, over which information was assumed to be available.

The initial state of the vehicle was arbitrarily set to $x_0^T = [0\ 0\ 0\ 0]^T$, i. e., it started at rest. The obstacle region was represented as a rectangle with $0.6 \leq x_1, x_3 \leq 1$. It is worth noting that, since only the discrete-time predictions of the position are considered in the inequalities, this does not avoid stretches of the continuous-time trajectory crossing the obstacle. One alternative to handle this issue is proposed in [8], which involves incorporating restrictions on the transition of the vehicle to each region of the space defined by obstacle inequalities. However, it entails the introduction of more binary variables, increasing the complexity of the MILP problem. As an alternative, in this work, the length and width of the obstacle were expanded. To this end, an amount determined through the maximal admissible absolute value of the velocity in each axis was used to expand the borders of the obstacle. Therefore, the adopted avoidance constraints were constructed based on the following expanded obstacle: $0.5 \leq x_1, x_3 \leq 1.1$.

The weight $\gamma$ of the fuel in the cost function was set to $0.1$. The maximal horizon was set to $\bar{N} = 35$. Two missions were simulated, each containing $N_{ts} = 3$ rectangular target sets. The inequalities describing the target sets are presented in Table I.

TABLE I
TARGET SETS OF THE TWO SIMULATED EXAMPLES

| Target set | Example | |
|---|---|---|
| | 1 | 2 |
| 1 | $0.2 \leq x_1 \leq 0.3$<br>$0.9 \leq x_3 \leq 1.0$ | $1.2 \leq x_1 \leq 1.3$<br>$0.8 \leq x_3 \leq 0.9$ |
| 2 | $0.5 \leq x_1 \leq 0.6$<br>$0.2 \leq x_3 \leq 0.3$ | $0.8 \leq x_1 \leq 0.9$<br>$1.7 \leq x_3 \leq 1.8$ |
| 3 | $1.2 \leq x_1 \leq 1.3$<br>$0.9 \leq x_3 \leq 1.0$ | $0.2 \leq x_1 \leq 0.3$<br>$0.7 \leq x_3 \leq 0.8$ |

All simulations were carried out in a personal computer equipped with a Pentium® Dual-Core E5400 processor with $2.7GHz$ clock. For solution of the MILP, the CPLEX toolbox from IBM ILOG was used in Matlab environment, under an academic license. The MPT toolbox [9] was employed to evaluate the $dist(\cdot,\cdot)$ function in Algorithms 2 and 3.

### B. Simulation Results

Figure 3 shows the trajectory in the horizontal plane obtained by employing the multitask planning technique with the target sets of Example 1 in Table I. It can be seen that all targets were visited once and the obstacle region was not crossed. It is worth remarking that the expanded obstacle region is crossed, but not the original obstacle, in agreement with the policy of expanding the obstacle in order to avoid collisions. From Table I, it is possible to note that the first visited target was target set number 2, the second was number 1 and the last, number 3. This shows that the order in which the targets are visited does not depend on the order that they are informed to the planner, because it chooses the visiting order in terms of the solution which provides the minimum cost as a compromise between the overall mission accomplishment time and the fuel spent. Moreover, as shown in Fig. 4, the accelerations $a_x$ and $a_y$ (which correspond to the controls $u_1$ and $u_2$, respectively) remained within the $\pm 5$ bounds.

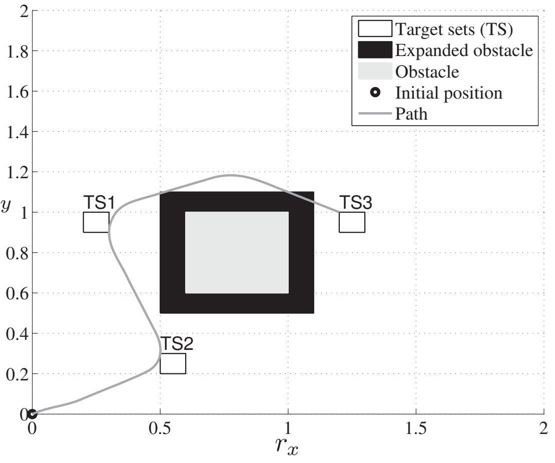


Fig. 3. Trajectory resulting from the use of the proposed multitask planning and control for Example 1.

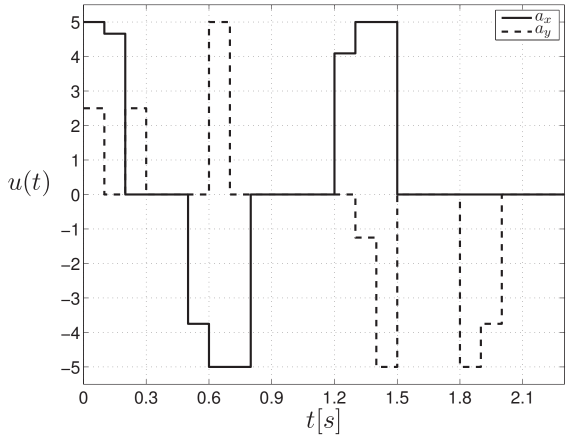


Fig. 4. Control action resulting from the use of the proposed multitask planning and control for Example 1.

The target sets were reached at times: $k = 7$, $k = 13$ and $k = 23$. The fuel cost was $62.50$ and the overall cost, $29.25$.

For comparison, Fig. 5 shows the trajectory in the horizontal plane generated by the minimum-distance trajectory planning algorithm described in Section IV. It can be seen that the minimum-distance choice for ordering the sets to be visited resulted in the same order as that of the multitask receding horizon planning and control shown in Fig. 3. However, the vehicle is now required to make a sharper turn after reaching the first target set and passes closer to the obstacle as compared to the first case, which is a result of the fact that the optimization is not global in this case, since the presence of a second target set is only informed for the planner/controller at the time the vehicle reaches the first target set. Again, as shown in Fig. 6, the accelerations $a_x$ and $a_y$ remained within the $\pm 5$ bounds.

The target sets were reached at times $k = 6$, $k = 14$ and $k = 26$. As can be seen, the overall time to accomplish the mission (26 sampling times) was three sampling times larger than the one observed with the multitask planning and control technique (23 sampling times). The fuel cost was $65.47$ and the overall cost was $32.55$, both larger than the corresponding values obtained with the multitask planning.

Figure 7 shows the trajectory in the horizontal plane obtained by using the multitask planning technique in Example 2. Again, the obstacle was not crossed and all targets were visited once. The first visited target was target set number 3, the second was number 2 and the last, number 1. It can be seen in Fig. 8 that the accelerations $a_x$ and $a_y$ remained within

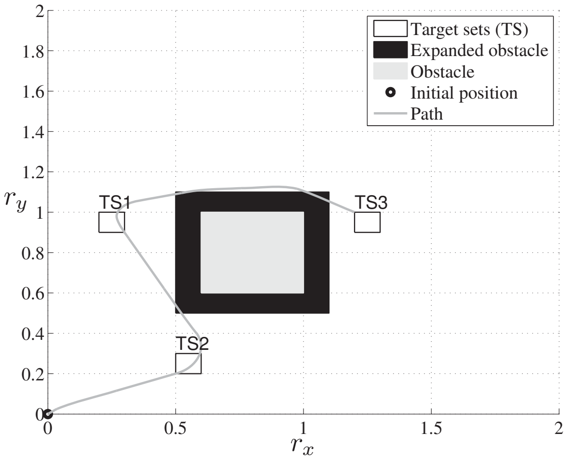


Fig. 5. Trajectory resulting from the use of the minimum-distance target ordering for Example 1.

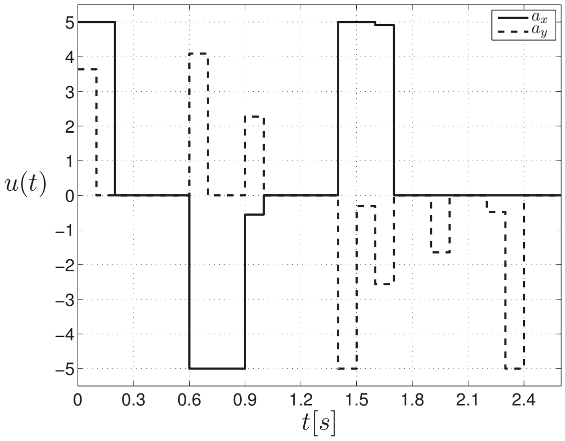


Fig. 6. Control action resulting from the use of the minimum-distance target ordering for Example 1.

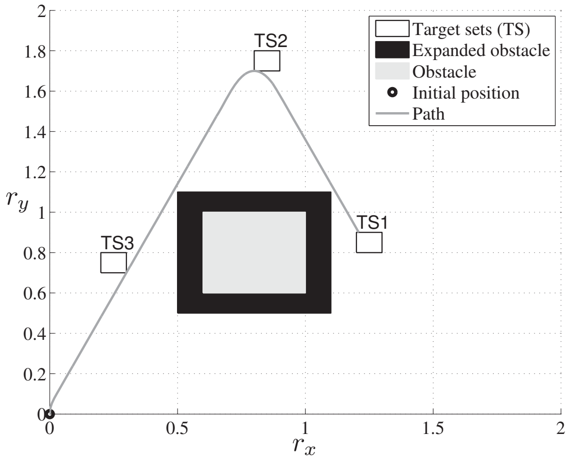


Fig. 7. Trajectory resulting from the use of the proposed multitask planning and control for Example 2.

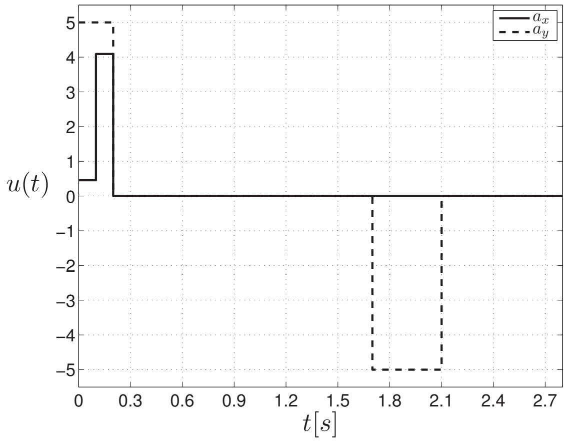


Fig. 8. Control action resulting from the use of the proposed multitask planning and control for Example 2.

the $\pm 5$ bounds.

The target sets were reached at times $k = 8$, $k = 19$ and $k = 28$. The fuel cost was $34.55$, which is smaller than the corresponding value obtained in Example 1. This can be explained by the fact that the resulting trajectory in this case is composed of two straight lines connected by a curve. The straight lines are more economic regarding the sum of the absolute values of the control signals, since no increment is necessary in these signals in order to travel along straight lines. The overall cost was $31.46$, larger than the one obtained in Example 1 due to the larger time necessary to accomplish the maneuver in this case.

For comparison, Fig. 9 depicts the trajectory in the horizontal plane resulting of the employment of the minimum-distance trajectory planning algorithm described in Section IV. It can be seen that the minimum-distance choice for ordering the sets to be visited resulted in a different order as that of the multitask receding horizon planning and control shown in Fig. 7. The first visited target was target set number 3, the second was number 1 and the last, number 2. Figure 10 shows the accelerations $a_x$ and $a_y$, which once again remained within the bounds $\pm 5$.

The target sets were reached at times $k = 8$, $k = 19$ and $k = 31$ sampling times, which gives an overall time to accomplish the mission three sampling times larger than the one observed with the multitask receding horizon planning and control. The fuel cost was $73.26$ and the overall cost was $38.33$, both larger than the ones obtained with the multitask planning. It is interesting to note that the trajectory in this case involves more curves than the one in Fig. 7, and even a reversion in the direction after reaching the target set number 1. This explains the larger fuel cost obtained with the minimum-distance order algorithm $(73.26)$ as compared to the one obtained with the multitask algorithm $(34.55)$.

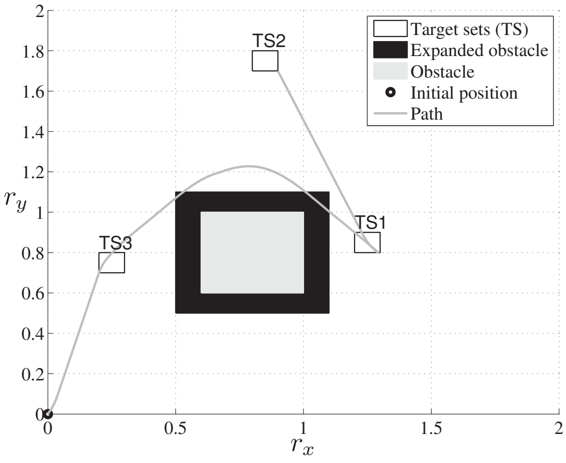


Fig. 9. Trajectory resulting from the use of the minimum-distance target ordering for Example 2.

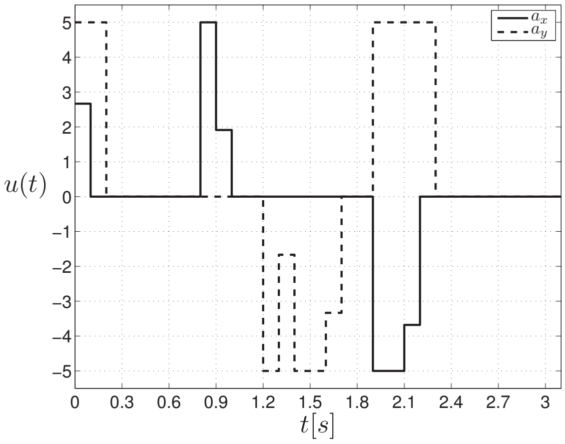


Fig. 10. Control action resulting from the use of the minimum-distance target ordering for Example 2.

## VI. Conclusions

This work proposed a novel formulation of a Predictive Control framework which builds upon the one presented in [4] in order to include multiple target sets in the optimization problem. In the context of autonomous vehicles, this is important due to the fact that some missions require the vehicle to visit a number of target sets.

Simulations were employed to validate the proposed technique. Another approach that first orders the list of sets to be visited based on a minimum-distance criterion and then uses each of them as a terminal target set in the formulation presented in [4] was used for comparison. The approach proposed in the present work outperformed the last one in terms of overall cost of the resulting trajectory.

Future works could include robustness regarding an unknown but limited disturbance to this multitask framework by using an adequate constraint tightening approach, as in [4].

## Acknowledgments

The authors acknowledge the support of FAPESP (grants 2011/18632-8 and 2011/17610-0) and CNPq (research fellowships).

## References

[1] C. Goerzen, Z. Kong, and B. Mettler, "A survey of motion planning algorithms from the perspective of autonomous UAV guidance," *J. Intell. and Robotic Systems*, vol. 57, no. 1, pp. 65 – 100, 2010.

[2] S. J. Qin and T. A. Badgwell, "A survey of industrial model predictive control technology," *Control Engineering Practice*, vol. 11, no. 7, pp. 733–764, 2003.

[3] J. M. Maciejowski, *Predictive Control with Constraints*. Harlow, England: Prentice Hall, 2002.

[4] A. Richards and J. P. How, "Robust variable horizon model predictive control for vehicle maneuvering," *Int. J. Robust and Non-linear Control*, vol. 16, no. 7, pp. 333 – 351, 2006.

[5] J. A. Rossiter, *Model-based Predictive Control: a practical approach*. Boca Raton: CRC Press, 2003.

[6] A. Bemporad and M. Morari, "Control of systems intregrating logic, dynamics, and constraints," *Automatica*, vol. 35, no. 3, pp. 407 – 427, 1999.

[7] A. Richards, T. Schouwenaars, J. P. How, and E. Feron, "Spacecraft trajectory planning with avoidance constraints using mixed-integer linear programming," *J. Guid. Control Dyn.*, vol. 25, no. 4, pp. 755 – 764, 2002.

[8] M. H. Maia and R. K. H. Galvão, "On the use of mixed-integer linear programming for predictive control with avoidance constraints," *Int. J. Robust and Non-linear Control*, vol. 19, pp. 822 – 828, 2009.

[9] M. Kvasnica, P. Grieder, and M. Baotić, "Multi-Parametric Toolbox (MPT)," 2004. [Online]. Available: http://control.ee.ethz.ch/ mpt/

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Connected Navigation of Non-Communicating Mobile Agents

Feza Kerestecioğlu
Department of Electrical-Electronics Engineering
Kadir Has University
Istanbul, Turkey
Email: kerestec@khas.edu.tr

Ahmet Cezayirli
Department of Research and Development
Forevo Digital Design Ltd.
Istanbul, Turkey
Email: cezayirli@ac.forevo.com

***Abstract*—This article discusses the connectivity of autonomous mobile robots that do not have communication capabilities. We show that if the group members follow the proposed Local Steering Strategy, which utilizes information only about the relative positions of neighbor robots, they can sustain their connectivity, even in the case of bounded position measurement errors and the occultation of robots by other robots in the group. To reduce the computational burden in the implementation of the proposed methodology, we used sub-optimal solutions.**

***Index Terms*—Mobile robots, autonomous motion, connectivity**

## I. Introduction

One important aspect of the navigation of multiple autonomous mobile agents is how they maintain connectivity. Loosely speaking, the connectivity of a robot group means that each robot can be contacted by any other robot in the group, either directly or via other robots. The method of making contact may differ according to the characteristics of the individual agents. For instance, being able to establish communications via a standard channel of communication, being visible, or being detected by ultrasonic waves are various ways of being contacted, and thus of being connected.

The navigation of autonomous agents may be the primary or secondary task of a group, depending on the application. The transportation of a group of mine-digging robots from one site to another is an example of the latter, in which navigation is a secondary task. However, navigation is the primary task of robots in such missions as defense patrols or underwater exploring. In both cases, connectivity is of vital importance, since it reflects the unity of the group. Thus, connectivity and its maintainability are fundamental concepts in almost any study regarding the decentralized group motion of autonomous agents.

In this paper, we present a methodology for the navigation of autonomous robot groups which maintain group connectivity. We assume that the robots have position sensors of limited range and with bounded measurement errors, but no communication capabilities. In studies related to connected navigation and the group behavior of mobile robots, many authors *assume* group connectivity or communication within the group during the period of motion as a prerequisite for the success of their methods. For example, graph theory or potential field techniques are employed in this way in [1]–[10].

Graph theoretic approaches to maintain the connectivity of mobile agents are mainly based on the maximization of the second smallest eigenvalue (Fiedler value) of the Laplacian matrix of the graph [5]–[8], [11]. Even if this maximization can be accomplished in a distributed manner as suggested in [5], this does not eliminate the necessity of communications between the robots. For example, the method introduced in [5] requires some data to be obtained from neighbor robots to update components of the supergradient of the Laplacian which are computed locally, and [8] provides an extensive literature survey about group connectivity. Only a few studies, however, have focused on the maintenance of connectivity without relying on information exchange or communication between robots [12]–[14]. The algorithmic methodologies in these studies assume that the robots are points, and are designed to work only in $\mathbb{R}^2$ with perfect measurements via sensors.

The approach proposed in a recent work by the authors [15] results in the navigation of a robot group having dynamic topology using only limited-range position sensors with guaranteed connectivity. In [16], an ad hoc method was proposed to resolve possible deadlock cases. In this study, we present an extension of these results by including measurement errors and the occultation of the robots as well as a modification of the navigation strategy to eliminate any deadlock problems.

In the following section, we describe the agents in the group and define the related navigation problems. Section III gives an overview of our theorem on connectivity and discusses a local steering strategy used for maintaining connectivity. A sub-optimal approach to the implementation of this strategy is given in Section IV and tested by simulations in Section V. Lastly, Section VI offers concluding remarks on the study.

## II. Problem Formulation

The robots in this study are assumed to be physically identical, and each robot has the capability of moving in all directions and is equipped with limited-range position sensors. These sensors have a known degree of accuracy. We assume that the sensing capability is omnidirectional, but the sensor results can bear both angular and radial measurement errors.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

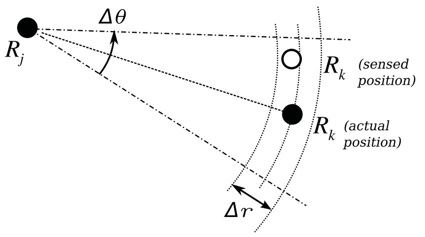


Fig. 1. Angular and radial measurement errors in $\mathbb{R}^2$

Fig. 1 depicts the bounds on the radial and angular components of position errors, $\Delta\theta$ and $\Delta r$, respectively.

It is important to note that sensing other robots means obtaining information about the position of the robots in the neighborhood via position sensors. We shall refer to such a mutual visibility between robots as a *link*. However, we should also note that such a link does not imply any communication or information exchange between the robots. As the robots move around, as long as they maintain visibility with their neighboring agents, they can avoid separating from the other robots, even if they do not communicate with them. Also, it should be pointed out that the robots have no labels, and as a result, sensing the other robots does not imply recognizing a specific robot.

Let us denote a group of autonomous mobile robots that are connected by links as discussed above as $\mathcal{G}$ and the robots in the group as $R_i$, $i = 1, \ldots, N$. Note that the subscripts are arbitrary and for the sake of analysis only. If we consider the robots $R_i, \ldots, R_N$ as vertices and the links between them as the edges of an undirected graph, group $\mathcal{G}$ will be *connected* if there is a path from any robot to any other robot in the group through the links [11]. A group which has at least one pair of robots without a path between them is therefore *disconnected.*

Since the range of the position sensors is limited, a robot may not sense all of the other robots in the group, especially when the total number of robots in the group is large. We refer to the set of robots sensed by $R_i$ as the subgroup $\mathcal{S}_i$. In this way, there are $N$ such subgroups of $\mathcal{G}$, and if $\mathcal{G}$ is connected, $\mathcal{S}_i$ $(i = 1, \ldots, N)$ are nonempty sets.

We denote the radius of the spherical region with $R_i$ at its center and which contains robots in $\mathcal{S}_i$ as $d_{max}$. In other words, $d_{max}$ is the maximum sensing distance for each robot. On the other hand, if the largest distance between the robots in $\mathcal{G}$ is denoted as $D_{max}$, $\mathcal{G}$ will be connected if $d_{max} \geq D_{max}$. However, nontrivial and more interesting cases emerge whenever $d_{max} \ll D_{max}$, which corresponds to groups of relatively large number of agents which have limited sensing ranges.

Fig. 2 depicts a group consisting of three robots. It is seen that $R_2 \in \mathcal{S}_1$, and $R_1 \in \mathcal{S}_2$, which means that $R_1$ and $R_2$ are linked. The links between $R_2$ and $R_3$ are formed likewise. Note that the robot $R_2$ has the position information of both $R_1$ and $R_3$, but $R_1$ and $R_3$ cannot sense each other, as the distance between them is larger than $d_{max}$.

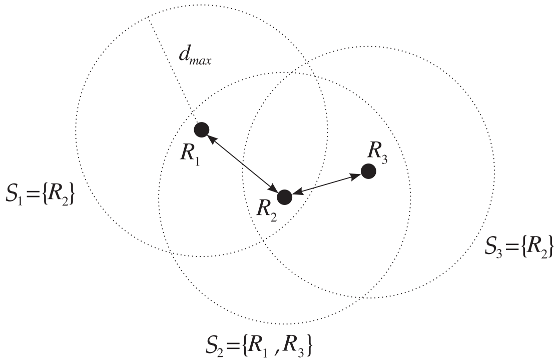


Fig. 2. A group of three robots and their subgroups

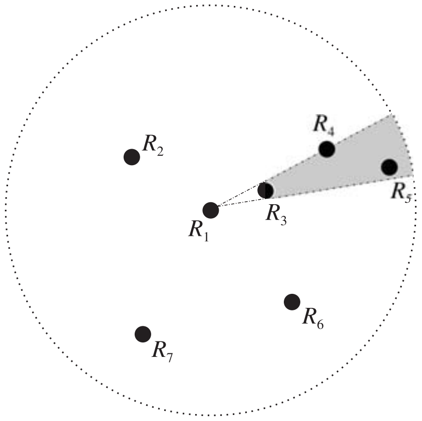


Fig. 3. $R_3$ occults $R_4$ and $R_5$ from $R_1$

In implementing position measurements, which could be performed using any kind of ultrasonic, laser or vision-based sensors, it is inevitable that some robots might occult others. In such a case, occulted robots are not sensed by another robot, say $R_1$ (and hence, are not in $\mathcal{S}_1$) although their distances to $R_1$ are less than $d_{max}$. Fig. 3 depicts an example of occulting in which the position measurements of $R_4$ and $R_5$ cannot be accomplished since $R_3$ prevents $R_4$ and $R_5$ from being "in sight" of $R_1$. Consequently, whenever occultation occurs, the positions of the occulted robots cannot be taken into account in the computation of local movement at that time instant.

Taking into account these sensory limitations and assuming that a set of mobile agents initially represents a connected group, our objective in this work is to develop a decentralized steering methodology that allows for the navigation of the group while preserving its connectivity without requiring any exchange of information between the robots.

In fact, once connectivity is assured, the target or navigation trajectory of the mission need not be known by all group members. It suffices if only one agent has this information [13]. We shall call this robot the *leader* of the group and denote it as $R_N$. Nevertheless, the leader has the same physical properties and capabilities as the other robots. The only difference is that the trajectory to be followed by the group is given to $R_N$. In fact, the leadership of the group is *hidden*. None of the robots recognize the leader as a distinguished group member. In other words, if $R_N$ is sensed by robot $R_j$, i.e. $R_N \in \mathcal{S}_j$, $R_j$ can only see it as one of its neighbors and the leadership of $R_N$ does not affect the local steering strategy of $R_j$.

In the following section, we consider the group of $N$ robots as having one leader, $R_N$, and $N-1$ *followers*, $R_j$ ($j = 1, \ldots, N-1$). Note that the indexing of robots is irrelevant as regards the problem under discussion and the solution we propose. Nonetheless, we utilize such a numbering of robots for the sake of notational simplicity.

## III. Autonomous Motion

Our goal is to develop a methodology for simple autonomous agents, such that a large group of them could navigate as a connected group. We assume that the robots update their information concerning the position of other robots within range of their sensors at every $\Delta t$ seconds. Also, to take into account measurement errors regarding distance, we define a positive scalar $d_m$ as

$$d_m \overset{def}{=} d_{max} - \Delta r$$

where $\Delta r$ is the bound of the distance measurement error with $d_{max} > \Delta r > 0$. We will denote the position of a robot $R_i$ at time $t$, as $X_i(t)$, $i = 1, \ldots, N$. Since all robots in the group move autonomously, we will set up local moving rules for each robot.

The motion of each robot is most conveniently described in terms of a coordinate system referring to itself, since each robot is at the center of its own local coordinate system. We denote the position vector in the local coordinates as $x(t)$ and use a notation such that the superscripts in $x$ indicate the robot to which the coordinate frame is attached, and the subscripts indicate which robot's position it represents. For example, $x_k^j$ represents the position vector of $R_k$ in the coordinate frame of $R_j$. For the robots in $\mathcal{S}_i$, $i = 1, \ldots, N$, we have

$$\|x_k^i(t)\| = \|X_k(t) - X_i(t)\| \leq d_m, \qquad k = 1, \ldots, M$$

where $M$ is the number of robots in $\mathcal{S}_i$. In the next section, we present a result on the sufficient conditions for the maintainability of connectivity.

### A. Main Theorem

According to the notation given above, $x_i^i(t+\Delta t)$ is the location, which $R_i$ targets (for the time instant $t+\Delta t$), in $R_i$'s own coordinate system at time $t$. For any $x_i^i(t+\Delta t)$, let us define two complementary subsets of $\mathcal{S}_i$ as

$$\begin{aligned}\mathcal{S}_{ip} &= \left\{R_p \in \mathcal{S}_i \mid [x_i^i(t+\Delta t)]^{\mathrm{T}} x_p^i(t) \leq 0\right\} \\ \mathcal{S}_{iq} &= \left\{R_q \in \mathcal{S}_i \mid [x_i^i(t+\Delta t)]^{\mathrm{T}} x_q^i(t) > 0\right\}.\end{aligned}$$

This means that if a displacement of $R_i$ to $x_i^i(t+\Delta t)$ will take $R_i$ closer to a robot, then this robot will appear in $\mathcal{S}_{iq}$. Otherwise, it will be a member of $\mathcal{S}_{ip}$. Using, $\mathcal{S}_{ip}$, $\mathcal{S}_{iq}$ and also the notation defined above, we can state the following theorem on group connectivity.

*Theorem 1:* Consider a group $\mathcal{G}$ of $N$ autonomous mobile robots which are connected at $t = 0$. If the motion of the robots is subject to the constraints

$$\|x_i^i(t+\Delta t)\| \;\leq\; \frac{1}{2}\left(d_m - \max_{R_p \in \mathcal{S}_{ip}} \|x_p^i(t)\|\right) \qquad (1)$$

and

$$\|x_i^i(t+\Delta t)\|^2 \;\leq\; \min_{R_q \in \mathcal{S}_{iq}} \left\{[x_i^i(t+\Delta t)]^{\mathrm{T}} x_q^i(t)\right\} \qquad (2)$$

for $i = 1, \ldots, N$, the group preserves its connectivity for $t > 0$.

*Proof:* Note that the position of each robot in $\mathcal{S}_i$ can constrain the motion of $R_i$ either via (1) or via (2), based on whether this robot appears in $\mathcal{S}_{ip}$ or $\mathcal{S}_{iq}$. Let $R_a$ and $R_b$ be any two robots within their mutual sensing range, that is, $R_a \in \mathcal{S}_b$ and $R_b \in \mathcal{S}_a$ at time $t$.

First, suppose that $R_b \in \mathcal{S}_{ap}$ and $R_a \in \mathcal{S}_{bp}$. Then, it follows from (1)

$$2\|x_a^a(t+\Delta t)\| + \max_p \|x_p^a(t)\| \leq d_m \qquad (3)$$

and

$$2\|x_b^b(t+\Delta t)\| + \max_p \|x_p^b(t)\| \leq d_m. \qquad (4)$$

Noting that $\max_p \|x_p^a(t)\| \geq \|x_b^a(t)\|$, $\max_p \|x_p^b(t)\| \geq \|x_a^b(t)\|$, and $\|x_b^a(t)\| = \|x_a^b(t)\|$, we obtain from (3) and (4),

$$\|x_a^a(t+\Delta t)\| + \|x_b^b(t+\Delta t)\| + \|x_b^a(t)\| \leq d_m.$$

Further, by triangle inequality, we get

$$\|x_a^a(t+\Delta t) - [x_b^a(t) + x_b^b(t+\Delta t)]\| \leq d_m. \qquad (5)$$

Note that the term $x_b^a(t) + x_b^b(t+\Delta t)$ is the position of $R_b$ at time $t+\Delta t$ as expressed in the local coordinate frame attached to $R_a$ at time $t$. Therefore, (5) shows that the distance between the robots $R_a$and $R_b$ will not be larger than $d_m$ at time $t+\Delta t$.

Next, we assume that $R_b \in \mathcal{S}_{aq}$ and $R_a \in \mathcal{S}_{bq}$. In this case, we have to proceed using the constraint in (2). Namely,

$$\|x_a^a(t+\Delta t)\|^2 \leq [x_a^a(t+\Delta t)]^{\mathrm{T}} x_b^a(t). \qquad (6)$$

Since

$$\left\|x_a^a(t+\Delta t) - \frac{x_b^a(t)}{2}\right\|^2 = \|x_a^a(t+\Delta t)\|^2 + \left\|\frac{x_b^a(t)}{2}\right\|^2 - [x_a^a(t+\Delta t)]^{\mathrm{T}} x_b^a(t), \qquad (7)$$

using (6), we obtain

$$\left\|x_a^a(t+\Delta t) - \frac{x_b^a(t)}{2}\right\| \leq \frac{\|x_b^a(t)\|}{2}. \qquad (8)$$

Noting that $x_b^a(t) = -x_a^b(t)$ and using the triangle inequality along with (8), it follows that

$$\begin{aligned}
&\|x_a^a(t+\Delta t) - [x_b^a(t) + x_b^b(t+\Delta t)]\| \\
&\quad = \left\| \left( x_a^a(t+\Delta t) - \frac{x_b^a(t)}{2} \right) - \left( x_b^b(t+\Delta t) - \frac{x_a^b(t)}{2} \right) \right\| \\
&\quad \le \left\| x_a^a(t+\Delta t) - \frac{x_b^a(t)}{2} \right\| + \left\| x_b^b(t+\Delta t) - \frac{x_a^b(t)}{2} \right\| \\
&\quad \le \|x_b^a(t)\| \\
&\quad \le d_m,
\end{aligned} \tag{9}$$

which asserts the link between $R_a$ and $R_b$ at time $t + \Delta t$ in the same way as (5).

To complete the proof, we also must analyze cases where $R_b \in \mathcal{S}_{aq}$ while $R_a \in \mathcal{S}_{bp}$, and $R_b \in \mathcal{S}_{ap}$ while $R_a \in \mathcal{S}_{bq}$. Without loss of generality, we consider only the former, since the proof for the latter can be obtained by an interchange of subscripts $a$ and $b$ only.

In other words, the motion of $R_a$ and $R_b$ will be constrained by (6) and (4), respectively. Similar to (7), we can state:

$$\begin{aligned}
&\|x_a^a(t+\Delta t) - x_b^a(t)\|^2 \\
&\quad = \ \|x_a^a(t+\Delta t)\|^2 + \|x_b^a(t)\|^2 - 2[x_a^a(t+\Delta t)]^{\mathrm{T}} x_b^a(t).
\end{aligned}$$

In light of (6), we get

$$\|x_a^a(t+\Delta t) - x_b^a(t)\| \le \|x_b^a(t)\|. \tag{10}$$

Therefore, (4) with (10) yields

$$2\|x_b^b(t+\Delta t)\| + \|x_a^a(t+\Delta t) - x_b^a(t)\| \le d_m$$

or

$$\|x_a^a(t+\Delta t) - x_b^a(t) - x_b^b(t+\Delta t)\| \le d_m - \|x_b^b(t+\Delta t)\|. \tag{11}$$

Hence, the validity of (5) is maintained in this case as well.

The results in (5), (9) and (11) show that any two robots $R_a \in \mathcal{S}_b$ and $R_b \in \mathcal{S}_a$ sensing each other at time $t$ will still be linked when they move to their new locations at $t + \Delta t$. Hence, we conclude that if the group is connected at $t = 0$, it will also be connected for $t > 0$. ■

Note that if a robot $R_i$ is occulted by another robot in $\mathcal{S}_i$, the number of robots in $\mathcal{S}_i$ might decrease. Nevertheless, this does not disturb the overall connectivity, as the existence of the occulting robot itself is the evidence of the connection between $R_i$ and the occulted robot.

### *B. Local Steering Strategy*

As long as the constraints in (1) and (2) are satisfied when a given navigation trajectory is followed, formation control and other mission-oriented tasks can be accomplished by minimizing suitable cost functions. Therefore, in view of Theorem 1, once group connectivity is assured by (1) and (2), the following Local Steering Strategy can be applied for navigation by a group which is composed of follower robots and a leader.

*Local Steering Strategy:* Subject to the constraints (1) and (2)

- The follower robots $R_j$ ($j = 1, \ldots, N-1$) move towards a target location $x_j^j(t+\Delta t)$, which minimizes the cost function $J(x_j^j(t+\Delta t))$ related to the positions of the robots in $\mathcal{S}_j$.
- The leader $R_N$ follows the navigation trajectory.

Several types of cost functions can be used in implementing the local steering strategy. An example may be given as

$$J(x_j^j(t+\Delta t)) = \max_k \|x_j^j(t+\Delta t) - x_k^j(t)\|, \tag{12}$$

which makes the $j^{\text{th}}$ robot try to decrease the distance to the farthest robot that it senses. Another possible approach could be to employ

$$J(x_j^j(t+\Delta t)) = \sum_{k=1}^{M} \Big( \|x_j^j(t+\Delta t) - x_k^j(t)\| - d_0 \Big)^2 \tag{13}$$

as the cost function in order to force the robots to keep their distances with the robots in their subgroups as close to a desired distance as possible. Here, $d_0$ ($d_0 < d_m$) denotes the desired distance.

Note that both (12) and (13) are defined in terms of local coordinates to ensure a distributed algorithm. The choice of suitable cost functions will depend on the requirements of the mission. It is possible to take into account fixed as well as time-varying cost functions, and these can incorporate the position information of all or only some of the neighbor agents. Furthermore, it is possible for each agent to minimize a different cost function.

Note that the constraint in (1) delimits the distance to be travelled by each robot at every sampling interval. It is a direct consequence of the requirement that the distance between any two robots that are connected at time $t$ must be less than $d_m$ at $t+\Delta t$, even in the worst case, which happens when the robots are moving in opposite directions. Therefore, (1) alone is sufficient to maintain connectivity. Nevertheless, it was shown in [15] and [16] that if the motion of the robots is constrained only by (1), the group may get stuck in situations where none of the robots can move. Such situations are called *deadlock*. A typical example of a deadlock is when all inter-robot distances are $d_{max}$, so that $\|x_i^i(t+\Delta t)\| = 0$, $i = 1, \ldots, N$ and, hence, none of the robots can move. In avoiding deadlocks, it is essential to allow the outermost robots in the group to move towards their neighbors. However, (1) depends only on the magnitude of $x_i^i(t+\Delta t)$ and does not account for its direction. The direction of $x_i^i(t+\Delta t)$ can be utilized by considering the neighbor robots in two distinct subsets as implied by (1) and (2). In fact, it is shown in [17] that the constraints (1) and (2) prevent deadlocks under reasonably mild conditions as regards the cost functions, which are fulfilled by, for example, those in (12) and (13).

## IV. A SUBOPTIMAL SOLUTION TO A LOCAL STEERING PROBLEM

The robots in this study are quite simple and limited devices especially from the computational point of view. Our purpose is to provide a decentralized control methodology which can be applied to such simple robots yet still lead to satisfactorily good group navigation. Below, we propose a gradient-descent-based iterative method to reduce the computational burden in the implementation of the Local Steering Strategy with the cost function in (13).

The minimum of the cost function given in (13) is the location to which follower robot $R_j$ aims to arrive at each sampling time. In a general case, the minimization of (13) subject to the constraints in (1) and (2) requires solving a set of nonlinear equations and this approach may result in several local minima, which should be further checked to ensure they are the absolute minimum. In other words, it requires high computational power, especially when the number of the robots in $\mathcal{S}_j$ increases.

It should be noted that the optimal points are computed for each follower robot $R_j$ at every sampling time. The location of the optimal points depends on the positions of the robots in the subgroup $\mathcal{S}_j$. Since the sensed robots in $\mathcal{S}_j$ also move autonomously, these local targets will be updated every $\Delta t$ sec, possibly before reaching them. Since (1) and (2) constrain the magnitude of $x_i^i(t+\Delta t)$, they will not be violated as the robots are moving towards their local targets. Hence, the solution will only provide a direction to the optimal points, because the solution will be updated before $R_j$ arrives at that location.

This fact can be exploited to introduce an iterative method to implement the Local Steering Strategy in a sub-optimal way. Rather than solving for the minimum points of the cost function, each robot $R_j$ can move in the direction of the negative gradient of the cost function, evaluated at the position of $R_j$ for each sampling instant. That is,

$$x_j^j(t+\Delta t) = x_j^j(t) - \gamma \left. \frac{\partial J(x_j^j(t+\Delta t))}{\partial x_j^j(t+\Delta t)} \right|_{x_j^j(t+\Delta t) = x_j^j(t)} \qquad (14)$$

where $\gamma > 0$ is a positive gain, and $x_j^j$ is the position vector of $R_j$ in its local coordinates. From (13), (14) and the fact that $x_j^j(t) = 0$, it follows that

$$x_j^j(t+\Delta t) = 2\gamma \sum_{k=1}^{M} \left( \|x_k^j(t)\| - d_0 \right) \frac{x_k^j(t)}{\|x_k^j(t)\|}. \qquad (15)$$

The gain $\gamma$ in (15) should be treated as a parameter by which one can choose the distance of the local target so as to satisfy the constraints in (1) and (2), rather than as a constant to be determined a priori. In that respect, there is no reason why $\gamma$ must be kept constant during navigation. Therefore, the direction of the next movement can be obtained by (15) and then implemented in this direction with a magnitude so as to satisfy the inequalities (1) and (2).

## V. Simulation Results

In this section, we present our computer simulations to verify the theoretical results of the previous sections. In the simulations, the working space was taken as a section of $xy$-plane in $\mathbb{R}^2$. The sensor range ($d_{max}$) was 15 units. The bounds on the measurement errors were $\Delta\theta = 12°$ for angle and $\Delta r = 0.03 d_{max}$ for distance measurements. The following scenario was applied: The leader was given a trajectory and as the leader started navigation, the rest of the group also moved so as to stay connected under the Local Steering Strategy. The simulation was done for a group consisting of 20 robots where (15) is used with $\gamma = 0.2$.

The snapshots of the simulation can be seen in Fig. 4. With the initial locations shown in Fig. 4 and $d_{max} = 15$, the group was connected and accommodated 62 links at the beginning. The trajectory of the leader is shown by the solid line through the graph. As soon as the simulation started with $d_0 = 11$, the group widened but kept its connectivity. Occulting happened rarely. For 20 robots, the maximum possible number of links is $20 \cdot (20-1)/2 = 190$. Fig. 5 indicates that at least 63 of 190 possible links were preserved until the end of the navigation. The sharp turns in the trajectory of the leader are important, as they could disrupt the shape of the group.

Also, the impact of $d_0$ on connectivity can be assessed by Fig. 5, in which the number of links in the group throughout the entire simulation is plotted for three different values of $d_0$. It is seen that for $d_0 = 5$, the number of links lies between those for $d_0 = 11$ and $d_0 = 8$. This is an interesting result and deserves some interpretation. Whenever $d_0$ is small, the group is dense and occupies a smaller area. Conversely, the group widens with increasing $d_0$. However, when the robots are confined to a smaller volume due to a low value in $d_0$, the number of occulted robots increases significantly. Hence, the number of links decreases because visibility is reduced by occulting. However, as was noted in Section III, this does not imply any weakness for overall group connectivity.

## VI. Conclusions

This work examined the navigation of mobile robot groups and the methodology presented does not require communication between the robots. Rather, a local steering strategy, which uses only information regarding the position of neighbor agents, is employed to sustain the connectivity of group members. The limited-range sensors are modeled in such a way that they produce angular and radial position measurement errors to better reflect a realistic situation. Moreover, robots may be occulted by other robots. In this way, the methodology accounts for all of the fundamental difficulties that can arise in real-life implementation.

This study demonstrates that once the robots start their motion as a connected group, the steering strategy assures their connectivity without any risk of deadlock. Also, the fact that no communication or hierarchy among the robots is required makes it possible for new members to be easily accepted into the group. The simulations verified the success of the methodology.

## References

[1] H. G. Tanner, A. Jadbabaie and G. J. Pappas, "Stable flocking of mobile agents, Part I: Fixed topology", in *Proceedings of the Conference on Decision and Control*, Maui, Hawaii, pp. 2010–2015, 2003.

[2] H. G. Tanner, A. Jadbabaie and G. J. Pappas, "Stable flocking of mobile agents, Part II: Dynamic topology", in *Proceedings of the Conference on Decision and Control*, Maui, Hawaii, pp. 2016–2021, 2003.

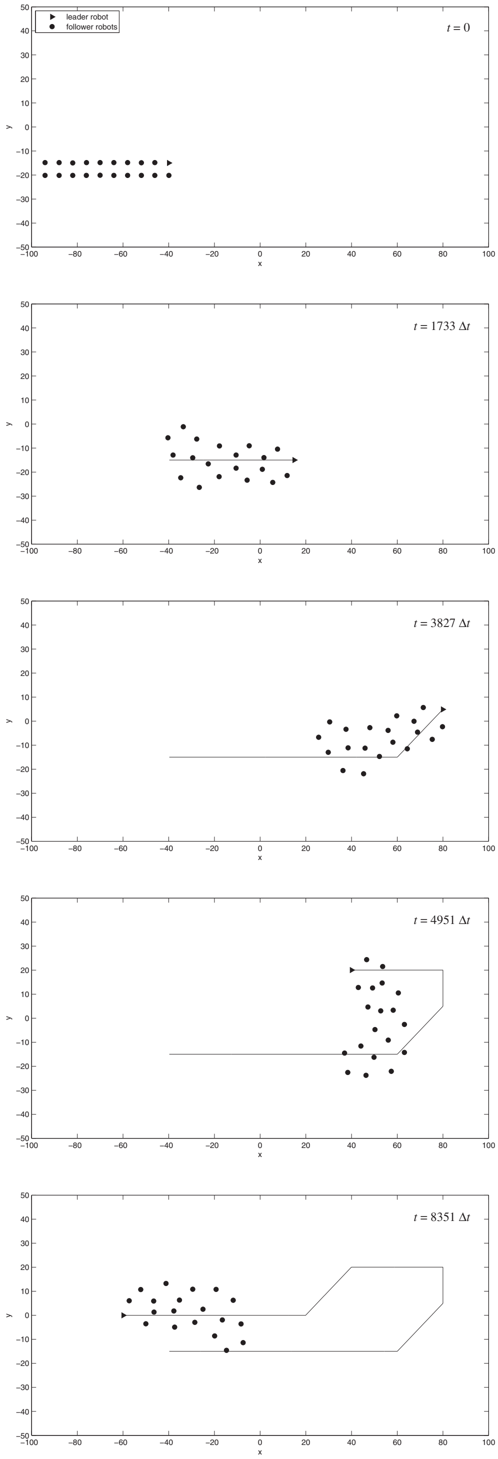


Fig. 4. Navigation of 20 robots ($d_{max} = 15$, $d_0 = 11$)

[3] Z. Lin, M. Broucke and B. Francis, “Local control strategies for groups of mobile autonomous agents”, *IEEE Transactions on Automatic Control*, vol. 49, pp. 622–629, 2004.

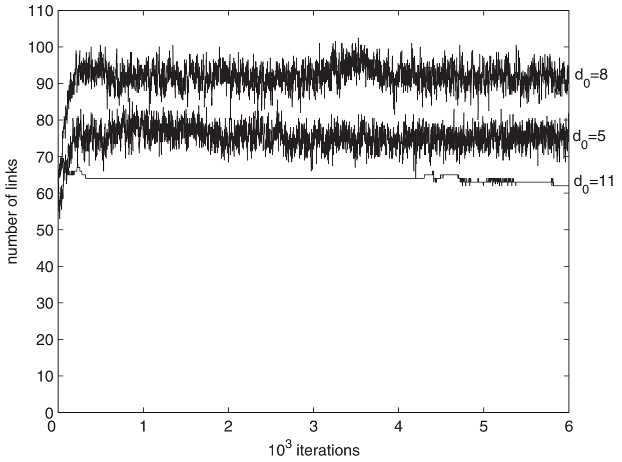


Fig. 5. Total number of links in a group of 20 robots ($d_{max} = 15$)

[4] G. A. S. Pereira, V. Kumar and M. F. M. Campos, “Closed loop motion planning of cooperating mobile robots using graph connectivity”, *Robotics and Autonomous Systems*, vol. 56, pp. 373–384, 2008.

[5] M. C. De Gennaro and A. Jadbabaie, “Decentralized control of connectivity for multi-agent systems”, in *Proceedings of the 45th Conference on Decision and Control*, St. Diego, CA, USA, pp. 3628–3633, 2006.

[6] A. Cornejo and N. Lynch, “Connectivity service for mobile ad-hoc networks”, in *Proc. of the 2nd IEEE International Conference on Self-Adaptive and Self-Organizing Systems Workshops*, pp. 292–297, 2008.

[7] N. Ayanian and V. Kumar, “Decentralized feedback controllers for multiagent teams in environments with obstacles”, *IEEE Transactions on Robotics*, vol. 26, pp. 878–887, 2010.

[8] M. M. Zavlanos, M. B. Egerstedt and G. J. Pappas, “Graph-theoretic connectivity control of mobile robot networks”, *Proceedings of the IEEE*, vol. 99, pp. 1525–1540, 2011.

[9] M. M. Zavlanos and G. J. Pappas, “Potential fields for maintaining connectivity of mobile networks”, *IEEE Transactions on Robotics*, vol. 23, pp. 812–816, 2007.

[10] A. Jadbabaie, J. Lin and A. S. Morse, “Coordination of groups of mobile autonomous agents using nearest neighbor rules”, *IEEE Transactions on Automatic Control*, vol. 48, pp. 988–1001, 2003.

[11] N. Biggs, Algebraic Graph Theory, Cambridge, England: Cambridge University Press, 1993.

[12] H. Ando, Y. Oasa, I. Suzuki and M. Yamashita, “Distributed memoryless point convergence algorithm for mobile robots with limited visibility”, *IEEE Transactions on Automatic Control*, vol. 15, pp. 818–828, 1999.

[13] V. Gervasi and G. Prencipe, “Coordination without communication: the case of the flocking problem”, *Discrete Applied Mathematics*, vol. 144, pp. 324–344, 2004.

[14] P. Flocchinia, G. Prencipe, N. Santoroc and P. Widmayer, “Gathering of asynchronous robots with limited visibility”, *Theoretical Computer Science*, vol. 337, pp. 147–168, 2005.

[15] A. Cezayirli and F. Kerestecioğlu, “Navigation of autonomous mobile robots in connected groups”, in *Proceedings of the Third International Symposium on Communications, Control and Signal Processing*, St. Julians, Malta, pp. 162–167, 2008.

[16] A. Cezayirli and F. Kerestecioğlu, “On preserving connectivity of autonomous mobile robots”, in *IEEE International Conference on Control Applications/International Symposium on Intelligent Control*, St. Petersburg, Russia, pp. 677–682, 2009.

[17] A. Cezayirli and F. Kerestecioğlu, “Navigation of non-communicating autonomous mobile robots with guaranteed connectivity”, submitted to *Robotica*, 2012.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Low Cost Obstacle Detection System for Wheeled Mobile Robot

Ibrahim Alsonosi Nasir
*Department of Electronic and Computer Engineering,*
*Sebha University, Sebha, Libya*
*Ibrn103@yahoo.com*

## Abstract

This paper presents a new low cost obstacle detection system for wheeled mobile robot. The hardware implementation of the proposed detection system uses two Sharp GP2D02 infra-red range sensors which are placed cross over each other in front of the robot in order to detect any object and avoid collisions. The software implementation of the proposed detection system is implemented using timer with the overflow interrupt for generating the required waveform for driving the Sharp GP2D02 infra-red range sensor. As a result, the mobile robot is able to execute different tasks without a delay on the processing time, for example driving the robot wheels and drive the Sharp sensors for detecting an object. Experimental results show that the proposed detection system successfully detect an object placed in a path of the mobile robot.

**Keywords-** Obstacle, Detection system, Mobile robot, Sharp Infra-red Range Sensors.

## 1. Introduction

Robots are very powerful elements of today's life. They are widely used in many industries due to the high level of performance, reliability, capability of performing many different tasks and operations. Recently, mobile robots are more and more involved in our daily lives. They can be employed in several missions such as cleaning rooms, taking care of patients, playing with kids, etc. It is important for many mobile robots to have ability for understanding and be aware of their environment such as the ability to detect an object and avoid collisions. To achieve this, mobile robots have used several types of sensing sensors, such as ultrasonic, infra-red, laser rang finder, charge-coupled device (CCD), web camera, mini-radar and bump sensors [1, 2,3]. The visual sensors can provide the richer source of useful information about the surroundings. However, they are slow in computing data and more expensive [4]. The infrared sensors have widely used for object detection and avoidance collisions due to their low cost and ranging capability [5]. Authors in [5] presented a software implementation of an obstacle detection and avoidance system. In this system, three sharp GP2D12 infra-red range sensors were used to cover the large area in the front of the mobile robot. This type of the infrared sensor offers a analogue voltage corresponding to the distance measured. Therefore, an analogue to digital converter (ADC) is required using such this sensor, so extra cost is required for this system. Authors in [6] designed and developed a low cost mobile robot that designed in as a circle–shaped and five sharp GP2D12 Infra-red Range Sensors were used in order to cover the large area.

In this paper, a new low cost obstacle detection system for wheeled mobile robot is proposed. In comparison with existing approaches, the proposed system possesses the following advantage: (i) the proposed detection system requires only two infra-red range sensors to cover a large area. By contrast, method in [5] uses three infra-red range sensors and method in [6] uses five infra-red range sensors. (ii) The proposed detection system is implemented using the Sharp GP2D02 infra-red range sensor which is a digital sensor and does not required an analogue to digital converter (ADC). By contrast, methods in [5, 6] are implemented using the Sharp GP2D12 infra-red range sensors which is an analogue sensor and required an analogue to digital converter (ADC) and therefore, extra cost is required using such this sensor.

The rest of this paper is structured as follows. Section 2 describes the interfacing between the Sharp GP2D02 infra-red range sensors and the PIC microcontroller. Section 3 covers the details of the proposed obstacle detection system. Section 4 presents experimental results. Conclusions are drawn in Section 5.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

## 2. Interfacing the Sharp GP2D02 Infra-red Range Sensor to the PIC microcontroller

The Sharp GP2D02 infra-red range sensor uses an open-drain input which means it can not be interfaced directly to the PIC microcontroller due to the maximum characteristics of the open drain input which is in range -0.3 to 3 voltages. Therefore, a diode is used to only enable the current to flow when I/O pin is low [7]. Figure 1 shows the Interface between PIC and the sharp GP2D02 infra-red range sensor. As shown, this sensor requires two lines from the PIC in order to be controlled. One line provides the signal to begin a measurement and also is used to provide clock pluses, this line is called $V_{in}$ and the other line is called $V_{out}$ which is used to transmit the measurement back to the PIC microcontroller. The output of this sensor is 8 bit serial measured.

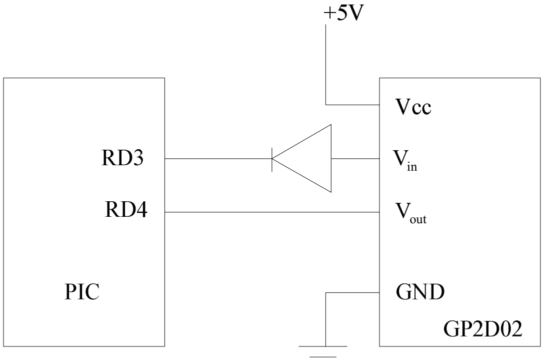


Fig. 1 Interfacing PIC with GP2D02

### 2.1 Driving the Sharp GP2D02 Infra-red Range Sensor

The sharp GP2D02 infra-red range sensor is driven according to the timing diagram which illustrates in figure 2. As shown, the measurement is initiated by forcing the $V_{in}$ signal to logic low for at least 70 ms or until the $V_{out}$ signal from GP2D02 sensor becomes logic high. Once that occurs, $V_{in}$ signal is toggled at the rate of 0.2 ms or less to start clocking in the serial bits from the sensor. Once the entire byte is read, $V_{in}$ is floated high for at least 1.5ms in order to reset the sensor for another reading.

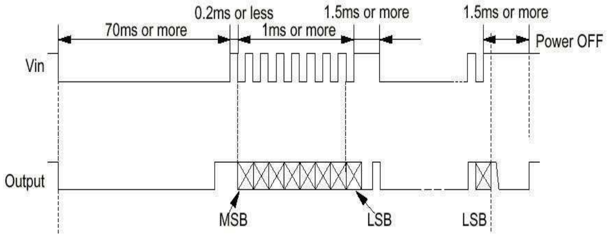


Fig. 2 the timing diagram for GP2D02 infra-red sensor

## 3. The proposed obstacle detection system

The proposed obstacle detection system uses two sharp GP2D02 infra-red range sensors that are mounted in front of the mobile robot cross over each other in order to provide coverage for a large area as shown in figure 3. The sharp GP2D02 sensors can measure a distance to an object by emitting infra-red pulses and then receives back the reflected signal. It can measure distances in the range 10cm – 80cm.

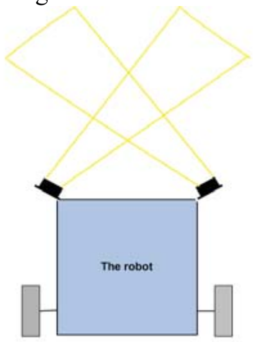


Fig. 3 Block the diagram of the mobile robot with two Sharp infra-red range sensors.

## 3.1 The implementation of the detection system

The implementation of the proposed detection system is divided into two stages: The first stage covers the hardware implementation and the second covers the software implementation. In the hardware implementation, two Sharp GP2D02 infra-red range sensors are used to implement the hardware part of the proposed detection system as shown in figure 4; RD3 pin of the PIC 18F452 is used as the output pin for controlling both of the Sharp sensors. Therefore, one timing waveform is used to initiate, read a byte from each sensor and reset both of the Sharp sensors. RD1 and RD2 are used as input pins in order to read a byte from each sensor.

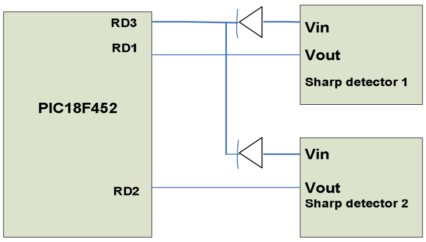


Fig. 4 the hardware implementation of the detection system

In the software implementation, the required waveform is generated for driving the Sharp GP2D02 Infra-red sensor as described in section 2. This required waveform can be implemented by using a delay function. However, this will take about 74ms, and will slow down the processing time on the PIC 18F452 microcontroller. If the required waveform is implemented using a delay function; the robot will be unable to execute different tasks without a delay on the processing time, for example driving the robot wheels and drive the Sharp Infra-red sensors to detect an object. To overcome this, the proposed detection system uses a timer with the overflow interrupt to generate the required waveform for driving the Sharp GP2D02 sensors. The software implementation is divided into three functions that are described as follows.

- **Initial function**

The aim of this function is to initiate the general and the overflow timer 0 interrupt. Also timer 0 is initiated as 16- bit timer/count. Two flags are used to distinguish between the first interrupt which should occur after the timer 0 reaches 70ms and other interrupts that should occur after the timer 0 reaches 0.2 ms and 1.5 ms as shown in figure 5.

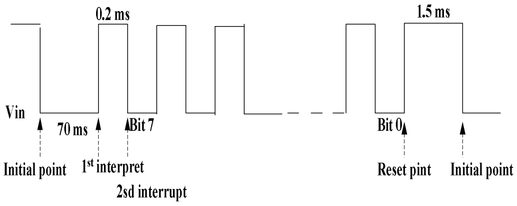


Fig. 5 Timing waveform and interrupt edges

The following Pseudo-codes describe the implementation of the initiate function.

```
void initGP2D02(void)
{
  Enable the interrupt (GIE& timer 0
   overflow INT)
   OpenTimer0( as 16BIT &  256 prescale value);
   WriteTimer0(TIMER_VAL_70MS);
   timer to overflow after 70 ms
   Set timer0_70ms_flag;
   Vin= 0;
   Clear timer0_15ms_flag;
   sensor value=0;
 }
```

- **Read byte from Sharp sensor**

The aim of this function is to read the distance information from the sensors. This information is serial bits and it is converted into a distance of centimeters. The sensors are calibrated using a linearization formula. The sensors output has an inverse relationship to the distance of an object. That is the further the object, the output values of the sensor will be small. The output values of the sensors are converted into distances using the equation (1) [8].

$$D = Kg/(X\text{-}Ko) \qquad (1)$$

Where D denotes a distance and it is given in units of centimeters (cm), X is the sensor output, Kg is the gain and Ko is an offset. The values of Kg and Ko can be determined as follows:

Let D and X be the distance and output, respectively of the first measurement. Let D' and X' be the distance and output, respectively, of the second measurement.

$$Kg = (X'-X)\, D'D \,/\, (D-D') \qquad (2)$$
$$Ko = (D'X' - DX) \,/\, (D' - D) \qquad (3)$$

The Pseudo-codes of this function is described as follows.

```
int readGP2D02(void)
{
 int ret_val = sensor_value;
 Disable Timer 0 interrupt
 ret_val =1560/(ret_val-82);
 Enable Timer 0 interrupt
 return ret_val;
}
```

The 'sensor_value' stores the output byte from the Sharp sensor. This value is updated via the interrupt function. The timer 0 overflow interrupt is disabled at the begging of this function and it is enabled at end by clear or set the timer 0 enable bit in the interrupt control register (INTCON) respectively.

- **Interrupt function**

The aim of this function is to deal with the timer 0 interrupt by generating the required waveform on the $V_{in}$ input pin of the Sharp sensor in order to initiate and read a byte from the output of the Sharp sensors also after the eight bits output from the Sharp sensors are read, it will generate the reset pulse to initiate the sensors to be ready for next measurement. This waveform is generated according to the waveform shown in figure 2. The design of the interrupt function is illustrated in figure 6.

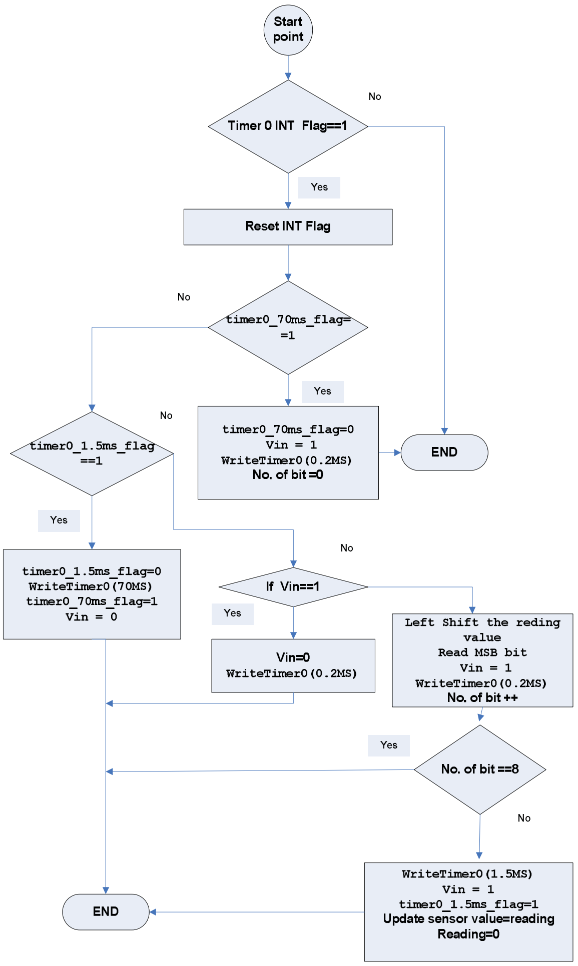


Fig. 6 flowchart of the interrupt function

## 4. Experimental Results

The experiments were organized in two phases. The first phase evaluated the relation between the output of the Sharp GP2D02 Infra-red sensor and the distance to the reflective object. The second phase evaluated the performance of the proposed detection system. To complete the first experimental phase, the Sharp Infra-red sensor was mounted as shown in figure 7 and a reflective object was placed in the front of the sensor in different distances in order to find out the relation between the output of the sensor and the distance to the reflective object.

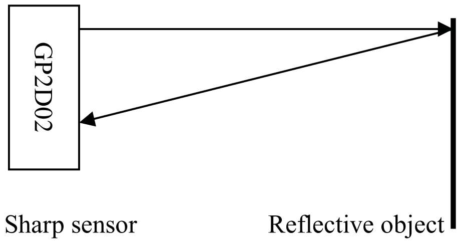


Fig. 7 the first experimental phase

The results of the first experimental phase are shown in figure 8. As shown, the output of the Sharp detector within the range 10 cm - 80 cm. The sensor gives a wrong value for distances less than 10 cm or more than 80cm [7] therefore, an object closer than 10 cm will appear to be further away. The measurement unit of the distance is centimeter and the outputs of the sensors are decimal values.

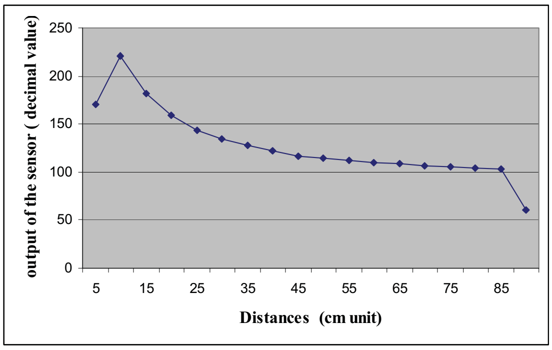


Fig. 8 the results of the first experimental phase

The second experimental phase is divided into two stages; the aim of the first stage is to test each Sharp sensor individually in order to find out if the beam of one sensor affects other sensor. The first stage was achieved by recording the output of each sensor according to the distance to the object without driving the robot; in case one  by activating one sensor and record its output; and in the case two by activating both sensors and record the output of one sensor. This was achieved by using the PORTB as output port and using an oscilloscope to read the output of the sensor from PORTB. The results of these tests are shown in table 1 and table 2. As shown in table 1 and table 2, the output of the Sharp sensor 1 is affected by the beam of the Sharp sensor 2. As a result of this, the output of the Sharp sensor returns a wrong distance measurement to an object. To overcome this problem, extensive experiments were carried out to find the best position for mounting the Sharp sensor. It has been found that the best angle for mounting the position of the Sharp sensors should be 5 degree. If this angle is increased, the sensors will affect each other and they will return wrong distance measurements to an object.

The aim of the second stage is to find out if the proposed detection system is able to detect different types of objects place in deferent positions in the front of the robot; this test was achieved by driving the robot and using a different type of object in different position in front of the robot.  The result of this test is carried out that the robot is able to detect any object even if a small object For example, a pen or a wire.

Table 1 Result of testing sensor 1

| sensor 1 is activated<br>Output of sensor 1 (Decimal) | Both sensors are activated<br>Output of sensor 1 (Decimal) | Distance in cm |
|---|---|---|
| 226 | 226 | 10 |
| 157 | 218 | 20 |
| 135 | 27 | 30 |
| 133 | 14 | 40 |
| 134 | 124 | 50 |
| 95 | 94 | No object |

Table 2 Result of testing sensor 2

| sensor 2 is activated<br>Output of sensor 2 (Decimal) | Both sensor are activated<br>Output of sensor 2 (Decimal) | Distance in cm |
|---|---|---|
| 212 | 212 | 10 |
| 145 | 145 | 20 |
| 126 | 125 | 30 |
| 111 | 110 | 40 |
| 97 | 98 | 50 |
| 32 | 39 | No object |

## 4. Conclusion

A new mobile robot detection system based on two Sharp GP2D02 infrared range sensors is presented. To detect an object in front of the mobile robot, the proposed detection system is implemented using two sharp sensors cross each other in order to cover all the area in front of the mobile robot. The experimental results show that the best angle for mounting the position of the Sharp sensors should be 5 degree. Moreover, experimental results show that the proposed detection system is able to detect any object placed in front of the mobile robot.

## References

[1] H.R. Everett, Sensors for Mobile Robots: Theory and Application. AK Peters, Wellesley, MA, 1995.

[2] M. D. Adams, Sensor modelling, design and data processing for autonomous navigation. River Edge, NJ, World Scientific, 1998.

[3] S. Yue and F. Claire, A Collision Detection System for a Mobile Robot Inspired by the Locust Visual System, Proc. of the IEEE Int. Conf. on Robotics and Automation, Spain, pp. 3832-3837, 2005

[4] S. Yamada, and M. Murota, Unsupervised Learning to Recognize : Environments from Behaviour Sequences in a Mobile Robot. Proc. of the IEEE Int. Conf. on Robotics & Automation, Belgium, PP. 1871- 1876, 1989.

[5] Nwe, A.A., Aung, W.P., and Myint, Y.M., Software implementation of obstacle detection and avoidance system for wheeled mobile robot. World Academy of Science, Engineering and technology 42, PP. 572-577, 2008.

[6] S. Nurmaini, Intelligent Low Cost Mobile Robot and Environmental Classification, International Journal of Computer Applications, Vol. (35)12, 2011.

[7] Acroname, Acroname Articles-Demystifing the Sharp IR Rangers, http://www.acroname.com/robotics/info/articles/sharp/sharp.html

[8] www.barello.net/Papers/GP2D02

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Fuzzy Expert Rule-Based Airborne Monitoring of Ground Vehicle Behaviour

Hyondong Oh, Hyo-Sang Shin,
Antonios Tsourdos and Brian A. White
Department of Engineering Physics
School of Engineering, Cranfield University, UK
Email: h.oh; h.shin; a.tsourdos; b.a.white@cranfield.ac.uk

Seungkeun Kim
Department of Aerospace Engineering
Chungnam National University
99 Daehak-ro, Yuseong-gu
Daejeon 305-764, Republic of Korea
Email: skim78@cnu.ac.kr

***Abstract*—This paper proposes an airborne monitoring methodology of ground vehicle behaviour based on a fuzzy logic to identify suspicious or abnormal behaviour reducing the workload of human analysts. With the target information acquired by unmanned aerial vehicles, ground vehicle behaviour is firstly classified into representative driving modes and then a string pattern matching theory is applied to detect pre-defined suspicious behaviours. Furthermore, to systematically exploit all available information from a complex environment and confirm the characteristic of behaviour, a fuzzy rule-based decision making is developed considering spatiotemporal environment factors as well as behaviour itself. To verify the feasibility and benefits of the proposed approach, numerical simulations on moving ground vehicles are performed using both synthetic and realistic car trajectory data.**

***Index Terms*—Airborne monitoring, Target tracking, Trajectory classification, Behaviour recognition, Fuzzy decision making**

## I. Introduction

Recently, autonomous airborne surveillance and reconnaissance systems become a challenging and emerging problem in the area of aerospace and robotics with the rapid improvement of the UAV (unmanned aerial vehicle) operation and sensing technology. Airborne monitoring allows suspicious or unusual behaviour to be identified and investigated promptly so that situational awareness can be increased in support of border patrol, law enforcement and protecting infrastructure. For this, many researchers have investigated a swarm of autonomous airborne sensor platforms having a long endurance as well as good spatial coverage with an appropriate level of decision making. In particular, surveillance by UAVs equipped with a MTIR (Moving Target Indicator Radar) sensor can provide a certain level of accurate estimation of a large number of moving targets as well as capability to respond possible threats from the air. However, for detection of suspicious behaviours, the operators have to manually analyse the gathered mass data and construct a coherent picture of events. With these backgrounds, this paper focuses on the development of a high-level analysis algorithm to process target information acquired by UAVs which provide awareness of abnormal behaviour.

In general, detecting anomalous behaviour can be classified into two categories: The first approach codifies the behaviours using experience and domain knowledge of experts and the behaviours are learned from data in the second approach [1]. Purely learning based approaches can provide a good performance [2], [3], however, they require massive data set in advance or tend to suffer from the high computation burden for real-time applications. On the other hand, there are several algorithms to deal with behaviour or activity analysis in the context of codified (or classified) behaviour model with the aid of learning. Srivastava *et. al.* [4] introduced the method to detect anomalies of the ground vehicle by observing the patterns in its velocity called as velocity trajectory using hypothetical co-ordinated system in which the axes are specified with respect to the road segment. Besides, Fraile and Maybank [5] proposed the idea of dividing the trajectories of the ground vehicle into several driving modes using video images which can be exploited for ground traffic surveillance. However, this classification is limited to car manoeuvres in an urban parking space with slow speed. Similarly, Kim *et. al.* [6], [7] proposed the trajectory classification codified with more detailed driving modes, and applied it to string matching theory to detect suspicious behaviour defined from expert knowledge.

This paper proposes a fuzzy expert rule-based airborne monitoring methodology of ground vehicle behaviour as an extension of our previous works mentioned above [6], [7]. In those works, a primary source for the behaviour recognition is a single deterministic cost obtained from the string matching. Although this cost can provide the measure of suspiciousness computing similarity between pre-defined suspicious strings and driving mode history from trajectory classification within a certain time window, additional information needs to be considered to finally confirm the characteristic of behaviour while avoiding frequent false alarms. Therefore, in this study, to systematically exploit all available information interconnected and influenced by each other obtained from complex environment, a fuzzy system is applied considering its ability to classify complex sources into simple and intuitive form resulting in the final decision with some degree of confidence through expert rules. The proposed fuzzy expert rule-based decision making allows to concurrently accommodate several aspects of behaviour as well as spatiotemporal environment factors providing a level of alert to operator monitoring complex scenes.

The overall structure of this paper is given as: Section

II briefly introduces a target tracking filter, trajectory classification, and behaviour recognition algorithm using string matching. Section III introduces rule-based decision making algorithm to find suspicious behaviour based on a fuzzy logic. Section IV presents numerical simulation results of behaviour monitoring for both synthetic and civilian traffic scenario using realistic ground vehicle trajectory data. Lastly, conclusions and future works are addressed in Section V. An overall flow chart the technique presented in this paper is shown in Fig. 1.

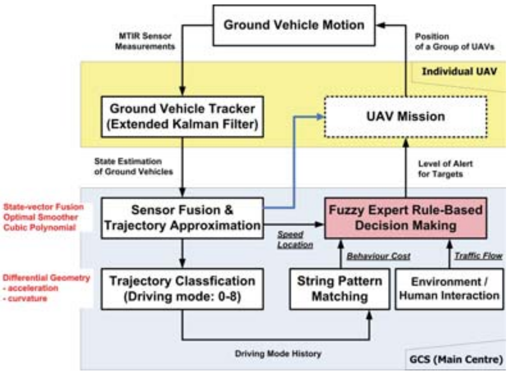

Ground Vehicle Motion
MTIR Sensor Measurements
Position of a Group of UAVs
Individual UAV
Ground Vehicle Tracker (Extended Kalman Filter)
UAV Mission
State Estimation of Ground Vehicles
Level of Alert for Targets
State-vector Fusion
Optimal Smoother
Cubic Polynomial
Sensor Fusion & Trajectory Approximation
Fuzzy Expert Rule-Based Decision Making
Speed
Location
Behaviour Cost
Traffic Flow
Differential Geometry
- acceleration
- curvature
Trajectory Classfication (Driving mode: 0-8)
String Pattern Matching
Environment / Human Interaction
Driving Mode History
GCS (Main Centre)

Fig. 1. An overall flow chart of fuzzy expert rule-based behaviour monitoring

## II. Behaviour Modelling and Recognition

This section briefly introduces our previous works on airborne monitoring of ground vehicle behaviour based on [6], [7]. Ground target tracking filter using UAVs is firstly explained. Trajectory classification is followed to model the behaviour of ground vehicles, and lastly behaviour recognition algorithm using string matching theory is presented.

### A. Target tracking

This study considers acceleration dynamics to apply it to tracking of the moving ground vehicle. This model regards the target acceleration as a process correlated and exponentially decreasing in time, which means if there is a certain acceleration rate at a time $t$, then it is likely to be the same jerk also at a time instant $t+\tau$ as:

$$\mathbf{x}_k^t = F_k \mathbf{x}_{k-1}^t + \eta_k \tag{1}$$

where $\mathbf{x}_k^t = (x_k^t, \dot{x}_k^t, \ddot{x}_k^t, y_k^t, \dot{y}_k^t, \ddot{y}_k^t)^T$, $\eta_k$ is a process noise which represents the acceleration characteristics of the target, and the state transition matrix $F_k$ can be represented as:

$$F_k = \begin{bmatrix} 1 & T_s & \Phi & 0 & 0 & 0 \\ 0 & 1 & \frac{(1-e^{-\alpha T_s})}{\alpha} & 0 & 0 & 0 \\ 0 & 0 & e^{-\alpha T_s} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & T_s & \Phi \\ 0 & 0 & 0 & 0 & 1 & \frac{(1-e^{-\alpha T_s})}{\alpha} \\ 0 & 0 & 0 & 0 & 0 & e^{-\alpha T_s} \end{bmatrix} \tag{2}$$

where $\Phi = (e^{-\alpha T_s} + \alpha T_s - 1)/\alpha^2$, and $\alpha$ is a correlation parameter which allows for the modelling of the different classes of targets. The details of the covariance matrix $Q_k$ of the process noise $\eta_k$ can be found in [8].

Besides, this study assumes the UAV are equipped with a MTIR to localise the position of target. Because the measurement of MTIR is composed of range and azimuth of the target with respect to the radar location, the actual measurements are the relative range and azimuth with respect to the position of the UAV airborne. The radar measurement $(r, \phi)^T$ can be defined as the following nonlinear relation using the target position $(x_k^t, y_k^t)^T$ and the UAV position $(x_k, y_k)^T$ as:

$$\begin{aligned} \mathbf{z}_k &= \begin{pmatrix} r_k \\ \phi_k \end{pmatrix} = h(\mathbf{x}_k^t) + \nu_k \\ &= \begin{pmatrix} \sqrt{(x_k^t - x_k)^2 + (y_k^t - y_k)^2} \\ \tan^{-1} \frac{y_k^t - y_k}{x_k^t - x_k} \end{pmatrix} + \nu_k \end{aligned} \tag{3}$$

where $\nu_k$ is a measurement noise vector, and its noise covariance matrix is defined as:

$$V[\nu_k] = R_k = \begin{bmatrix} \sigma_r^2 & 0 \\ 0 & \sigma_\phi^2 \end{bmatrix}. \tag{4}$$

Considering that the measurement equation is nonlinear, the localisation of target is designed using the EKF (Extended Kalman filter). In addition, assuming a pair of UAVs track the same targets, sensor fusion technique using a Covariance Intersection algorithm is applied. Lastly, as the behaviour of ground vehicle will be analysed with a moving horizon history, the optimal fixed-interval smoother is used to improve the accuracy of past state estimation of the target. The details for sensor fusion and optimal smoother can be found in [6].

### B. Trajectory classification

To model a driver's behaviour, the trajectory is classified into driving modes. The purpose of the classification is to categorise characteristics of manoeuvres associated with forward or lateral driving by assigning them to driving modes as will be explained in the following. This allows not only to recognise characteristic fragments of the trajectories, but also to enable recognition of ground traffic behaviour in an intuitive, computationally-efficient and flexible way.

Since the driving manoeuvre does not happen for a single sampling time, the trajectories for a certain length of time need to be considered. For this, a moving-window-based trajectory approximation [6] is applied using a polynomial function which generates trajectory with a virtually increased sampling time for a certain time interval. Let us assume a new time sequence within a moving window, $0 < T_n < 2T_n < \ldots < (N_T - 1)cT_n = (N_T - 1)T_s$ where $T_s$ is an original sampling time of tracking filter, $T_n$ is a new virtual sampling time, and $N_T$ is the number of samplings for a moving window. In this study, it is assumed that $N_T = 4$, $T_s = 0.5$, and $c = 5$, and thus the new virtual sampling time is 0.1 seconds. The selection of $N_T = 4$, i.e. 1.5 seconds' moving window reflects that the bandwidth for lane changing is at least 1.0Hz according to the reference [9]. Then, velocity $(\dot{x}^t(i), \dot{y}^t(i))$ and acceleration $(\ddot{x}^t(i), \ddot{y}^t(i))$ histories with a new time sequence are used to compute the minimum speed $U$, the rate change of orientation $\theta(i)$, and forward acceleration $a_f(i)$ of the vehicle at current time step $k$ for each $i$ in a moving window (i.e. $k-c(N-1)+1 \leq i \leq k$) as:

$$
\begin{aligned}
U &= \min v(i) = \min \sqrt{\dot{x}^t(i)^2 + \dot{y}^t(i)^2} && (5)\\
\theta(i) &= v(i)\kappa(i) \\
&= \sqrt{\dot{x}^t(i)^2 + \dot{y}^t(i)^2}\frac{\dot{x}^t(i)\ddot{y}^t(i) - \dot{y}^t(i)\ddot{x}^t(i)}{(\dot{x}^t(i)^2 + \dot{y}^t(i)^2)^{3/2}} && (6)\\
a_f(i) &= \ddot{x}^t(i)\cos\psi(i) + \ddot{y}^t(i)\sin\psi(i) && (7)
\end{aligned}
$$

where $\kappa$ is a curvature, and $\psi = \tan^{-1}(\dot{x}^t/\dot{y}^t)$ is the heading angle from the North. Using above equations, a driving mode $m_k^d$ among driving mode set $M^d = \{0, \cdots, 8\}$ at time step $k$ can be obtained for each moving window with a frequency of $1/T_s$ as:

- **Stopping (0)**, $U < 1$: Since 1 m/s equals to 3.6 km/h, it can be assumed that the car does not move or is about to stop or start moving.
- **Left turn (1)**, $\max(\theta)\min(\theta) > 0$ and $\max(\theta) > \theta_{th}$: The inspection of sign change of $\theta$ is used to distinguish the pure turning maneuver from the lane changing.
- **Right turn (8)** $\max(\theta)\min(\theta) > 0$ and $\max(\theta) < -\theta_{th}$
- **Left lane change (2)** $\max(\theta)\min(\theta) < 0$, $\max(|\theta|) > \theta_{th}$, and $\theta(0) > 0$: The difference to the left turn of this condition is the sign change of the rate of orientation change. The sign of curvature transits from positive to negative in case of the left lane change.
- **Right lane change (7)** $\max(\theta)\min(\theta) < 0$, $\max(|\theta|) > \theta_{th}$, and $\theta(0) < 0$: The sign of curvature transits from negative to positive in case of the right lane change.
- **Closing gap (6)** $\max(a_f)\min(a_f) < 0$, and $a_f(0) > 0$: When the driver wants to close gap to the preceding vehicle, the sign of acceleration transits from positive to negative.
- **Widening gap (3)** $\max(a_f)\min(a_f) < 0$, and $a_f(0) \leq 0$: Contrary to the case of closing gap, the sign of acceleration transits from negative to positive.
- **Accelerating ahead (5)** $\max(a_f)\min(a_f) > 0$, and $a_f(0) > 0$: The sign of acceleration keeps positive.
- **Decelerating ahead (4)** $\max(a_f)\min(a_f) > 0$, and $a_f(0) \leq 0$: The sign of acceleration keeps negative contrary to the case of the accelerating ahead.

### C. Behaviour detection

This section introduces behaviour detection scheme to find suspicious behaviour using driving mode histories of ground vehicles. The key tools for this detection scheme are symbolic dynamics and string matching. The mathematical subject of symbolic dynamics originally arose in the theory of dynamical systems and was motivated by the qualitative approach to dynamics in which the character of trajectories is more important than their numerical values. String matching theory is a well-developed area of text processing. String matching consists in finding all the occurrences of a string (called a pattern) in a text where the pattern is a string $x$ of length $m$, while the text is a string $y$ of length $n$. In this study, using the driving mode set $M^d = \{0, \cdots, 8\}$, a symbolic time series of driving modes $y_k^d = \{m_l^d \in M^d | l = 1, \ldots, N_{sm}\}$ is generated by trajectory classification for each time step $k$, where $N_{sm}$ represents a moving window length for string matching. The suspicious behaviour is also expressed as strings $x_s$ consisting of nine numbers.

Intuitive string matching method we can apply is the exact matching which detects exactly the same pattern in the driving mode history as the pre-defined suspicious string. However, assuming a reference suspicious string of '145048', '145548' or '145448' cannot be ignored as well in the detection scheme, whose fourth element of the string might be one of the following forward driving modes: '3'; '4', '5', '6', instead of '0'. To tackle this, an approximate matching is applied by defining a cost which is called as Edit distance measuring distance or similarity between reference and test patterns. The Edit distance $D(S_1, S_2)$ [10] between two string patterns $S_1$ and $S_2$ is defined as the minimum number of editions including changes $C$, insertions $I$, and deletions $R$ required to change pattern $S_1$ into $S_2$. The details of approximate string matching can be found in [6].

Although above edit distance can provide the measure of suspiciousness computing similarity between pre-defined suspicious strings and current driving mode history within a certain time window, additional information needs to be considered to finally confirm characteristic of behaviour while avoiding frequent false alarms. From the following section, what types of information can be used and how to combine them will be dealt with.

## III. Fuzzy Expert Rule-Based Decision Making

For airborne behaviour monitoring, this section proposes a decision making algorithm to find suspicious or anomalous vehicle based on a fuzzy logic. To systematically exploit all available information interconnected and influenced by each other obtained from complex environment, fuzzy system is applied considering its ability to classify complex sources into simple and intuitive form (fuzzification) resulting in the final decision (defuzzification) with some degree of confidence (rather than single certain decision) through expert rules (fuzzy inference). The proposed fuzzy expert rule-based decision making allows to concurrently accommodate several aspects of behaviour as well as spatiotemporal environment factors with supervision of human providing a level of alert to operator monitoring complex scene. Fuzzy system used in this study consists of four fuzzy membership functions for inputs and one output with 36 expert inference rules.

### A. Fuzzification

A fuzzy input for behaviour monitoring includes four aspects: location, behaviour cost, speed of the vehicle, and environmental aspect as:

- **Location:** A time history of the location which is a relative position of the suspicious ground vehicle to the critical area (e.g. the centre of complex activities and the base walls of military facilities) or an index of road

that the ground vehicle has moved along is an important source for behaviour monitoring. Assuming that the local roadmap information is readily available in advance, the indexes of the local roads in region of interest can be annotated by a sequence of road numbers. If the vehicle travelling on one of identified roads of interest, the location is categorised as ’Region of interest (R)’; otherwise is ’General (G)’ as shown in Fig. 2(a).
- **Behaviour cost:** As a key factor for the behaviour monitoring, the edit distance $D$ is used resulting in time history of a behaviour cost. Let $X_s = \{x_s^1, \cdots, x_s^{N_{su}}\}$ be the set of pre-defined suspicious behaviours. Then, the behaviour cost $C_k^b$ with respect to current time series of driving modes $y_k^d$ and suspicious behaviours at time step $k$ can be defined as:

  $$C_k^b = \frac{1}{\min_{i \in X_s} D\left(x_S^i, y_k^d\right) + 1} \tag{8}$$

  Three fuzzy membership functions with linguistic variable ’Normal (N)’, ’Suspicious (Su)’, and ’Worrying (W)’ are used to categorise the behaviour cost as shown in Fig. 2(b).
- **Speed:** The velocity profile of the vehicle with respect to its position or time step also needs to be investigated since it can provide the measures of the suspicious or abnormal behaviour inherently. Three functions with ’Slow (Sl)’, ’Moderate (M)’, and ’Fast (F)’ are used as shown in Fig. 2(c).
- **Environment:** The last input considers an environmental condition with a human interaction for the behaviour decision process. Depending on the traffic flow density, two functions with ’Normal traffic (Nt)’ and ’Congestion (C)’ are used as membership functions as shown in Fig. 2(d). Even though only traffic flow is used in this study, it can be easily replaced with time zone such as day/night or weekday/weekend or any other environmental aspects. This input allows for incorporating human supervision on a certain environment into decision making instead of relying only fully autonomous decision process which can be vulnerable to unexpected and dynamic environments.

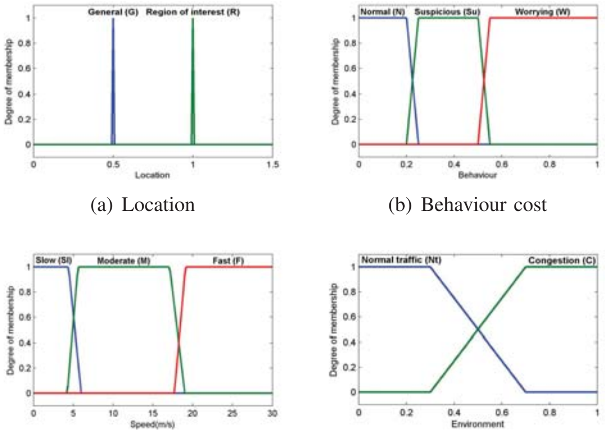
A fuzzy output for behaviour monitoring is the level of alert of each ground vehicle consisting of four membership functions with linguistic variable ’Allow’, ’Monitor’, ’Investigate’, and ’Respond’ as shown in Fig. 3.

### B. Fuzzy inference

In this study, a fuzzy inference system is designed by using a Mamdani model [11]. Inhere, expert knowledge can be expressed in a natural way using linguistic variables defined above as Table. I∼II. In the table, rules can be interpreted as:

- Rule 1: If Location is ’G’ and Behaviour is ’N’ and Speed is ’Sl’ and Environment is ’Nt’, then Alert is ’Allow’.

Note that depending on the location and environment, rules are changed slightly. For instance, if location of the vehicle is ’G’ (i.e. general area), speed ’Sl’ does not mean something significant leading to alert ’Allow’, whereas if the location is ’R’ (i.e. region of interest), slow speed or stopping of the vehicle can be identified as suspicious one (monitoring the specific facility or placing of improvised explosive devices) leading to alert ’Investigate’ as Rule 1 and 19. However, even though the location is ’R’ and speed is ’Sl’, if the environment is ’C’ (i.e. congestion), its alert level should be alleviated as Rule 20 since slow speed is more likely to be observed.



(a) Location

(b) Behaviour cost

(c) Speed

(d) Environment (Traffic flow)

Fig. 2. Membership functions for fuzzy inputs

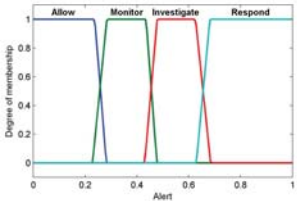


Fig. 3. Membership function for fuzzy output

### C. Deffuzification

By using input variables and defined fuzzy rules, the fuzzy outputs for all rules are then aggregated to one output fuzzy set. Finally, to obtain a crisp decision value for level of alert, a defuzzification process needs to be performed. Even though there are several algorithms for this defuzzification,

TABLE I
FUZZY RULE 1∼18: LOCATION IS ’G’ (GENERAL ROAD)

| *Rule No.* | *Behaviour* | *Speed* | *Environment* | *Alert* |
|---|---|---|---|---|
| 1 / 2: | N | Sl | Nt / C | Allow / Allow |
| 3 / 4: | N | M | Nt / C | Allow / Allow |
| 5 / 6: | N | F | Nt / C | Monitor / Investigate |
| 7 / 8: | Su | Sl | Nt / C | Monitor / Allow |
| 9 / 10: | Su | M | Nt / C | Monitor / Allow |
| 11 / 12: | Su | F | Nt / C | Investigate / Investigate |
| 13 / 14: | W | Sl | Nt / C | Investigate / Monitor |
| 15 / 16: | W | M | Nt / C | Investigate / Monitor |
| 17 / 18: | W | F | Nt / C | Respond / Respond |

TABLE II
FUZZY RULE 19∼36: LOCATION IS 'R' (REGION OF INTEREST)

| *Rule No.* | *Behaviour* | *Speed* | *Environment* | *Alert* |
|---|---|---|---|---|
| 19 / 20: | N | Sl | Nt / C | Investigate / Monitor |
| 21 / 22: | N | M | Nt / C | Allow / Allow |
| 23 / 24: | N | F | Nt / C | Investigate / Monitor |
| 25 / 26: | Su | Sl | Nt / C | Investigate / Monitor |
| 27 / 28: | Su | M | Nt / C | Monitor / Monitor |
| 29 / 30: | Su | F | Nt / C | Investigate / Investigate |
| 31 / 32: | W | Sl | Nt / C | Respond / Investigate |
| 33 / 34: | W | M | Nt / C | Investigate / Investigate |
| 35 / 36: | W | F | Nt / C | Respond / Respond |

this study uses the method of taking the centre of gravity of the aggregated output fuzzy set [12].

## IV. NUMERICAL SIMULATIONS

This section carries out a numerical simulation for both synthetic and civilian traffic scenario using the proposed fuzzy expert rule-based airborne monitoring algorithm for moving ground targets using UAVs loitering over a certain area.

### A. Synthetic scenario

Figure 4 shows the scenario description where a ground vehicle is moving around region of interest. In the map, at the southern area of a river, there is a stadium of strategic importance to be protected, which has a surrounding roadmap to be passed by a civilian ground vehicle near the base wall. A ground vehicle considered in this scenario circles clockwise round the stadium twice. During that time, the vehicle stops for ten seconds on the mid of road 3 near 420s. After that it crosses on the bridge and then travelling on the general road network. The vehicle trajectory data are used to generate virtual MTIR measurements composed of the relative range and azimuth angle adding the white Gaussian noise having the standard deviation of $(\sigma_r, \sigma_\phi) = (10m, 3deg)$.

The trajectory classification histories shows a reasonable performance capturing the turning or stopping manoeuvre timely as shown in Fig. 5 in conjunction with the trajectory estimation result with blue lines and numbered time history in Fig. 4. In this scenario, only road 3 and 4 are assumed to be of interest (i.e. location is 'R') as red line in Fig. 6(a), and suspicious behaviour $x_s$ is selected as '4 4 0 0 0 0' (which means deceleration and then stopping) to detect the vehicle which stops around stadium suspiciously. In addition, the size of driving mode history $y^D$ is set to $N_{sm} = 6$ which is the same as that of $x_s$.

Figure 6 shows the fuzzy rule-based decision making result including location information, behaviour cost and speed of the vehicle. Note that even if $y^D$ and $x_s$ are totally different, since edit distance $D$ between them is six in this case, the lowest behaviour cost would be 1/7 instead of zero according to Eq. (8) as shown in Fig. 6(b). In normal traffic shown as blue line in Fig. 6(d), level of alert has high value when the location is 'R', behaviour cost is high as well as speed is slow. Besides, if there is congestion in the traffic, the effect of the behaviour cost and slow speed on level of alert is reduced as red line in Fig. 6(d) since those conditions are more likely to happen due to the congestion.

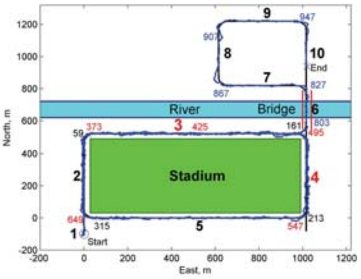


Fig. 4. Trajectory estimation for synthetic scenario

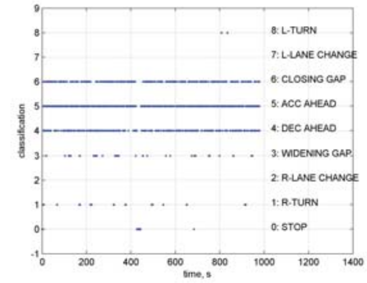
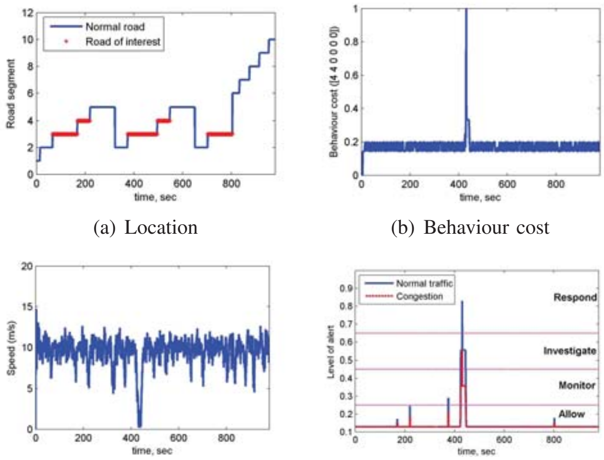


Fig. 5. Trajectory classification for synthetic scenario



(a) Location (b) Behaviour cost

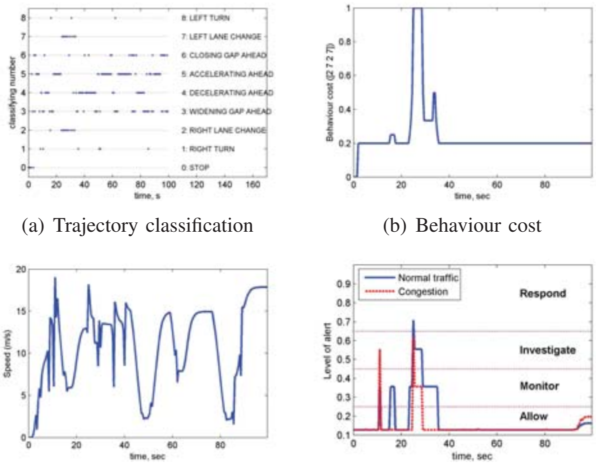
(c) Speed (d) Level of alert

Fig. 6. Fuzzy rule-based decision making results for synthetic scenario

### B. Civilian traffic scenario

The ground target trajectory is obtained from S-Paramics [13] traffic model of Devizes map in the UK at 2 Hz as shown in Fig. 7. Figure 8(a) shows trajectory estimation result of a given S-Paramics data with frequent lane changes inserted artificially (to generate suspicious behaviour) as shown in Fig. 8(b). This manoeuvre is called weaving or evasive, and can be viewed as one of the most dangerous behaviours in civilian traffic. In this scenario, every road is assumed to be general road (i.e. location is ’G’ only), and suspicious behaviour $x_s$ is selected as ’2 7 2 7’ and ’7 2 7 2’ (2: right lane change and 7: left lane change) to detect evasive manoeuvre.

Figure 9 shows the fuzzy rule-based decision making result including behaviour cost with trajectory classification and speed of the vehicle. In normal traffic shown as blue line in Fig. 9(d), level of alert has high value when the behaviour cost is high (which means evasive manoeuvre is likely to be happening) around 20∼40 seconds or velocity is fast around 10 second. In case there is congestion as red line in Fig. 9(d), although level of alert shows the same tendency, the effect of the behaviour cost is reduced as frequent lane change is more likely to happen due to the congestion. On the contrary, the effect of fast speed is enhanced around 10 second since fast ground vehicle in the congested traffic could be regarded as dangerous one.



(a) Trajectory classification

(b) Behaviour cost

(c) Speed

(d) Level of alert

Fig. 9. Fuzzy rule-based decision making result for civilian traffic scenario

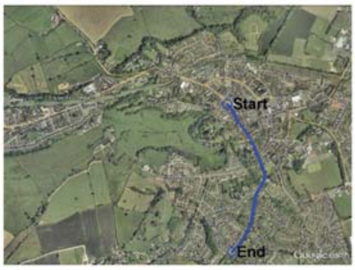
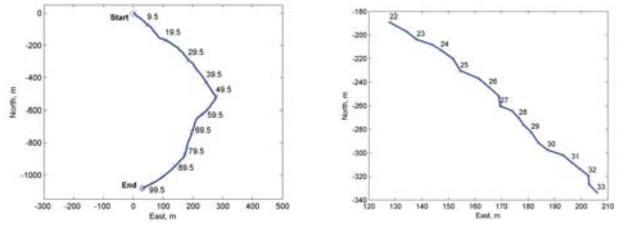


Fig. 7. Trajectory of a ground vehicle within the Devizes road network with GIS satellite data overlaid thanks to Google earth



(a) Entire trajectory

(b) Evasive manoeuvre part

Fig. 8. Trajectory estimation result and artificial evasive manoeuvre

## V. Conclusions

This paper proposed a fuzzy expert rule-based airborne monitoring methodology of ground vehicle behaviour to identify suspicious or abnormal behaviour considering spatiotemporal environment factors as well as behaviour itself. Numerical simulation results using synthetic scenario and realistic car trajectory data showed the feasibility of the proposed approach successfully providing a recommended level of alert. The study could be applied to various scenarios in view of both military and civil applications: monitoring urban/rural area of interest, detecting unknown intent of terrorists, providing a protective surveillance around military facilities, and enhancing situational awareness of traffic movements both on land and at sea. As a future work, additional relevant aspects of behaviour will be considered as fuzzy inputs such as cultural background related to driving habits and deviation from general behaviour.

## References

[1] J. Will, L. Peel, and C. Claxton. Fast maritime anomaly detection using kd-tree gaussian processes. In *2nd IMA Mathematics in Defence*, Defence Academy, Shrivenham, UK, 2011.

[2] F. Johansson and G. Falkman. Detection of vessel anomalies - a bayesian network approach. In *3rd International Conference on Intelligent Sensors, Sensor Networks and Information*, 2007.

[3] C.C. Loy, T. Xiang, and S. Gong. Modelling multi-object activity by gaussian processes. In *British Machine Vision Conference*, 2009.

[4] S. Srivastava, K.K. Ng, and E.J. Delp. Co-ordinate mapping and analysis of vehicle trajectory for anomaly detection. In *IEEE International Conference on Multimedia and Expo (ICME)*, July 2011.

[5] R. Fraile and S.J. Maybank. Vehicle trajectory approximation and classification. In *British machine Vision Conference*, 1998.

[6] S. Kim, R.W. Zbikowski, A. Tsourdos, and B.A. White. Airborne monitoring of ground traffic behaviour for hidden threat assessment. In *13th International Conference on Information Fusion*, Endinburgh, UK, July 2010.

[7] S. Kim, A. Tsourdos, and B.A. White. Behaviour monitoring on ground vehicles by uavs. In *Mathematics in Defence*, Swindon, UK, 2011.

[8] S. Kim, H. Oh, and A. Tsourdos. Nonlinear model predictive coordinated standoff tracking of moving ground vehicle. *AIAA Guidance, Navigation, and Control Conference, Potland, Oregon, USA*, 2011.

[9] C. F. Lin and A. G. Ulsoy. Calculation of the time to lane crossing and analysis of its frequency distribution. In *American Control Conference*, 1995.

[10] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, San diego, USA, 2006.

[11] E.H Mamdani. Applications of fuzzy logic to approximate reasoning using linguistic synthesis. *IEEE Transactions on Computers*, 26(12):1182–1191, 1977.

[12] H. Hellendoorn D. Driankov and M. Reinfrank. *An Introduction to Fuzzy Control*. New York: Springer-Verlag, 1993.

[13] SIAS Limited. S-paramics software. http://www.sias.com, Jan 2011.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Processing of Bars-Converging Joint and Engineering Application

Wanjing Luo
School of Mechanical Engineering
Tianjin University
Tianjin China
E-mail: luo494115063@163.com

Jiangping Mei
School of Mechanical Engineering
Tianjin University,
Tianjin China
E-mail: ppm@tju.edu.cn

***Abstract*-The complex structure of the shaped surface formed by a kind of bars-converging single-layer reticulated shell joints is analysed, and the characteristic of higher precision in production is considered, the patter decomposing – patter processing - patter assembling –patter casting joint processing flow is put forward. According to the joint design data, the joint spatial axis model is built, the spatial orientation description method is proposed, the developed SolidWorks 3D drawing software is used to import the design data to visualize the joints in 3D model. According to the joints splitting and assembling principle, the mathematical model of joint processing is derived, based on the mathematical model, the joint processing equipment based on the TriVariant series robots is developed, the real time control of the joint NC machining system is implemented. The results of this research have been verified in the large complex steel structure building of the Sun Valley for Shanghai World Expo and the roof of Star Mall in Shenyang, now the research is being used in the Cell Wall of the new natural museum in Shanghai which is under construction.**

***Keywords-reticulated shell joint; 3D model reappearance; splitting and assembling; TriVariant series robots; engineering application***

## I. INTRODUTION

Spatial reticulated shell structure is a kind of architectural structure connected by spatial bars with multiple joints, in order to adapt to the changing space angles, a type of bars-converging joint for single-layer reticulated shell appeared[1].As shown in Fig. 1,this type of joint has many advantages, such as the shape can be changed with the overall surface of the building, and can be connected easily with bars with various special-shaped sections (such as triangle, rectangle, trapezoid), so it can constitute a variety of fashionable spatial curved surface of architectural modeling, it becomes an important development trend of the highly decorative spatial reticulated shell structure building[2] .

In order to satisfy the whole smooth surface and continuous changes, joints for reticulated shells have different shapes and complex space angles. Through topological mesh the single layer netrack is composed by triangular grids, each grid is mutual restricted, each joint has at least four brackets, each bracket is connected with other components[3]. Any bracket has little error, it will affect the installation of the components, so the joint itself requires high machining precision. At present, at home and abroad the production methods of the bars-converging single-layer reticulated shell joint can be put into the following two categories:1）CNC cutting- assembling-processing, this facture process of joints has been applied in the latticed dome of the London Shopping Center called Beispiel Westlife (Fig. 2)by SEELE, a German company. The basic process is firstly steel plate cutting, edge processing and forming different units by pressure,

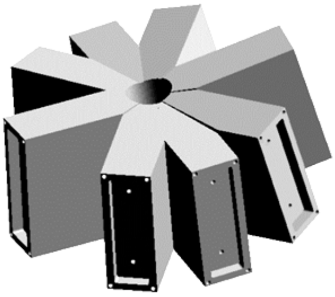


Figure 1. Bars-converging reticulated shell joint

Figure 2. Beispiel Westlife Shopping Center

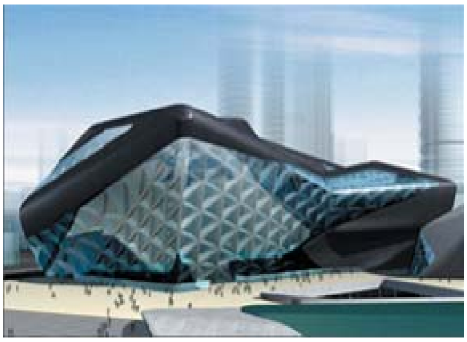


Figure 3. Guangzhou Opera House

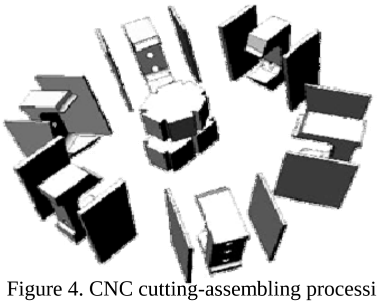


Figure 4. CNC cutting-assembling processing

then the next step is assembling, using clamping fixture for welding, finally the last step is completing the machining of surface / hole combined with bars using 5 axis machining center, as shown in Fig. 4. As the reticulated shell structures often contain thousands of different geometry joints, and each joint is divided into several units, it requires a lot of pressure molding and welding clamping fixture, so it results in high difficulty in joints production，high complexity in information management ， high cost, low processing efficiency. In addition, it is difficult to guarantee the manufacturing accuracy.2 ） Patter processing-casting-processing, this process of joints has been applied in the national stadium Bird’s Nest in Beijing for 2008 Olympic Games and Guangzhou Opera House. First the joint mould is made, then the next step is getting the joints semi-finished products by using precision casting, finally the last step is completing the machining of surface / hole combined with bars using 5 axis machining center[4]. Compared with the first method, this process needs mold making for each joint, if there is no quick, low cost molding technology and equipment available, this process can also lead to low efficiency, high cost, poor precision. In the Guangzhou Opera House project (Fig. 3), the sand casting is used in the step of mold making, as it needs a lot of wood patterns, the working efficiency is low and the production cost is high. It also leads to the difficulty in mold making and complex angles formation. So this research adopts the lost foam casting, which can raise the efficiency, lower the cost and guarantee the required precision.

In order to solve the engineering difficulties and make up for the previous study defects, this paper presents a joint processing method which includes deepen design- patter decomposing – patter processing - patter assembling–patter

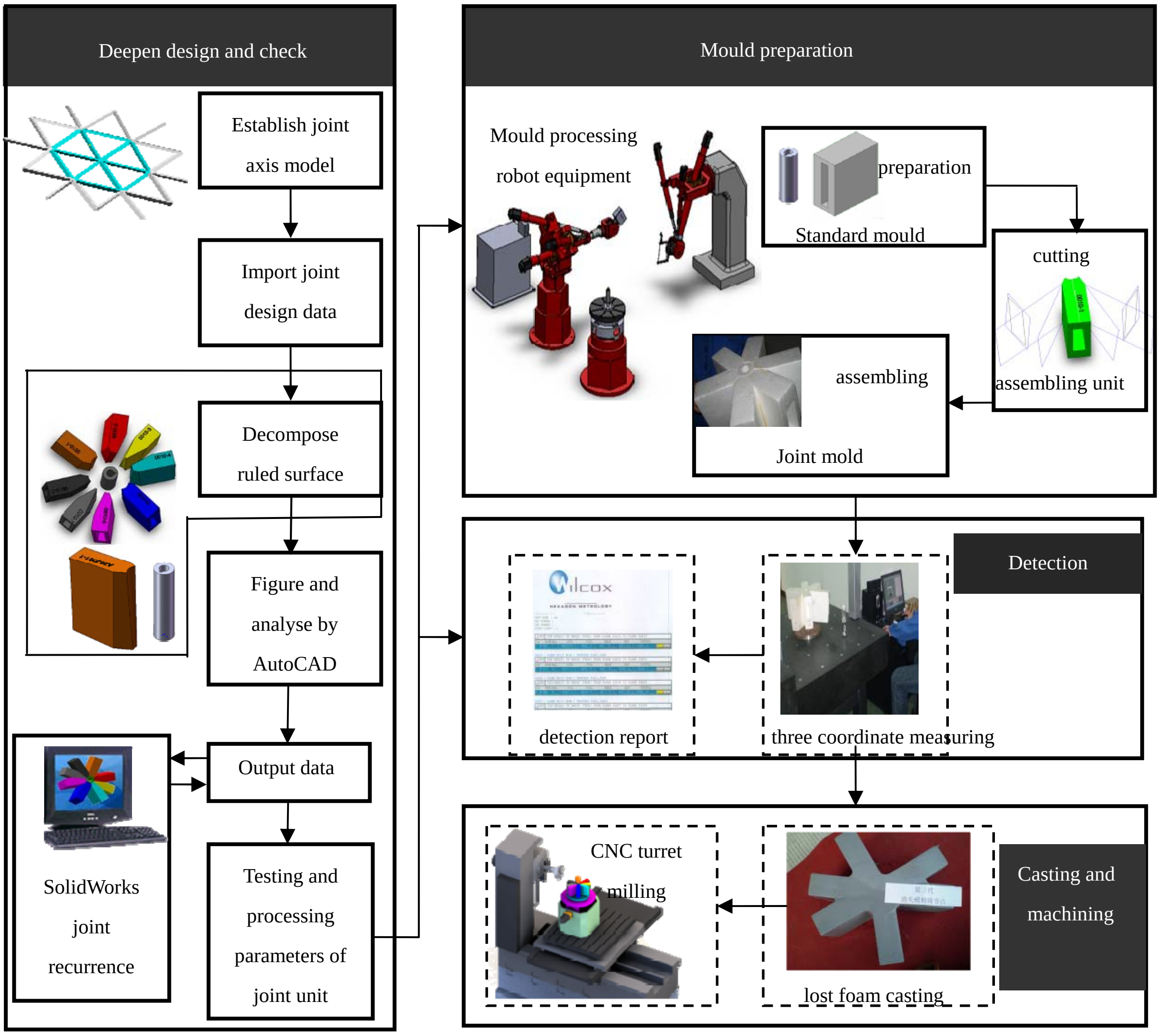


Figure 5. Joint CNC processing flow

casting- machining. At first the key point coordinates of the bracket is extracted from reticulated shell structure axis model, based on the developed SolidWorks software the joints are visualized in 3D model and virtual preassembled to verify the joint deepen data, then according to the principle of joint splitting - assembling, the method of ruled surface for parting and positioning is used, with the 5 degree of freedom TriVariant series hybrid robots as the joint mold processing equipment, the cutting and assembling of joint mould is completed.

## II. JOINT PROCESSSING FLOW

According to the bars-converging single-layer reticulated shell joint processing project, this paper proposed a joint NC machining process based on the multi functional robot.

Steps are as follows: At first the joints are deepen designed, the axis model of bars-converging single-layer reticulated shell structure is set up, the initial designed coordinate parameters and dimension parameters of all joints and bars are input into the computer, and each joint is divided into a number of tectonic units, AutoCAD is used to figure and analyse, the size data and processing parameters of each unit are gotten, by the programming language VC++ , Solidworks 3D drawing software is developed, every joint in 3D model is visualized based on the size data and processing parameters, the correctness of joint deepen data is verified. Then the size data and its processing parameters after testing of each joint structure unit are input into the joint processing robot CNC equipment, the robot is controlled to complete the cutting and assembling of joint mold units, three coordinate measuring instrument is utilized to detect the size and accuracy of the joint mold, the detection report is gained. Finally the qualified joint mold after measuring and testing is obtained, through the lost foam casting process, the molten steel is casted, through machinery processing the prepared products are obtained by using the CNC turret milling.

The process for joint products has many advantages, such as good mechanical properties, low production cost, and it can guarantee the required precision, the joint processing flow is shown in Fig. 5.

## III. JOINT SPACE ATTITUDE

In order to unified describe the spatial different joints, the characteristic that a large number of joints are similar but not identical is used. For every different joint, the joint number of unit N and the pitch angle $\alpha_i$ 、 the torsion angle $\beta_i$ and the angle of circumference between units $\theta_{i,i+1}$ (Where i is the joint unit number)and bracket unit length L and width B are used to describe the joint spatial attitude, as shown in Fig. 6, But in actual machining, in order to be measured and expressed easily, this study uses joint end face point coordinates to describe the spatial attitude. According to the joint spatial attitude, the coordinate system $O-xyz$ is established ,the point $O$ is the point of intersection in the theory diagram model, as $O$ is the origin, the axis of the center column is axis z, upward is positive, according to the clockwise direction ( z upward direction ) to define each joint unit NT-1、NT-2、……、NT-i（determined by the unit quantity）,the bottom center line of the numbered NT-1 unit is the axis x, the direction departs from the central cylindrical axis and points outwards. In accordance with the right-hand rule in Cartesian coordinate system to determine the y axis direction, the joint model is shown in Fig. 7. According to the joint initial designed space attitude angles and the bracket unit parameter data, the three vertices coordinates on every rectangular cross section of the numbered joint in the coordinate system $O-xyz$ are figured out(as $P_{11}, P_{12}, P_{13}$ in NT-1）,a total of 3i vertices coordinates can express the positional relationship between each unit and the space angle relation in the same joint.

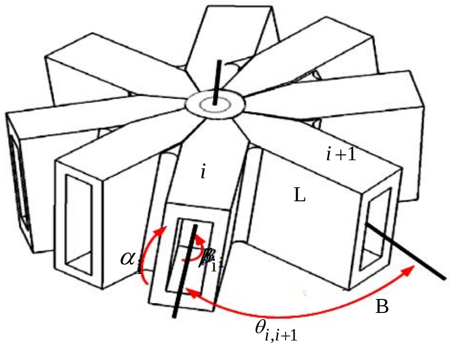


Figure 6. Joint space angle attitude

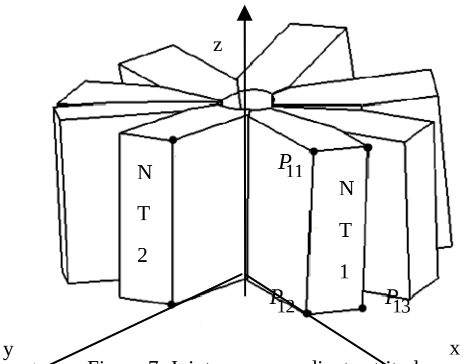


Figure 7. Joint space coordinate attitude

## IV. JOINT THREE-DIMENSIONAL REAPPEARANCE

In order to meet the requirement of high precision of joints, it needs to test and check the deepen design data, the programming language VC++ is used to develop SolidWorks 3D drawing software. According to the characteristics that the joints are similar but not identical, by dimension driving method, the joint model structure is maintained unchanged, all the dimension which affect the model structure are defined as the dimension variables, by means of giving these dimensions variables with different values, it is able to obtain a series of a different sort of junction joints with identical structure and different sizes.

The concrete steps are as followed: In Visual C++ 6 integrated developed environment for programming, through the dynamic link library DLL and SolidWorks API functions

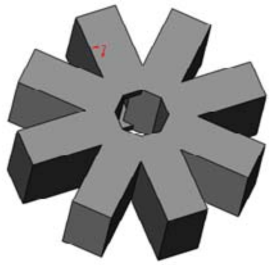


Figure 8. Three-dimensional reappearance results

to develop SolidWorks software, then the data tables which contain batch joint features are imported into the 3D drawing software automatically, automatic batch drawing is carried out, the joint model is reconstructed, at last virtually the reticulated shell structure model is assembled in advance, the initial visual inspection of joint and structure is realized. Because of the visualization and intuitionism of the model, the joint data with obvious errors can be found, reappearance results are shown in Fig. 8.

## V. JOINT SPLITTING-ASSEMBLING PRINCIPLE

Joint splitting - assembling principle which is adopted in this paper is that the joint are split into assembling units and the cylinder which is connected to each unit, the split surface between the unit is the angle bisector plane, the robot is used to cut the standard module into the assembling units, and with circular arc and plane surface to position each unit and the units are combined into complete joint mold.

The processing of each unit of joints is the cutting of the angle bisector plane and cylindrical ruled surface of the standard foam mold, the key technology is the transformation from the ruled surface mathematical expression in space to the trajectory data of the processing equipment. The overall principle of robotic cutting: Through the geometric relationship, the space equation of processing ruled surface in the joint coordinate system can be obtained, the conversion between joint coordinate system, workpiece coordinate system and processing device coordinate system is created, equipment's trajectory control data is gotten by post processing ,the end-effector space pose when cutting ruled surface can be obtained, the track pulse is input into the servo control system, the robot servo control is realized.

Standard module cutting mathematical principle: the center cylindrical surface equation and each unit's two plane equations are united, two arc $\alpha$ and $\beta$ equations can be obtained. Similarly the space plane equations are united, the angle bisector planes $\lambda_1$ and $\lambda_2$ equations can be obtained. Considering that the joint cutting data is easy to use and express, in each unit corresponding vertex the coordinate system $O'-x'y'z'$ is established, as shown in Fig. 9. According to the coordinate system conversion, the rotation matrix can be expressed as ***R***=

$$\begin{bmatrix} \sin\theta_1 \sin\theta_2 & -\cos\theta_2 & \cos\theta_1 \sin\theta_2 \\ \cos\theta_1 & 0 & -\sin\theta_1 \\ \sin\theta_1 \cos\theta_2 & \sin\theta_2 & \cos\theta_1 \cos\theta_2 \end{bmatrix}$$

The $\theta_1$ and $\theta_2$ are the attitude angle of the joint unit. Coordinate translation vector $P = \overrightarrow{OP_{4-2}} = \left[x_0, y_0, z_0\right]^T$ .The coordinate transformation equation is T

$$T = \begin{bmatrix} R & P \\ 0 & 1 \end{bmatrix}$$

So the relation is $T^{-1}\begin{bmatrix} x & y & z \end{bmatrix}^T = \begin{bmatrix} x' & y' & z' \end{bmatrix}^T$

Similarly, with the intersection method the equation of cylindrical surface and each unit plane in a new coordinate system $O'-x'y'z'$ can be obtained, then the equation of intersecting lines between upside and downside section and the cylinder surface $\alpha'$ and $\beta'$ can be obtained, and the equation of $\lambda_1{}'$ and $\lambda_2{}'$ which is the angle bisector planes by adjacent bracket units intersecting in the new coordinate system $O'-x'y'z'$ are achieved. Thus the spatial vector and the position which the robot end effector needs to arrive at in the unit coordinate system $O'-x'y'z'$ can be expressed.

In actual production, in order to control the cutting trajectory of the cutting robot, it needs to transform the measurement of joint mold unit ruled surface for processing (two spatial surfaces and a cylindrical surface) in the unit coordinate system $O'-x'y'z'$ to the same reference coordinate system with the cutting robot, the space relation between the two coordinate systems is shown in Fig. 10, the rotation matrix is R,

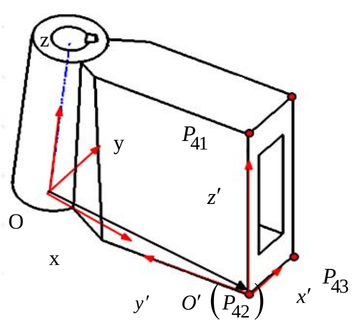


Figure 9. Joint coordinate system conversion

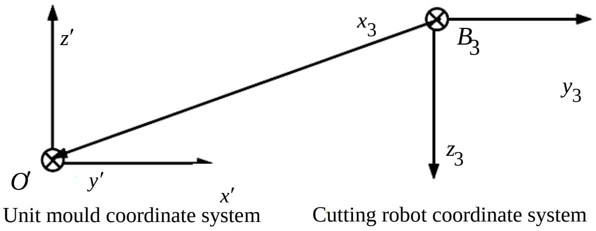


Figure 10. Relation between unit mould coordinate system and cutting robot reference coordinate system

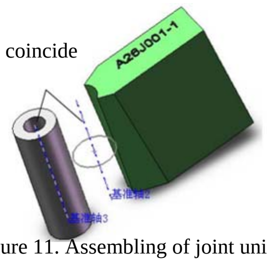


Figure 11. Assembling of joint unit mold

$$R = R(x_3, 180^\circ)R(z_{3'}, -90^\circ) = \begin{bmatrix} 0 & 0 & -1 \\ 0 & -1 & 0 \\ -1 & 0 & 0 \end{bmatrix}$$

Translation vector $P = \overrightarrow{B_3O'}\,BO$, the space vector $L_i$ and position $E_i$ which the end executor needs to reach in the cutting robot reference coordinate system $B_3 - x_3y_3z_3$ respectively are $L_i = RL_i', E_i = RE_i' + P, (i = 1, 2 \ldots n)$ Where n is the number of sampling points, $L_i'$ and $E_i'$ respectively are the space vector and the end position of the end-effector needs to arrive at in the unit coordinate system $O' - x'y'z'$ .

Through this two quantities and the robot inverse kinematics model, the relevant parameters of the branched chains and the joint angle values of TriVariant-A cutting robot can be obtained for completing processing, thereby the robot control during the cutting process is realized[8] .

Unit assembling process is essentially the process that the TriVariant-B robot module terminal control unit axis is coincident with the center cylinder axis by means of the rotary with the single degree of freedom platform to the assembling area, as is shown in Fig. 11. When the center cylinder axis of joint mold is expressed as $S = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T$ in the coordinate system $O$ - $xyz$ ,the vector $\overrightarrow{P_{42}P_{41}}$ is the robot target vector, point $P_{42}$ namely is the target point. Similar with the cutting process of mold unit, in the assembling process, each vector is required to be measured in the robot coordinate system, in order to convert to the robot recognizable data, through the inverse kinematics solution for assembling robot, as is similar with the coordinate system transformation in the cutting process, so is not discussed here.

The joints splitting -cutting - assembling process method not only realizes joint batch production, but also transfers the processing difficulty to the digital cutting of various unit's bisecting planes and curved surfaces, the arc and planes are used for positioning to ensure the precision of the joint shape. Finally through the lost foam casting the joint processing is finished.

## VI. CONCLUSION

This paper presents a method based on the basic idea of joint NC processing process flow which includes deepen design- patter decomposing – patter processing - patter assembling –patter casting- machining, the developed SolidWorks software is used to carry out the batch joints three-dimensional reappearance, and to test the correctness of the deepen data, the joint mold cutting- assembling model is established, on the basis of the mathematical model the

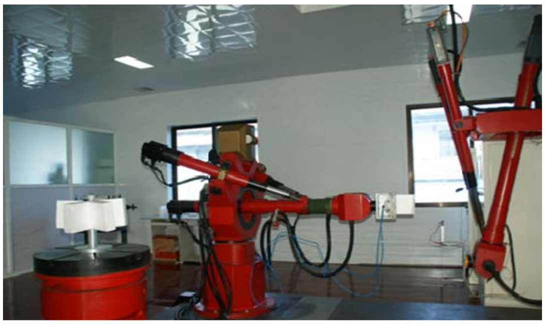


Figure 12. Robot equipment for processing joint mold

Figure 13. Shanghai World Expo Sun Valley

Figure 14. Star Mall in Shenyang

Figure 15. The design sketch of natural museum in Shanghai

control of the robot end-effector pose is realized. The robot NC equipment(Fig. 12) for processing the joint mold is used to complete cutting and assembling. The novelties of this method are as followed:1)The application of the TriVariant series robots in the mold cutting and assembling; 2)The utilization of lost foam casting in producing joints;3)The proposal of the whole process flow for bars-converging joints. Based on the multi functional robot, the digital joint manufacturing process has been successfully applied in the Sun Valley in 2010 [8] (Fig. 13)and the roof structure of Star Mall in Shenyang in 2011(Fig. 14). In addition now the Cell Wall of the new natural museum in Shanghai is under construction by using the results of this research (Fig. 15). It proves that this joint processing method not only reduces the joint production cost, but also improves the production efficiency, and it creates favorable conditions for the development of spatial reticulated shell structure technology.

## ACKNOWLEGMENT

This research is sponsored by the 863 High-Tech Development Program (2009AA04Z219). We would like to thank the anonymous reviewers for their useful comments.

## REFERENCES

[1] P J Trebilcock, Building design using culd formed steel, An Architect'B Guide, UK: The Steel Construction Institutu, 1994

[2] Bao Wei, Xing Litao, Qiu Jianhui, The use of cast steel in steel structure, Advanced Materials Research, 2011, 183~185

[3] Mei Jiangping, Tan Yang, Wang Lan, et al, Robot based manufacturing technology of cast-steel joints of Sunny Valley at Shanghai World Expo Axis, Advanced Materials Research, 2011, 385~389

[4]Dai Chun, Super scale construction in 2010 Shanghai Expo Axis design and construction, Times Architecture, 2010, (3): 56~59

[5]Sheng Linfeng, Sun Valley at Shanghai World Expo Axis monolayer steel structure joint of reticulated shells production technology research, Building Construction, 2009, 31(12): 1015~1018

[6] Wang Fan, Zhang Zhigang, Luo Min, Study on the overall stability of single-layer spherical latticed shell structures and its applications in engineering, Advanced Materials Research, 2011, 137~142

[7]Gao Ben, Research on patternmaking technology and equipments of cast steel bars-converging valve-type joints, [Master degree thesis], Tianjin; Tianjin University，2011

[8] Huang T, Li M, Zhao X Y, et al, Kinematic design of a reconfigurable miniature parallel kinematic machine, Chinese Journal of Mechanical Engineering, 2003, 16(1): 79~82

[9] Ming Wen, Xin Fang Wang, Zi Chen Deng, The study on performance of single-layer cylinder shells with semi-rigid bolt-ball joints, Advanced Materials Research, 2011, 243~249

[10] Ramme, Reitinger R, Shape optimization of shell structures, Bulletin of the International Association for Shell and Spatial Structures, 1993, 34(2): 103~121

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Modelling of a Pan and Tilt Servo System

Samuel Sharp and Adam Wicks
*Control Systems Department*
*Chess Dynamics Ltd.*
*Horsham, United Kingdom*
{Samuel.Sharp & Adam.Wicks}@chess-dynamics.com

Andrzej Ordys and Gordana Collier
*Faculty of Science, Engineering and Computing*
*Kingston University*
*Kingston upon Thames, United Kingdom*
{A.Ordys & Gordana.Collier}@kingston.ac.uk

***Abstract*—Two-axis pan and tilt systems are widely used in surveillance applications for high accuracy positioning of sensor payloads such as cameras and laser pointers. In order to develop advanced control algorithms to improve the performance of these systems, a model of the system must be developed. This model should include the dynamics of the system to include effects such as compliance and account for friction effects in the drive. This paper discusses the development of the overall model of the system using National Instruments LabVIEW, and in particular, the models for friction and the drive train that will be used.**

***Keywords-Servo systems; Modelling; Software tools; Harmonic drives***

## I. INTRODUCTION

Chess Dynamics Ltd. is a supplier of high performance Radar and Electro-Optic systems for the defence industry. It designs and manufactures precision two-axis pan & tilt camera systems for land, naval and airborne applications.

The performance demands of these systems are very high; requiring high pointing resolution and high positional accuracy with fast tracking speeds and smooth low speed control.

Research and development is being conducted within the company through a Knowledge Transfer Partnership (KTP) scheme with Kingston University to develop advanced control methods that will improve the accuracy and tracking performance of the company's products.

Initially, simulation models of the system will be developed to allow advanced control and estimation algorithms to be evaluated and tested before being implemented into the actual system. Subsequent stages of the project will develop estimation algorithms to improve the quality of sensor measurements and then control algorithms will be developed and evaluated as to their suitability for increasing system performance.

This paper discusses the models of the system that have been developed and their implementation in National Instruments' LabVIEW software using the Control Design and Simulation toolkit.

Finally, simulation results are compared to results obtained from a representative experimental test rig.

## II. SYSTEM MODELLING

In this section, we will begin by introducing the system to be modelled before going on to discuss the models for the individual components of the system. Finally, the implementation of the overall system model in different forms using LabVIEW is discussed.

### *A. System Overview*

The system to be modelled, shown in Figure 1, is a compact two-axis pan and tilt system designed to carry an optical sensor payload of up to 25kg.

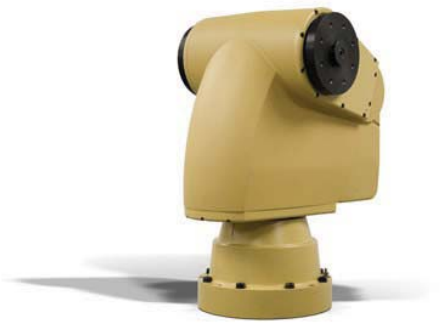


Figure 1 - Cobra Product [1]

The system consists of two identical drive assemblies, with integrated encoders, mounted on the azimuth and elevation axes. Housed within the body of the system are two current-mode servo amplifiers to drive each axis and the main control board to interface to the amplifiers and the outside world via a

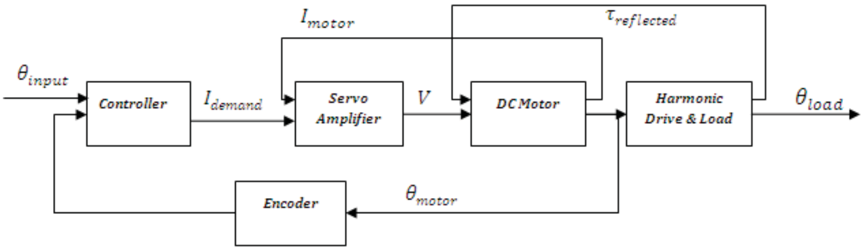


Figure 2 - Single Axis Servo System Control Block Diagram

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

serial interface. The software on this board implements the digital controller that controls the position and rate of the platform.

Initial investigations will consider a single axis only as the dynamics of each axis will be similar due to having the same drive components and electronics. It also provides a simplified system that can be easily tested and used for validating the models. It is, however, understood that the azimuth and elevation dynamics are not independent (for example, the azimuth inertial load will change as the elevation axis rotates).

The overall block diagram showing the feedback relationships between the main components of the system for a single axis is shown in Figure 2.

The details of modelling the components of the servo system from the servo amplifier to the load will now be discussed.

### B. Servo Amplifier

The servo amplifier used is a current-mode linear DC amplifier capable of providing up to 2A of current. The DC relationship was found by experimentation and connecting the amplifier to a large resistive load and measuring the current across it. This provided information on the DC Gain but did not give any information about higher frequency dynamics.

The experiment was conducted using LabVIEW, a CompactRIO and a Fluke 8808a Digital Multimeter (DMM) as shown in Figure 3. The 8808a DMM was connected to the CompactRIO via the serial interface, using plug and play instrument drivers to configure and read the current measurements. Meanwhile, the CompactRIO fed an analogue voltage to drive the servo amplifier across its full range.

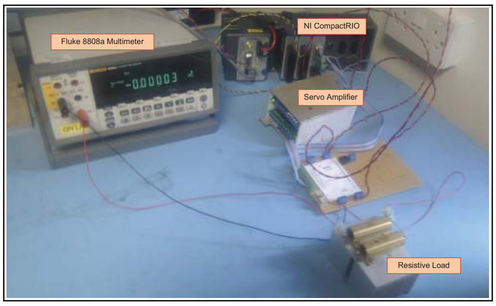


Figure 3 - Servo Amplifier Characterisation Experiment Setup

The servo amplifier was modelled as a proportional current feedback loop with a voltage saturation to represent the maximum voltage supply that can be supplied to the motor.

### C. DC Motor

The DC Motor used in the system is a high performance brushed DC motor. The motor is modelled as a two-input (voltage and load torque) and single output (motor shaft velocity) system. The DC motor model consists of two sub-processes, an electrical and mechanical, as detailed in [1].

The electrical sub-process is described by Equation 1, where $I_a$ is the armature current, $U_a$is the armature voltage, $L_a$ is the armature inductance, $R_a$ is the armature resistance, ω is the angular velocity and Ψ represents the magnetic flux.

$$\begin{aligned} I_a(t) = \frac{1}{L_a}\int_0^t \big(U_a(t) - R_a I_a(t) \\ - \Psi\omega(\mathrm{t})\big)\mathrm{dt} + \mathrm{I_a}(0) \end{aligned} \tag{1}$$

The mechanical process is described by Equation 2, where $J$ is the inertia of the motor, $K_t$ is the motor torque constant from the electrical sub-process, $M_l$ is the load torque and $M_\mathrm{f}$ is the friction torque in the motor.

$$\omega(\mathrm{t}) = \frac{1}{J}\int_0^t (K_t I_a - M_l - M_\mathrm{f})\mathrm{dt} + \omega(0) \tag{2}$$

### D. Harmonic Drive

The gearing used in the system is a harmonic drive, or strain wave gear, developed in the 1950's for the aerospace industry. It is a compact, lightweight gearbox with high gear ratios and zero backlash. Disadvantages of this type of gearing include high static friction, nonlinear compliance and high flexibility. Considerable research has been conducted into the modelling, friction compensation and control of harmonic drives due to their favourable transmission attributes [3][4][5].

From the research conducted in this field, a number of different models have been proposed for harmonic drive gears. Some of these models will now be discussed.

#### 1) Hashimoto & Kiyosawa

The dynamic model proposed in [6] considers two primary nonlinear effects of harmonic drives, Coulomb friction ($\tau_f$) and a linear stiffening spring ($K$). The dynamic equations for the model are shown by Equation 3.

$$\begin{gathered} J_l\ddot{\theta}_l - K\left(\frac{\theta_m}{N} - \theta_l\right) = 0 \\ NJ_m\ddot{\theta}_m + K\left(\frac{\theta_m}{N} - \theta_l\right) + N\tau_f = N\tau_m \\ \tau_l = K\left(\frac{\theta_m}{N} - \theta_l\right) \end{gathered} \tag{3}$$

#### 2) Ghorbel & Gandhi

The models proposed in [7] and [8] are significantly more complex and take into account additional nonlinear effects and more advanced friction models. The model, represented by $\tau_{fr}$ and shown in Equation 4, uses a LuGre friction model that takes into account Stribeck, Dahl, viscous and Coulomb friction effects and an additional position dependent component. Kinematic error in the harmonic drive is represented by $\tilde{\theta}_p$ as a function of the motor position and an additional damping term, caused by the stiffness in the drive, is represented by $B_{sp}$.

$$
\begin{aligned}
J_m\ddot{\theta}_m + K\left(\theta_l - \frac{\theta_m}{N} + \tilde{\theta}_p\right)\left[-\frac{1}{N} + \frac{d\tilde{\theta}_p}{d\theta_m}\right] \\
+ B_{sp}\left(\dot{\theta}_l - \frac{\dot{\theta}_m}{N} \\
+ \dot{\tilde{\theta}}_p\right)\left[-\frac{1}{N} + \frac{d\tilde{\theta}_p}{d\theta_m}\right] + \tau_{fr} \\
= \tau_m
\end{aligned}
\tag{4}
$$

$$
\begin{aligned}
J_l\ddot{\theta}_l + K\left(\theta_l - \frac{\theta_m}{N} + \tilde{\theta}_p\right) + B_l\dot{\theta}_l \\
+ B_{sp}\left(\dot{\theta}_l - \frac{\dot{\theta}_m}{N} + \dot{\tilde{\theta}}_p\right) = 0
\end{aligned}
$$

### 3) *Ghorbel & Dhaouadi*

The model suggested in [9] uses a nonlinear function, represented by $\tau(\dot{\theta}, \theta)$, to model a nonlinear torsional spring with nonlinear viscous damping and a function $\tau_f$ to represent a dry friction torque that includes the static friction, as shown in Equation 5.

$$
\begin{gathered}
J_m\ddot{\theta}_m + B_m\dot{\theta}_m + \tau_f + \frac{\tau(\dot{\theta}, \theta)}{N} = \tau_m \\
J_l\ddot{\theta}_l + B_l\dot{\theta}_l + \tau_l - \tau(\dot{\theta}, \theta) = 0 \\
\tau(\dot{\theta}, \theta) = g^{-1}\left[\frac{\dot{\theta}}{h(\theta)}\right] + f(\theta) \\
\theta = \frac{\theta_m}{N} - \theta_l
\end{gathered}
\tag{5}
$$

### 4) *Taghirad*

A more in-depth analysis of the friction effects in harmonic drives are proposed in [3]. The frictional components of the wave generator bearing ($\tau_{f1}$), gear meshing ($\tau_{f2}$) and output bearing are all considered separately in a friction model comprising velocity-dependant Coulomb, viscous and Stribeck friction. $T_{meas}$ represents the compliance model between the motor and load side positions. In comparison to the other models discussed, the motor and load inertias are considered as an overall effective inertia, $J_{eff}$. The overall model of the harmonic drive was therefore as per Equation 6.

$$
\begin{aligned}
K_m i - \frac{1}{N}(T_{meas}) = J_{eff}\ddot{\theta}_{wg} + (\tau_{fm} + \tau_{f1} \\
+ \tau_{f2})
\end{aligned}
\tag{6}
$$

### 5) *Tjahjowidodo, Al-Bender & Van Brussel*

The work carried out in [10] proposed a model for the torsional compliance of the harmonic drive and then performed identification on the drives in two test setups using a low and high-torque class of harmonic drive. The harmonic drive model was then described by Equation 7, with equations for the input (motor) side and load side dynamics.

$$
\begin{gathered}
J_m\ddot{\theta}_m + \frac{\left[T_b(\Delta\tilde{\theta}) + T_h\left(\Delta\tilde{\theta}, \Delta\dot{\tilde{\theta}}\right)\right]}{N+1} = \tau_m \\
J_l\ddot{\theta}_l + T_h\left(\Delta\tilde{\theta}, \Delta\dot{\tilde{\theta}}\right) + T_f(\Delta\theta_l, \Delta\theta_l) = 0
\end{gathered}
\tag{7}
$$

In Equation 7, $\tilde{\theta}$ is the difference between the motor and load angles, $T_b$ is a third order polynomial function to represent a nonlinear spring, $T_h$ represents the torsional stiffness and $T_f$ is the bearing friction torque of the harmonic drive.

## *E. Overall System Model*

This section deals with the different representations of the overall system model built from the components previously discussed and their implementation in LabVIEW.

The initial model of the system developed encompassed the models of the DC motor and servo amplifier discussed and the harmonic drive model detailed in Section 1) of II.D. It consists of a load inertia and a linear spring to represent the compliance between the motor and load side angles. Friction in the system was considered for each of the motor and harmonic drive separately as a linear damping function. The initial overall model of the mechanical system is shown in Figure 5.

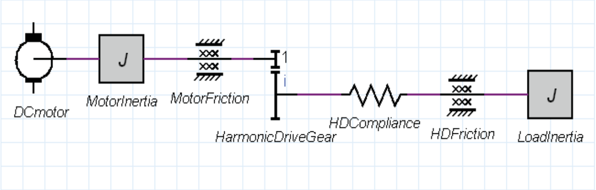


Figure 5 - Overall System Model Diagram

### 1) *Simulation Model*

Firstly, a simulation model was developed to represent the system from the dynamical equations as discussed previously. This simulation model was implemented in LabVIEW using the Control Design and Simulation toolkit as shown in Figure

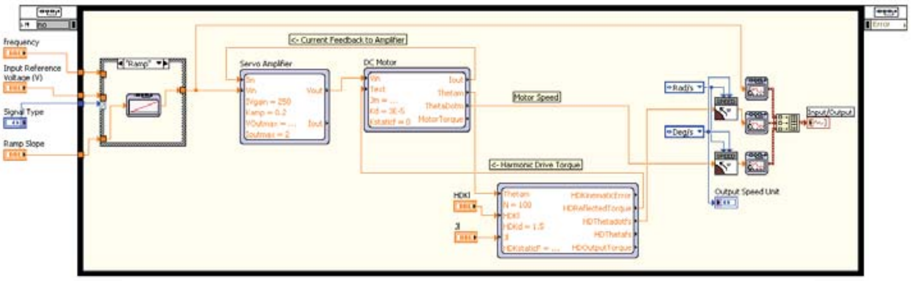


Figure 4 - Simulation Model of System in LabVIEW

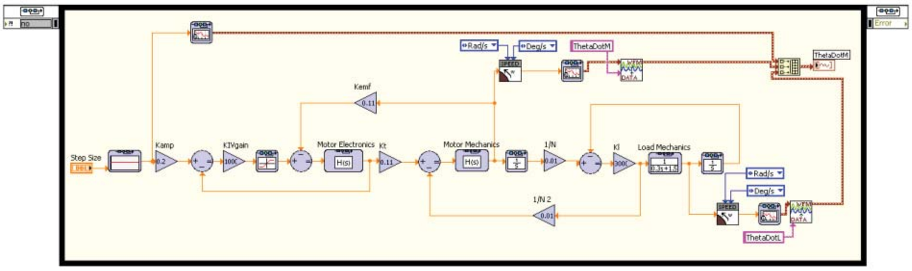


Figure 6 - Transfer Function Representation

4.

*2) Transfer Function Model*

Having a transfer function representation of the system is important to allow the use of classical design and analysis tools. From the system models, a transfer function representation using the Laplace transform was developed and implemented in LabVIEW, as shown in Figure 6.

*3) State Space Model*

The need for a state space model of the system allows for the use of modern control design and analysis tools and estimation algorithms such as the Kalman Filter rely on a state-space model of the system. They are also more suited to multiple-input multiple-output (MIMO) systems.

The state space model was implemented in LabVIEW using a MathScript node, allowing the state space equations, parameters and state space model to be defined in a textual way, as shown in Figure 7.

## III. Simulation Results

To test the correct operation of the model, simulations were run with a range of input signals. Known parameters of the system were used along with values for unknown parameters that demonstrate the operation of the model. Responses of the system to a step and sinusoidal input for the transfer function model are shown in Figure 8 and Figure 10.

Experiments were carried out for a small number of input signals to an experimental test rig containing a drive assembly of the type being modelled connected to a representative inertia for the Cobra product (including payload). This test rig will be used to validate the models and then controllers developed in the simulation will be verified using the physical test rig.

The results of running the experiments for the same input signals as shown in Figure 8 and Figure 10 are shown in Figure 9 and Figure 11.

The linear model proposed in this paper does not take into account the most important non-linear characteristic of the harmonic drive unit, the high static friction. This is shown by the discontinuity around the zero-output crossing point of the sine response, as shown in Figure 11. Current work is looking to include this effect in a future iteration of the model.

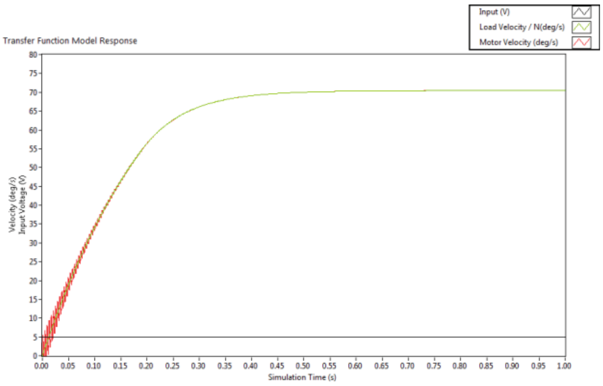


Figure 8 - 5V Step Response – Simulation

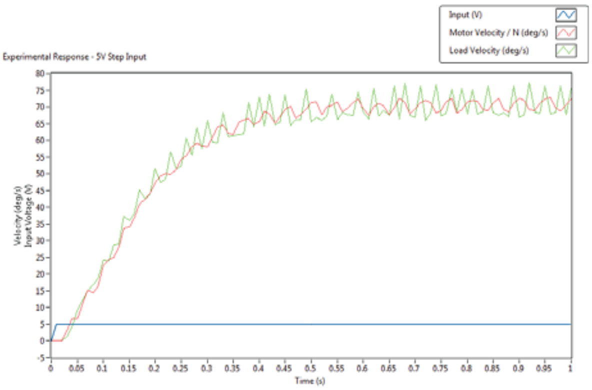
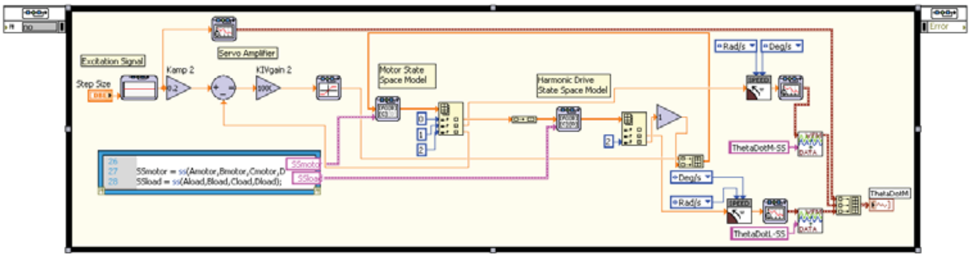


Figure 9 - 5V Step Response – Experiment



```
SSmotor = ss(Amotor,Bmotor,Cmotor,D
SSload = ss(Aload,Bload,Cload,Dload);
```

Figure 7 - State Space Representation of System Model using MathScript

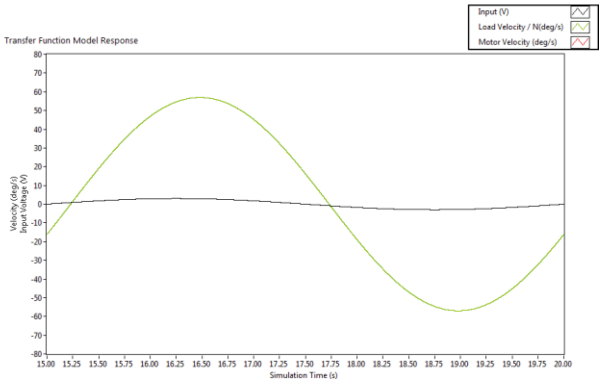


Figure 10 - 3V Low-Frequency Sine Response - Simulation

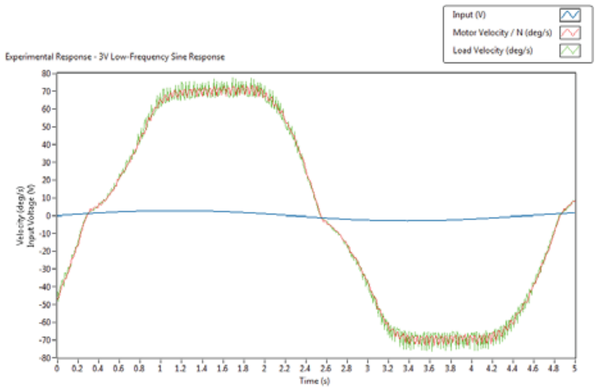


Figure 11 - 3V Low-Frequency Sine Response – Experiment

Comparing the results of the simulation and experiment, the general response of the simulation is similar, however, the experimental results exhibit measurement noise due to the quantisation error in the quadrature encoder. Also, in the experimental sine response, the maximum velocity is reached during the cycle and this limit is not reached in the simulation, as shown by the smooth sine response in Figure 10.

## IV. Further Work

Now that a model of the system has been developed and implemented using LabVIEW, further work will look at validating the model to ensure that it reflects the physical system being modelled. This will be conducted on the experimental test rig to provide a known system with which to test the system. The models can then be enhanced to include further dynamical effects and also to evaluate some of the other harmonic drive models discussed previously.

The initial model included only linear friction models, but it will be possible to extend the model to evaluate some of the nonlinear friction models that incorporate additional effects such as static and Coulomb friction, as surveyed in [11]. One example is the positional dependent friction as discussed in [8]. It was confirmed that this exists within the experimental test rig; this will be quantified and added to a future iteration of the model.

Within the larger scope of the full KTP project, the models developed will be used to test and evaluate control and estimation algorithms to evaluate the potential benefits in terms of increased performance of Chess Dynamics' products.

## V. Conclusions

In this paper, a model for a two axis servo platform has been introduced. Firstly, an outline of the system was presented, followed by models for the individual components of the system with the main area of interest being the harmonic drive gear system.

The overall model of the system was then presented and different representations of the model were then derived (simulation, transfer function and state space).

Results of running the simulation for various input signals were shown and the model was shown to lack one of the key nonlinearities of the physical system which has the most significant impact on its performance. The simulation results were then compared to the results from the physical test rig using the same input signals.

Finally, the direction of further work in the area was discussed as part of the overall KTP project.

## References

[1] Chess Dynamics Ltd. Cobra product [Image]. 2012.

[2] D. Vrancic, D. Juricic, T. Hofling. Measurements and mathematical modeling of a DC motor for the purpose of fault diagnosis, University of Ljubljana. Rep. DP-7091. 1994.

[3] H. Taghirad and P. Bélanger. An experimental study on modelling and identification of harmonic drive systems, in *Proc. 35th Conf. Decision and Control*, 1996, pp. 4725-30.

[4] J. P. Hauschild, G. Heppler and J. McPhee. Friction compensation of harmonic drive actuators, in *Proc. 6th Int. Conf. Dynamics and Control of Systems and Structures in Space*, Italy, 2004, pp. 683-692.

[5] R. Dhaoudi. Torque control in harmonic drives with nonlinear dynamic friction compensation. *Journal of Robotics and Mechatronics*, Vol. 16, No. 4, pp. 388-389, 2004.

[6] M. Hashimoto, Y. Kiyosawa. Experimental study on torque control using harmonic drive built-in torque sensors. *Journal of Robotic Systems*, Vol. 15, No. 8, pp. 435-445, 1998.

[7] F. Ghorbel and P. Gandhi. On the kinematic error in harmonic drive gears, in *Journal of Mechanical Design Trans. ASME*, Vol. 123, pp. 90-97, 2001.

[8] P. Gandhi, F. Ghorbel, J. Dabney. Modelling, identification and compensation of friction in harmonic drives, in *Proc. 41st IEEE Conf. Decision and Control,* 2002, pp. 160-166.

[9] R. Dhaouadi and F. Ghorbel. Modelling and analysis of nonlinear stiffness, hysteresis and friction in harmonic drive gears. *Int. Journal Modelling and Simulation,* Vol. 28, No. 3, pp. 329-336, 2008.

[10] T. Tjahjowidodo, F. Al-Bender and H. Van Brussel. Nonlinear modelling and identification of torsional behaviour in harmonic drives, in *Proc. ISMA,* 2006, pp. 2785-2796.

[11] B. Armstrong-Helouvry, P. Dupont and C. Canudas de Wit. A survey of models, analysis tools and compensation methods for the control of machines with friction. *Automatica,* Vol. 30, No. 7, pp. 1083-1138, 1994.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Deployment of full vehicle simulator for electrical control system validation

Georgios Tsampardoukas
Jaguar Land Rover
Warwick, UK, CV35 0RR

Alexandros Mouzakitis
Jaguar Land Rover
Warwick, UK, CV35 0RR

*Abstract*—**Development and testing of automotive embedded control systems traditionally depended on the availability of prototype vehicles. Automotive manufactures adopted model based approaches in order to produce quality products faster. Thus, the need of more integrated testing using virtual environment in an automated manner becomes a vital element of product development. The full vehicle simulator aims to provide a fully integrated environment for verification and validation of the embedded automotive software to avoid the dependency of the prototype vehicles. The execution of different automated test scenarios aims to increase the development of the product faster without compromising robustness and quality. This paper deals with both the development of full vehicle simulator and the concept of the automated modelled test cases.**

***Keywords-** Hardare-in-the-loop; full vehicle simulator; automated modeled test cases, automotive control systems.*

## I. Introduction

The embedded software complexity and the number of electronic control units (ECU) integrated in modern luxury vehicles are radically increased due to the increasing distributive functionality, safety requirements and legislation for lower emissions. Today's luxury vehicles include more than 60 interconnected ECUs using various network systems [1]. Four main domains such as body systems, chassis, powe rtrain and infotainment constitute a typical vehicle electrical architecture. Mots vehicle functionality is distributed among these domains as shown in [2],[3] and [4].

The power train and chassis domain contains ECUs responsible to control systems such as engine management, anti-lock brake system, hybrid systems, transmission and vehicle dynamics [5] and [6]. These are generally continuous control systems and are interconnected using the high speed CAN network.

The body system domain, however, is responsible to deal with systems like security, locking, wipers, mirrors, start authorization, etc. These control systems are mainly event driven with response time slower than the powertrain domain. The characteristic of the body domain is that the overall functionality (i.e. locking) is distributed to more than one ECU. These ECUs are located on one or more network buses [7] and [8].

The infotainment domain is one of the most popular systems nowadays due to its human machine interface nature between and the driver and the vehicle. This system usually consists of DVD player, amplifier, human machine interface console, TV modules and navigation. Other infotainment features require communication with the external world using external media such as Bluetooth and Wi-Fi. Typically, the response time of these systems is very slow due to the nature of consumer electronics functionality. The data bandwidth required to run an infotainment system is high compared to powertrain and body domains. The driver interaction with the vehicle makes the infotainment system one of the hottest topics in modern automotive industry.

Distributive functionality shared amongst four domains can impact customer's perception about vehicle quality. Software complexity and programme development cycle substantially reduced due to continuously customer's demand for new features. Competition amongst vehicle manufactures radically increased due to demand for robust and quality vehicle systems. Advanced and sophisticated techniques (i.e. hardware-in-the-loop) commonly employed to validate the embedded software in real time early at the product development [4]. The drive to reduce dependency in prototype vehicles is still an important initiative for most vehicle manufacturers. The usage of prototype vehicles is therefore aimed mainly for verification and validation activities close to mass production data. Automated virtual testing environment promotes more robust, systematic, time efficient and cost effective way for software testing. It has the potential to uncover possible software failure modes and to perform fault diagnostics automated tests prior to development of prototype vehicles [8].

This paper is organised in the following manner: The first section deals with introduction and brief literature review in the area of automotive control system development and test. The second section presents the development of the full vehicle simulator within Jaguar Land Rover (JLR). Two case studies are considered in the next section. The results and benefits of the automated testing are depicted in section four. Section five illustrates the discussion about the main benefits of the automated modelled test cases. The last section gives some concluding remarks of the work presented in this paper.

## II. Full vehicle simulator overview

The main scope of this section is to describe the structure of the full vehicle simulator (i.e. fully integrated hardware-in-the-

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

loop platform) suitable for automated functional and non-functional testing.

### *A. Simulator setup and schematics*

The full vehicle simulator consists of three 21 inches cabinets and two load tables. The load tables hold vehicle real loads and ECUs. Figure 1 show the full vehicle simulator which is currently used within the premises of JLR.

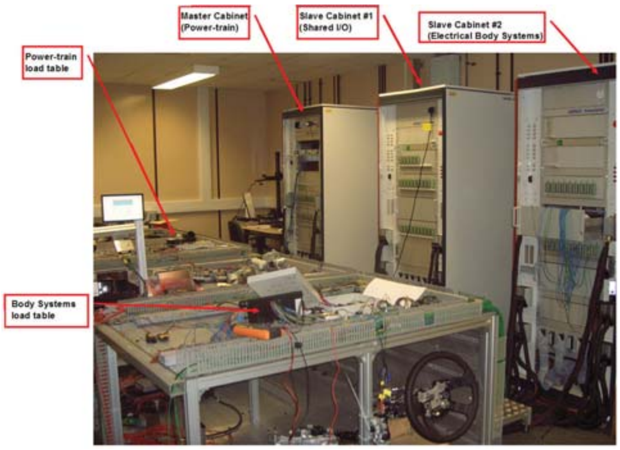

Master Cabinet (Power-train)
Slave Cabinet #1 (Shared I/O)
Slave Cabinet #2 (Electrical Body Systems)
Power-train load table
Body Systems load table

Figure 1. Full vehicle simulator at JLR.

In the following sections the high level requirements including the number of input and output (i.e. I/O) channels specified for the full vehicle simulator are given. The amount of the I/O required to interface with the ECUs to form the vehicle architecture clearly demonstrates the complexity of the system.

The main features of this simulation platform can be summarized as follows.

149 ADC; 145 DAC; 275 digital input; 79 PWM , 195 digital out, 126 digital relay output, 88 PWM output, 24 resistive channels, 22 special channels for powertrain simulation, 110 dedicated power lines and GND, 8 CAN channels, 32 LIN channels used to interface with ECU hardware. Quiescent current measurements an all power lines (i.e. 2 power switches), 1 power supply (i.e. 400A@20V), access to CAN and LIN channels for measurements, Serial ports from processors (i.e. 3 DS1006 quad processors), Integration of the low voltage tester (i.e. LVT), On/Off capability on all power lines, Fault insertion and load boards on the input channels w.r.t (with regards to the) simulator, fault insertion capability on all CAN and LIN channels, ABS valve detection unit, colour coding on the on each I/O type, Special software to interact with the simulator (i.e. Control Desk, Automation Desk, Motion Desk, CAN/LIN multi-message), Measuring point for the main power supply (i.e. 4 banana plugs on the front side of the first cabinet).

The first cabinet (i.e. master cabinet) dedicated for the powertrain, chassis and driveline domain (i.e. high speed CAN network domain). For instance, engine management system (i.e. EMS), chassis control and transmission control ECUs are integrated on the master table. Peripheral loads such as electronic throttle, injectors, differential and transmission solenoids interfaced to ECUs via the master load table to the simulator. The aim of load integration is to enable the ECUs to functionally operate in as close as possible to the environment of the vehicle. Thus, the number of logged diagnostic trouble codes substantially reduced. Engine and transmission plant models developed and executed in real time in order to provide dynamic closed loop control between the EMS module and the models of the engine/driveline.

The second cabinet (Figure 1) is known as first slave to the master cabinet. The use of this cabinet is to provide bulk I/O (i.e. input and output) channels to powertrain and body tables. These I/Os are distributed on both load tables to provide enough channels to ECUs for the interface with the simulator.

The third cabinet (i.e. body systems cabinet) dedicated to electronic body systems (i.e. body control module, door modules, keyless vehicle module, etc) and is known as the second slave of the master cabinet. This domain is dedicated to medium speed CAN modules and their LIN slaves (i.e. intrusion monitoring system). The second load table (i.e. body systems table as shown in Figure 1) is used to accommodate all the medium speed ECUs and their peripheral components..

### *B. Integrated model to control the full vehicle simulator*

The VITAL framework is a generic model that built to interface with twelve different core processors (three quad processors DS10006) of the full vehicle simulator. The aim of this framework is to create a structured environment for integration of potentially of vehicle electrical components. Additional features of this model are summarised below:

- The model structure allows integration of three dSPACE Quad core processors or more.
- Enables simulation of Multi-CAN/LIN architectures and other network protocols (i.e. Ethernet)
- Exchange of signals between cores of each processor.
- Selection mechanism for switching between real and modelled ECUs
- Common interface for simulation of ECU loads and actuators.
- Hardware interface between ECUs and simulator arranged per dSPACE IO board..

The above features offer less development time, more optimal model structure resulting in more efficient real time execution. In addition to that the model promotes consistency and reduces the risk for development errors.

Figure 2 show only an example of the VITAL model and the closed loop integration method between the real driver door ECU and the latch plant model. The plant model in this example provides the feedback signals to the driver's door control unit (i.e. DDCU). The door ECU outputs are fed back to the door latch model and the feedback of the door latch is fed back to the door ECU via the simulator IO channels.. The door latch plant model developed in Stateflow® is a true functional representation of the real latch [9].

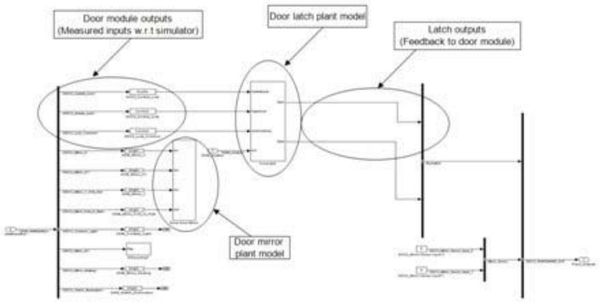


Figure 2. Hardware interface on VITAL framework for closed-loop control

More detailed description about the VITAL framework is presented in [9]

## III. CASE STUDY (CLOSED LOOP CONTROL OF BODY SYSTEM FUNCTIONALITY)

The aim of this section is to demonstrate two automated test scenarios using the fully integrated platform described previously. The first scenario is the drive away door locking and the second one is the valet mode.

The drive away door locking function is a security feature that locks all the doors automatically when the vehicle speed exceeds the threshold speed (i.e. 32Km/h). This feature is selectable by the driver, and any operation of the door locks by any other means (i.e. master locking switch on the facial panel) will unlock the doors [10].

Valet mode is also a security feature that allows the vehicle to be driven with the luggage compartment locked with restricted touch screen functionality. This feature is accessible directly from the home menu on the touch screen or from the vehicle settings screen. The vehicle owner enters a four digit Personal Identification Number (PIN) to a soft key pad displayed on the touch screen. This PIN must be entered twice in order for the valet mode to be enabled. A pop-up screen is displayed, confirming that the vehicle is now in valet mode. To cancel the valet mode operation, the PIN number must be entered once again [10].

### *A. System Overview*

The deployment of the drive away door locking and the valet mode features is depicted in Figure 3. Several ECUs are required to exchange data amongst different domains in order to interpreter the customer's operation into low level software command (i.e. set the threshold of drive away door locking).The simulation environment to deliver these features required the integration of the following ECUs and models, engine controller; gateway controller; door controllers; infotainment control units; engine and driveline real-time models.

The complexity of system integration significantly increases due to the following reasons, number of ECUs; the scale of the system integration; inter-dependencies across functional areas and domains. For instance, the DDCU receives the command from the gateway to lock the door latch when the EMS transmit the correct vehicle speed from the high speed CAN network. An engine model and accurate sensor simulation required for the real EMS hardware to assume that an engine is in operation and the vehicle is in a drive cycle. Although drive away door locking feature appears to the customer to be a simple operation the effort required to develop the automated virtual environment is certainly a challenging engineering task.

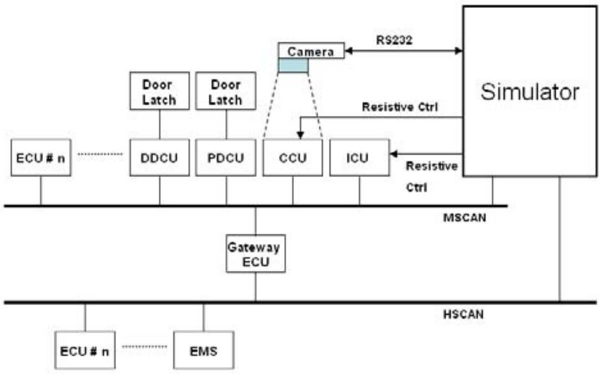


Figure 3. Schamatic of vehicle feature demployment

### *B. System Integration*

The full vehicle simulator is interfacing (Figure 3) the ICU using resistive signals for simulation of the driver switch-pack component which is hardwired back to the ICU (this switch-pack controls the navigation menu shown in the CCU'display). The simulator is also providing resistive signals to the ICU in order to emulate driver touch screen selection. High resolution camera is integrated with the simulator via serial (i.e. RS232) connection to provide feedback acquired by the image processing software. This camera is used as feedback sensor to capture the results from the instrument cluster and infotainment display. Detailed description about multi-camera vision system is out of the scope of this paper.

The two front door modules (i.e. driver and passenger), instrument cluster and infotainment display have an interface to the medium CAN bus. The two door latches are integrated to each door module using hardwire connection. The gateway ECU (i.e. ECU that accommodates the core body system functionality) is used to pass network signals from medium speed CAN bus to high speed. The EMS is interfaced with the engine model to high speed CAN in order to provide engine and vehicle speed to the rest of the vehicle systems. The rear door modules are not assessed on this paper since their behaviour is very similar to front doors.

### *C. Event driven control Logic*

Event driven control logic is a discrete programming method that is based on the conditional transition between operating modes. This method is used in this paper to model the automated control sequence of the infotainment displays selection.

### 1) *Automatic navigation on instrument cluster*

This control logic is employed for automatic selection of the drive away door locking threshold. The model is divided in two main parts.

The first part deals with the simulation of a button (i.e. up, down, left, right and OK button) from the driver's switch pack. Figure 4 shows a snapshot of this model which represents one event of the OK press button. The event sequence is required to alter the simulator's resistive output from idle (X Resistance) to pressed position (Y Resistance). The transition between the two states delayed for 500 milliseconds. This allows the instrument cluster to process the request received from the infotainment control unit. A counter is implemented to capture the complete sequence. It is also used as a transitional condition to another event (i.e. simulation of down button). In addition to counter, the camera feedback is used as an alternative transitional condition. The camera is trained to identify the main menu pattern. This menu is displayed after pressing the OK button. Figure 4 shows that either the camera feedback or the counter conditions must be satisfied in order to continue the sequence.

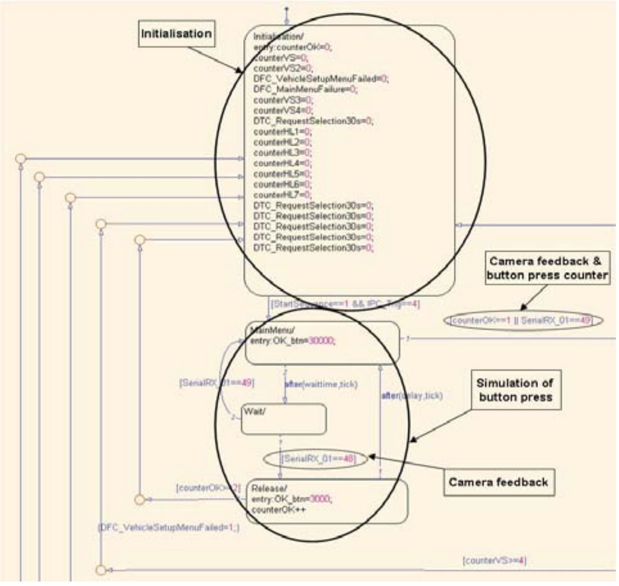


Figure 4. Automatic navigation through the instrument cluster menu

The second part (not shown in this paper) of this control logic is the evaluation of the menu position via the camera's feedback. The camera is placed to point directly the instrument cluster. It is trained to recognise five different positions of the drive away door locking settings menu. The control logic identifies the menu selection and responds to the driver's choice. For instance, different sequence is employed in order to set 32Km/h when the default position is on OFF mode and different sequence when the value is 5Km/h. The difference is due to the number of down or up button event presses. Similar concept is used to control the automatic selection of valet mode.

### 2) *Valet Mode*

The driver inserts twice a predefined four digit code to ICU touch screen in order to set the vehicle to valet mode. The automatic sequence requires selection steps on the ICU menu before the driver enters the PIN number. An event driven control logic is developed using Stateflow® to achieve this automated selection

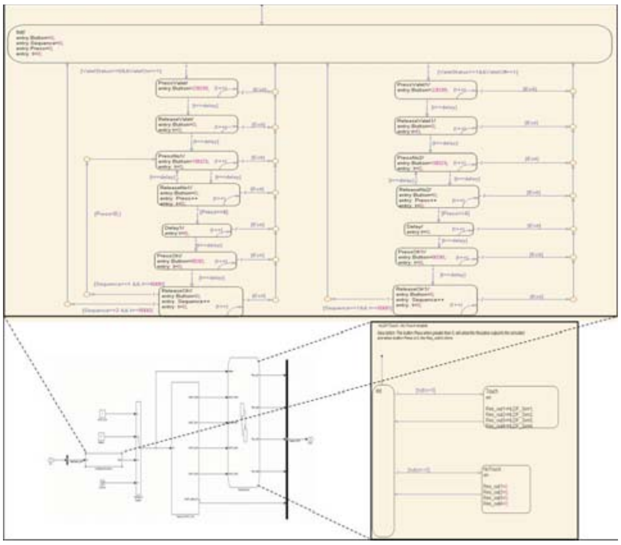


Figure 5. Automatic control of infotaiment display (i.e. valet mode selection)

Figure 5 shows only the sequence to set the valet mode on. The first state defines the initial conditions and determines the idle mode. On this mode the control logic provides the idle resistive output to ICU via the simulator's restive channel. The sequence starts when all the initial transitional conditions are satisfied. Delay of 800ms is implemented between the states after the idle mode. This allows the ICU to process the request from the driver and the infotainment graphics. Exit conditions are implemented on every stage of the sequence to ensure smooth execution avoiding stagnation points. On this particular example a medium speed CAN signal is used to inform about the valet mode status instead of camera feedback.

## *D. Control desk interface*

Control desk layouts are developed in order to control the full vehicle simulator (i.e. 20 controlDesk® tabs) via graphical user interface. Figure 6 shows only a sub-set of the main working layout. This is dedicated to control the automated selection of the drive away door locking. It is also shown that the camera mode and image processing job is controlled from this layout. The manual mode of the driver's switch pack is also part of this interface as shown in Figure 6.

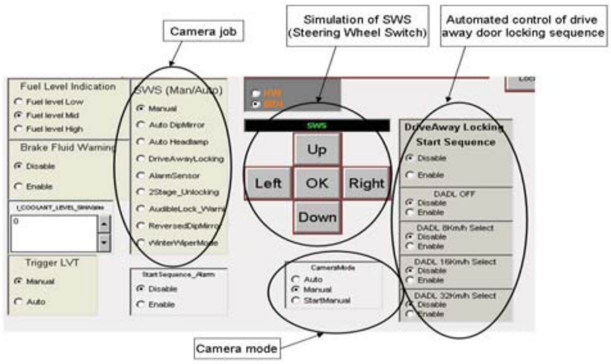


Figure 6. Graphical user interface for automated drive away door locking selection mode.

A Similar graphical user interface is developed to control the automated valet mode selection. However, this is not shown in Figure 6.

### E. *Manual test scenario description*

A manual test scenario is performed to set the drive away door locking threshold from OFF mode to 32Km/h. The flowchart in Figure 7 depicts the manual sequence.

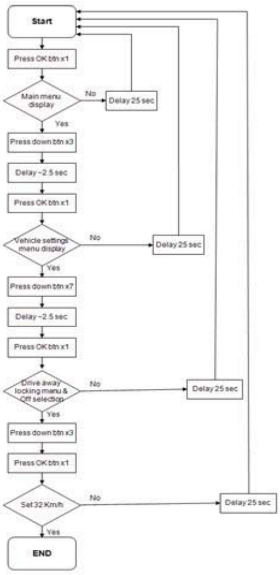


Figure 7. Flowchart of manual drive away door locking from OFF to 32Km/h

The sequence starts from a warning free instrument cluster. Figure 7 shows the exact sequence required for the instrument cluster to display the driver away door locking settings menu. Since this is a submenu, it is required to navigate three layers below the main menu. An extra delay is introduced on every conditional state (approximately 2.5 seconds) to allow the camera to process the captured image. Based on the camera feedback, the control logic decides about the status and the progress of the automated sequence. Similar manual test scenario is produced for the valet mode. The only difference is that the navigation is through the resistive touch screen of the infotainment display. The flowchart for the valet mode selection is not shown in this paper.

## IV. Test Automation

The purpose of the automated testing is to execute the existing manual test cases in a repeatable manner. MXvDEV® test automation software is used to model the automated test cases.

### A. *Execution of the automated test sequence*

Both automated sequences are executed and the results are captured in a graphical manner. Figure 8 and Figure 9 present the test results for drive away door locking and valet mode, respectively.

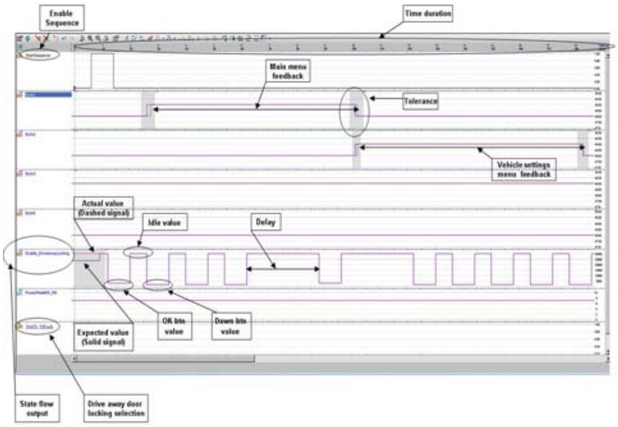

Enable Sequence
Time duration
Main menu feedback
Tolerance
Vehicle settings menu feedback
Actual value (Dashed signal)
Idle value
Delay
OK btn value
Down btn value
Expected value (Solid signal)
State flow output
Drive away door locking selection

Figure 8. Automated test results for drive away door locking to 32Km/h

For presentation purposes, only half duration of both scenarios are presented in Figures 8 and 9.

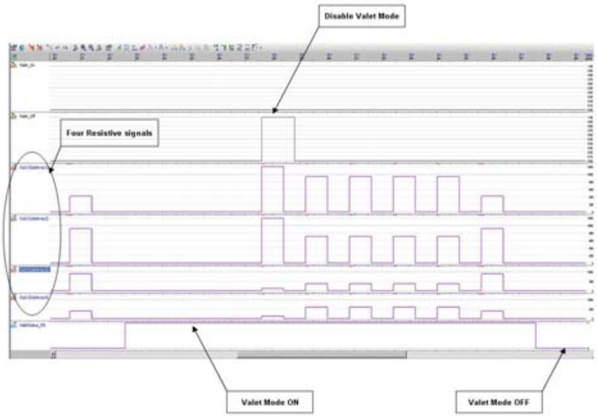


Figure 9. Automated test results to disable the valet mode.

The desired signals are graphically modelled as shown in Figures 8 and 9. It is difficult to distinguish the differences between the actual and desired signal. This is due to exact match of the two signals. For that purpose, the expected signal (i.e. sixth signal in Figure 9) is altered at the start of the sequence. A time delay is implemented on x-axis (time axis) at the first 500 milliseconds of the test in order to deliberately make the test cases result a fail (i.e. a shadowed area).

## V. Discussion

The evaluation of the test results has shown that both case studies (i.e. Figures 8-9) were executed successfully in an automated manner. These scenarios performed with zero tolerance on the y-axis. The expected signal has identically matched the actual measured. Occasional deviations observed on x-axis (i.e. time), where slight time variations (i.e. milliseconds) occurred. This is due to the time delays of the real time processor to fetch and process the results. Thus, signal delay of few milliseconds was introduced in the area of interests (i.e. shadowed areas on x-axis) to avoid failures of the test case.

The benefits of the modelled test cases are summarised below:

- Accurate definition of the test case in terms of time and signal definition.
- Graphical representation of the signal during the test execution.
- Reduced test case duplication due to test case re-use.
- Less specialised knowledge is required to analyse the test results.
- The graphical definition of the test case can uncover failure modes associated with the functional requirements prior to the test execution.
- Test cases can be linked to system requirements.
- Visualisation of signal tolerances (i.e. y-axis).
- The review of the automated test case and scenarios is significantly minimised.

The creation of the modelled test cases require well defined design verification plan and signal specification standards to be in place. This approach of automation helps JLR to move test creation using tabular format to one with graphical representation.

In addition to the above remarks regarding the creation of automated test cases, the deployment of the full vehicle simulator has demonstrated the following potentials.

- Most Electrical functional requirements can be validated prior to prototype build.
- Distributed functionality validation is decoupled from single software release.
- Drive cycles which have functional safety implications, can be executed in a controlled test environment.
- The product design can be evaluated and altered early in the programme before commitment to tier 1 is made.
- Early feature demonstration can help towards concept selection and decision making.
- Enables cross functional team working and explores opportunities towards "what can we do better and how?"

Although the aforementioned characteristics clearly deliver competitive advantage to an OEM, there are points to be considered before full deployment takes place.

- Significant upfront capital investment is required to purchase the hardware and software simulation components.
- Early engineering effort is required to develop product engineering specification.
- Engineering mind set shift from manual vehicle testing to automated simulation based testing.

## VI. Conclusion

The purpose of this paper was to present a fully integrated hardware-in-the-loop environment for validation of distributive vehicle functionality. The automated modelled test cases concept is introduced with the execution of two case studies. The successful execution of both scenarios has proven that the entire vehicle functionality can be modelled and executed on the full vehicle simulator.

## References

[1] Waltermann, J., 2009. Hardware-in-the-loop: The Technology for testing Electronic Controls in Automotive Engineering. *6th Paderborn Workshop: Designing Mechatronic Systems, Paderborn, April 2-3,Germany*

[2] Kendall,I,R. and Jones,R,P., 1999. An Investigation into use of hardware-in-the-loop simulation testing for automotive electronic control systems. Control Engineering Practice, 7 (1999), p.p 1343-1356.

[3] Lamberg, K.,Richert, J. and Rasche, R., 2003. A new environment for integrated Development and Management of ECU tests. Proceedings of the SAE World Congress, Detroit,USA.

[4] Mouzakitis,A.,Humphrey,R.,Bennett,P. And Burnham,J,K. 2006. Development, Testing and Validation of Complex Automotive systems. The 10th Mechatronic Forum Biennial International Conference,MX2006,Philadelphia,USA.

[5] Dhaliwal, A., Shreyas,C., Nagaraj,C, and Syed,A. 2009. Harware-in-the-loop Simulation For Hybrid Electric Vehicles-An Overview, Lessons Learnt and Solutions Implemented. SAE Technical Pepers 09AE-0198.

[6] Wu,K., Zhang,Q. And Hansen,A. 2004. Modeling and identification of a hydrostatic transmission hardware-in-the-loop simulator. International Journal of Vehicle Design, Vol.34,No. 1.

[7] Henselmann,H. 1993. Hardware-in-the-loop simulation as a standard approach for development, customisation and production test. SAE technical papers, 930207.

[8] Tsampardoukas,G., Mouzakitis,A. And Sydor,P., 2009. Design Methodology for Integrating Networked Automotive Electronic Control Units Using Hardware-in-the-loop. Proceedings of the 20th International Conference on Systems Engineering ,Coventry, UK.

[9] Huang, Y., McMurran, R., Dhadyalla, G., Jones' P. and Mouzakitis., 2009. Model-based testing of a vehicle instrument cluster for design validation using machine vision. Journal of Measurement Science and Technology, Vol. 20, No. 6.

[10] Jaguar Land Rover, User's Handbook, 2009.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Modelling of Hybrid Plus Retrofit Hybrid System

Xiao GUO,
University of Glamorgan
CF37 1DL, UK
Email xguo@glam.ac.uk

Jonathon Williams
University of Glamorgan
CF37 1DL, UK
Email jgwilliams@glam.ac.uk

Prof.Guoping LIU
University of Glamorgan
CF37 1DL, UK
Email gpliu@glam.ac.uk

***Abstract*—This paper focuses on addressing theoretical and practical considerations in developing a retrofit system for traditional internal combustion engine vehicle. Extensive simulation test results and theoretical analyses are presented and analyzed. Firstly, the basic theory of design is discussed in term of the market needs, target vehicle chosen, system components chosen, etc. A controller algorithm design is discussed in details. The performance test results are simulated later in several drive cycles. The results show that the retrofit kit can improve about 3% fuel consumption in a NEDC and 19% in a real world urban delivery drive cycle.**

***Keywords: mild hybrid, HEV, CVT, matlab/Simulink***

## I. INTRODUCTION

Nowadays, energy consumption and CO2 emissions of road transport are the main issues amongst all impacts. CO2 emissions threaten the world with climate-change. The increasing demand and the tight supply for oil is the problem every country need to face in the next decade, many research believe that the demand for oil will exceed the production capability (Figure 1) [1] Oil reserves will run short within the near future, experts are just arguing about the time. How to maintain the energy supply to catch up with the increase demand become an on-going concern and a high priority.

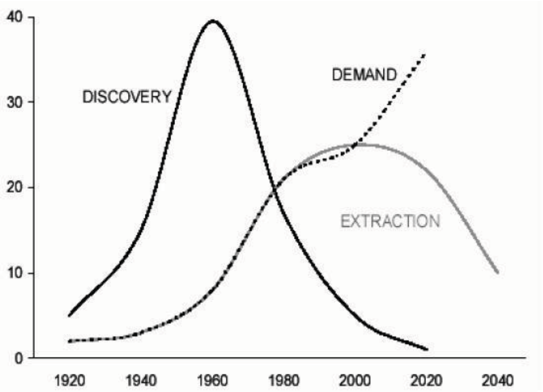


Figure.1 Actual and projected worldwide discovery, extraction and demand for conventional oil (From 1920 to 2040 in billions of barrels). [1]

The transport energy consumption takes a significant part of the oil consumption worldwide and it is continue to rise rapidly. In 2000 it was 25% higher than in 1990 and it is projected to grow by 90% between 2000 and 2030 as shown in Figure 2.[2]

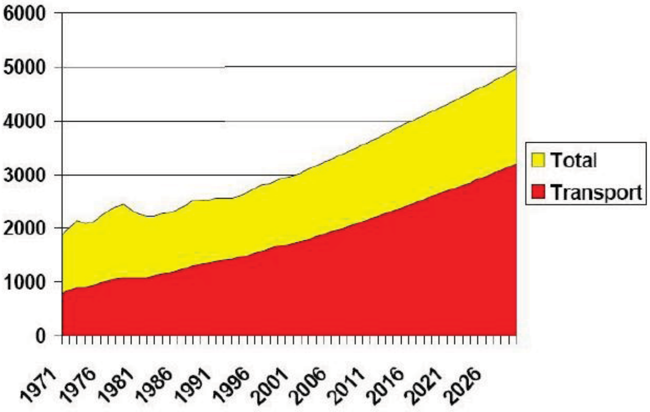


Figure. 2 Globe Oil Consumption Perspective. [2]

Some researchers indicate that Sweden and the UK are large potential markets of the HEV due to the potential energy savings and also California by the exceptionally high incentives. [3] Also the use of HEVs depends on the legislation which either increases the fuel cost or mandates high fuel economy. As a result, France and the UK are good potential markets because of the high fuel price.[4] The Daily Telegraph reported in February 2005 that “The value of home delivery has risen to £36.8bn in 2004, the story continued, with one-seventh of all spending in the retail sector now on items delivered to door”. More and more big retail companies offering the online shopping and home delivery services. This door to door shopping saved a lot of time and money for customers. The accelerating growth of the new retail pattern means that a new pattern of usage needs to be checked. This new pattern of usage of delivery vehicle, door-door and start-stop all the time, is potentially highly wasteful of fuel. Vans are most commonly used in urban areas for delivery and daily commercial use. In this way, the research start from the best potential type of vehicle, diesel Vans in UK.

## II. HYBRID PULS SYSTEM

The main objective of this research is modeling and validation a solution to retrofit the normal Diesel Van to mild-hybrid vehicle. Transfer the van into a vehicle which can do regenerate break and assist on start.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

The system is mainly designed for big delivery fleet companies, so it should match the need of the market. It is quite different between a private car customer and a commercial fleet manager. The former would like to pay for new technology and comfort features that they can afford, while the later one often resistive to new technology and more interested in a reliable, robust vehicle with the lowest cost during the ownership. In this way, the hybrid plus system should be with an affordable price, provides acceptable vehicle performance, supports versatility of the vehicle, has low impacts on the environment, consumes drastically less fuel, provides some additional value to the customer without a big cost. The lowest cost and best robust are considered mainly throughout the design. The system contains four main parts, A Supercapacitor, A CVT, A PMDC Motor and controller units. The system works as follows in Figure 3.

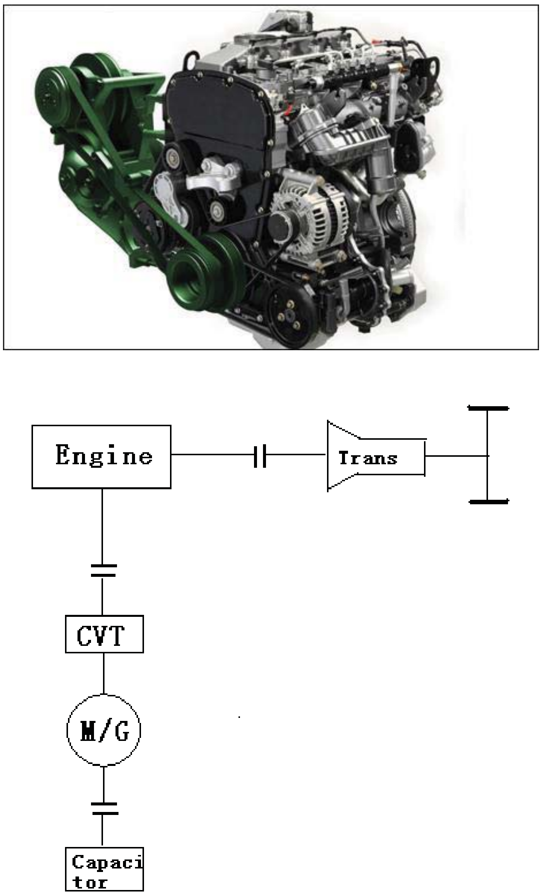


Figure.3 Hybrid plus system components

The retrofit kit will be installed in a VAN with a limit space allowed, so the size of the motor is limited. With a CVT installed between the motor and engine, the torque applied on the engine crankshaft will be increased, which means more assist and regenerate power will be provided. Normally the acceleration or brake process all finish within 3 seconds, to achieve the best assist or regenerate performance, a supercapacitor is used in this design mainly because of its quick charge/discharge performance is better than a lithium ion battery in this application

When the Engine is speed up from a very low speed, the CVT turns to a ratio which allows the motor speed same as engine speed at first. At this moment, the CVT out put the max torque and help engine accelerates. And then the engine speed is accelerating, the CVT need continuously change ratio allow the motor work at the slowest allowed speed which give a max torque on engine at every motor's speed. When the engine speed is equals to the max speed of the motor with the CVT max ratio. The clutch will off and the whole acceleration finished. This is how this system works on acceleration assist.

On the regenerate brake, it works just on the same theory. When the engine brake from a high speed to a low speed, the CVT turns to a ratio which allows the motor rotate at its max speed which will force it works as a generator. The motor generator will charge the Supercapacitor. With the engine slow down, the CVT should continuously change ratio allows the motor generator works at its max speed at every engine speed. When engine speed is decelerate to the min speed which can drive the motor as a generator, the clutch will off and the regenerate brake finish.

## III. VEHICLE POWERTRAIN MODELING

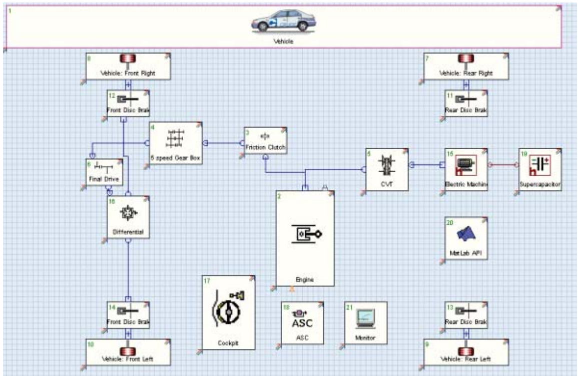


Figure.4 Hybrid plus system simulation layout

The system is basic on a typical Manual front wheel drive power-train, the system combines a CVT, a PMDC motor and a super-capacitor. With these units the system is retrofitted into a mild-hybrid vehicle. The simulation used Simulink/MATLAB run with other simulation software together build a co-simulation system.

## IV. MODELING AND SIMULATION RESULT OF THE VEHICLE

The basic theory of all hybrid vehicles is to regenerate the “free energy” from vehicle braking and use this energy to help a vehicle accelerate when it need more power. In this system the vehicle has four different states: assist mode, regenerate mode, curise mode and idle mode. The controller read CAN signal identify the vehicle mode and give the right command to motor and CVT.

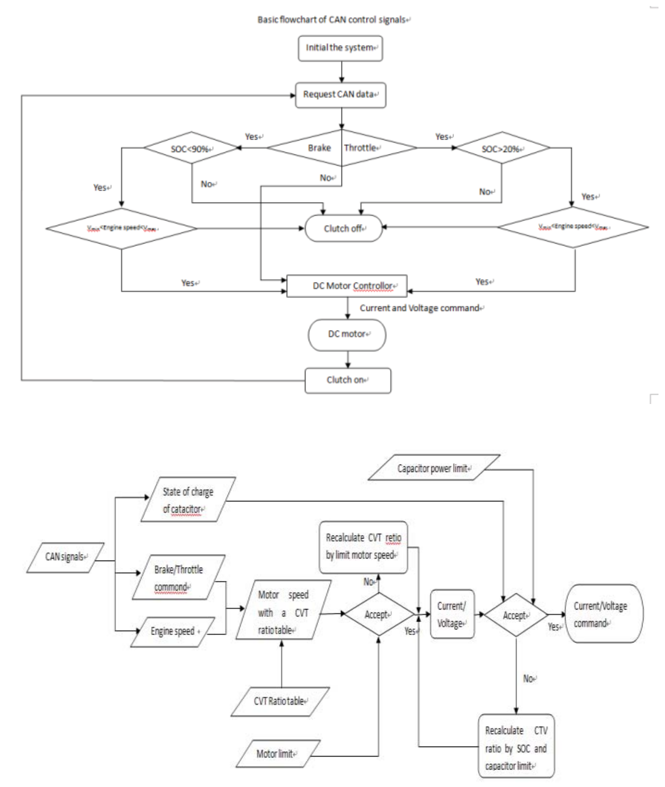


Figure.5 Hybrid plus system controller calculation piority chart

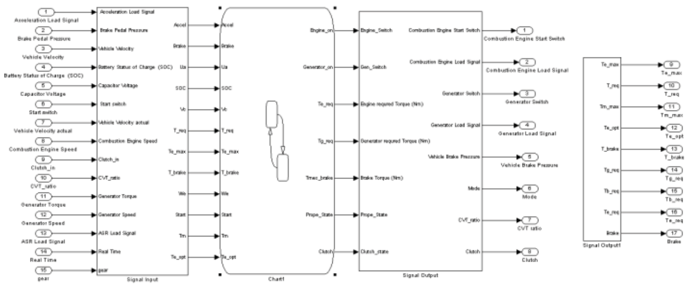


Figure.6 Hybrid plus system controller block in Matlab/SImulink

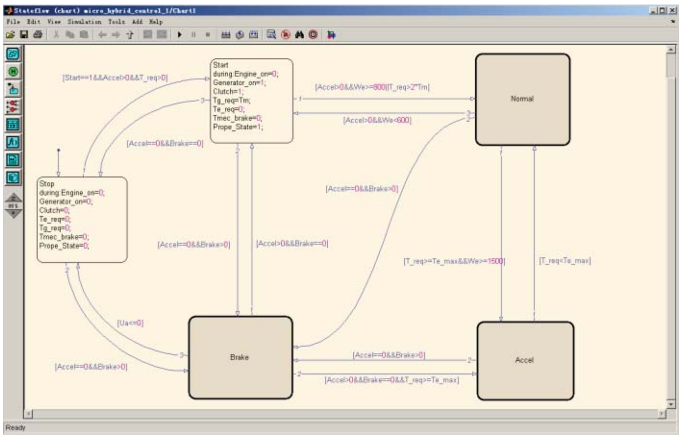
The simulation system is based on Simulink/MATLAB, the main inputs and outputs are shows in the following tables. There are three main calculation block in the model. The main part of the model is a stateflow chart. The main function of this chart is used the signals to identify the system states and gives out different commands to CVT and motor.

| | Name of inputs | unit |
|---|---|---|
| 1 | Acceleration Load Signal | % |
| 2 | Brake Pedal Pressure | bar |
| 3 | Vehicle Velocity Require | km/h |
| 4 | Battery Status of Charge (SOC) | % |
| 5 | Capacitor Voltage | V |
| 6 | Start switch | - |
| 7 | Vehicle Velocity actual | km/h |
| 8 | Combustion Engine Speed | 1/min |
| 9 | Combustion Engine Torque | Nm |
| 10 | CVT_ratio | - |
| 11 | Generator Torque | Nm |
| 12 | Generator Speed | rad/s |
| 13 | ASR Load Signal | % |
| 14 | Real Time | s |
| 15 | gear | - |

| | Name of outputs | unit |
|---|---|---|
| 1 | Combustion Engine Start Switch | - |
| 2 | Combustion Engine Load Signal | % |
| 3 | Generator Switch | - |
| 4 | Generator Load Torque | Nm |
| 5 | Vehicle Brake Pressure | Bar |
| 6 | Mode | |
| 7 | CVT Ratio | |
| 8 | CLutch | - |

Table.1 Veriables used in the controller

The stateflow chart shows as follows. There are 5 main states: start, stop, normal drive, acceleration and brake. The states are changed when the inputs changed.



Stateflow (chart) micro_hybrid_control_1/Chart1
File Edit View Simulation Tools Add Help
[Start==1&&Accel>0&&T_req>0]
[Accel>0&&We>=800||T_req>2*Tm]
[Accel>0&&We<600]
Normal
[Accel==0&&Brake==0]
[Accel==0&&Brake>0]
[Accel==0&&Brake>0]
[Accel>0&&Brake==0]
[T_req>=Te_max&&We>=1500]
[T_req<Te_max]
[Ua<=0]
Brake
[Accel==0&&Brake>0]
[Accel==0&&Brake>0]
[Accel>0&&Brake==0&&T_req>=Te_max]
Accel
Ready

```
Start
during:Engine_on=0;
Generator_on=1;
Clutch=1;
Tg_req=Tm;
Te_req=0;
Tmec_brake=0;
Prope_State=1;
```

```
Stop
during:Engine_on=0;
Generator_on=0;
Clutch=0;
Te_req=0;
Tg_req=0;
Tmec_brake=0;
Prope_State=0;
```

Figure.7 Figure.3 Hybrid plus system controller all modes in Stateflow

In the stop mode, the Acceleration signal is 0, Brake signal is 0, torque required is 0 and vehicle speed is 0, an engine off signal will be given.

In the start mode, the engine is on, clutch signal is on and car is about to move.

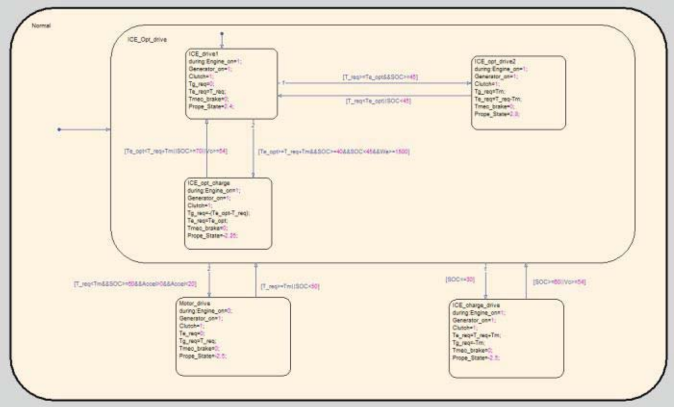


Figure.8 Hybrid plus system normal drive mode in  Stateflow

In the normal drive mode there are 3 different drive sub-modes. The motor drives only state, combine drive state and engine drive only state. When the vehicle get a start signal, the motor will spin up first drives the engine crankshaft and helps the engine to start. This state mode will be less than 1 second and then both of the engine and motor will working together or engine only. This depends on the SOC (state of charge) of the supercapacitor. When the SOC is lower than the minimum limit, the motor will work only as a generator to charge the battery.

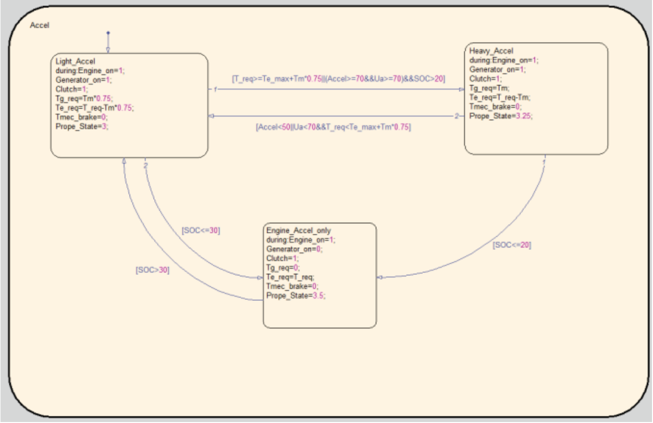


Figure.9 Hybrid plus system acceleration mode in  Stateflow

In the acceleration mode, there are 3 states. Engine accelerates only, light accelerate demand and heavy accelerate demand. When the SOC is lower than the minimum limit, the motor will stop working. In the heavy acceleration state, the motor will work in full load, in light acceleration mode the motor will work with a calculated load signal from the controller.

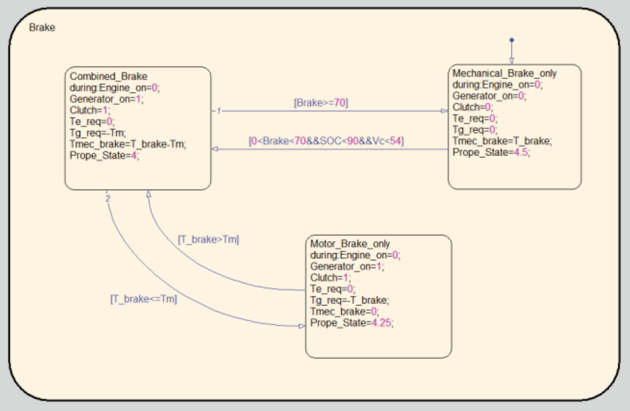
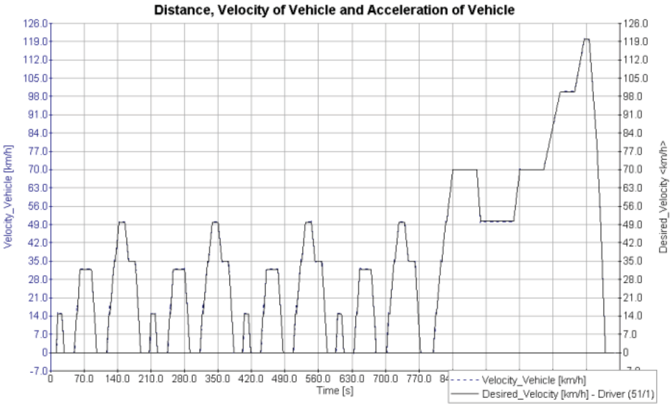


Figure.10 Hybrid plus system brake mode in Stateflow

In the brake mode, there are 3 modes. Motor brake only, combined brake and mechanical brake only. This is depends on the brake signal from the CAN bus.

The vehicle been tested with several drive cycles. One simulation running under NEDC-manual (New European Driving Cycle), the simulation result shows as follows, the vehicle follows the drive cycle quite good. It follows other ones quite good as well.



## V. SUMMARY AND CONCLUSIONS

The focus on this study is the feasibility analysis and evaluation of the retrofit set, and the improvement of the fuel economy of the Van after the kit installed. This study utilized a combination of simulation and some on-road test. This paper presents the results of modeling and simulation of the mild hybrid vehicle after the kit installed.

In the NECD drive cycle the test results show that the hybrid plug in system will increase about 3% fuel consumption.

But in the door-door start stop Urban Delivery Drive Circle (UDDC) the conventional ICE vehicle test results are

| | | |
|---|---|---|
| *Overall Fuel Consumption:* | *0.6229* | *[kg]* |
| *Idle Fuel Consumption:* | *0.0310* | *[kg]* |
| *Acceleration Fuel Consumption:* | *0.2858* | *[kg]* |
| *Constant Drive Fuel Consumption:* | *0.2457* | *[kg]* |
| *Deceleration Fuel Consumption:* | *0.0603* | *[kg]* |
| *Fuel Consumption:* | *7.56* | *[l/100km]* |
| *CO2 Emission:* | *178.23* | *[g/km]* |

The Hybrid Plus system one is

| | | |
|---|---|---|
| *Overall Fuel Consumption:* | *0.5016* | *[kg]* |
| *Idle Fuel Consumption:* | *0.0009* | *[kg]* |
| *Acceleration Fuel Consumption:* | *0.3093* | *[kg]* |
| *Constant Drive Fuel Consumption:* | *0.1846* | *[kg]* |
| *Deceleration Fuel Consumption:* | *0.0068* | *[kg]* |
| *Fuel Consumption:* | *6.09* | *[l/100km]* |
| *CO2 Emission:* | *143.59* | *[g/km]* |

The fuel consumption has reduced 19.4% and CO2 emission has reduced 19.4% as well. Based on the result, this system has a great potential on energy saving and reduce CO2 emission in urban delivery vans. This retrofit gear could also be applied on the HGVs, with a proper size of motor and capacitor. The system still could be improved with the engine start/stop function while idling.

## References

[1]. Gilbert R , Background paper for a post Kyoto transport strategy (06/07/02).

[2]. Lew, F., Reducing Oil Consumption in Transport: Combinning Three Approaches. 2004, Office of Energy Efficiency, Technology and R&D Internation Energy Agency.

[3] Ronning, J.J. and G.L. Grant, Global hybrid electric vehicle markets and missions, in SAE Future Transportation Technology Conference and Exposition. 1999: Costa Mesa, California, USA.

[4] West, J.G.W., Propulsion systems for hybrid electric vehicles, in IEEColloquium on Electrical Machine Design for All-Electric and Hybrid-Electric Vehicles. 1999. p. 1-9.

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# Model-base Predictive Control for Vibration Suppression of a Flexible Manipulator

Mehdi Abdolvand
Sama technical and Vocational Training College
Islamic Azad University, Islamshahr Branch
Islamshahr, Iran
Mehdi.abdolvand@gmail.com

Mohamad Hosain Fatehi
Department of Electrical Engineering
Islamic Azad University, Science and Research Branch
Tehran, Iran
Mh_fatehi@srbiau.ac.ir

***Abstract*—This paper focuses on the development of a multivariable predictive controller for vibration suppression of a flexible manipulator using piezoelectric actuator. To the best of the authors knowledge, the predictive controller for active vibration suppression has been rarely studied, so making it a prime area research to explore. A two-step procedure performed to develop the detailed model of the whole structure. First a dynamic model of the structure without piezoelectric film actuator is developed using the combined Lagrange-Assume modes method. Second the influence of the actuator is incorporated by calculating generalized applied force on the substructure.**

***Keywords-Active Vibration Control; Flexible Structures; Model Predictive Control***

## I. INTRODUCTION

Design of flexible smart structures for vibration suppression and noise control represents a major challenge in the past two decades. Such structures are used in a variety of applications including robots, industrial machine design and tooling, aircraft systems, civil engineering structures and the space stations.

The flexible structures are generally lighter in weight, designed to use less material in order to be more transportable. They need less power for motion and therefore can be driven by smaller actuators resulting in less cost. Flexible structures are characterized by a significant number of closely spaced, lightly damped low frequency modes.

We focus on the algorithms which were recently developed and applied on motion control and active vibration damping of flexible structures. Lee and Moon [1] used two separate feedback loops for position and damping. They concluded that the controller is simple but robust, since it cancels out nonlinear and uncertain dynamics by acceleration feedback, and adds more damping by base motion feedback. Knotnic [2] used the linear quadratic regulator (LQR) and the acceleration feedback control method for vibration suppression. Chevallereau and Aoustin [3] applied nonlinear control laws for vibration control. Banks [4] used a linear quadratic Gaussian (LQG) compensator control with proper orthogonal decomposition to control a cantilever beam with a piezoelectric patch. Chen [5] applied the LQR algorithm and studied the optimum layout of the piezoelectric materials based on control performance specifications and a cost function. Robust control approaches such as $H_\infty$ control design, robust pole assignment and D-stability constraints have often been applied to the problem of controlling large flexible space structures [6].

The strategies mentioned above all provide reasonable suppression of structural vibration. However, the feedback loop controller is difficult to apply to MIMO systems. Also, LQR and LQG controllers need a very accurate plant model for good control while the $H_\infty$ controller can't easily handle multivariable constraints in the manipulated variables.

## II. MODELING

### *A. Modeling of bending vibration*

In this section the vibration of a flexible beam in the direction perpendicular to its length is considered. Such vibration is often called transverse vibration or flexural vibrations. Fig.1 illustrates a cantilevered beam with the transverse vibration [7].

From mechanics of materials, the beam sustains a bending moment M(x,t), which is related to the beam deflection by:

$$M(x,t) = EI_{cs} w_{xx}(x,t) \tag{1}$$

A model of bending vibration may be derived from examining the force diagram of an infinitesimal element of the beam as indicated in Fig. 1.

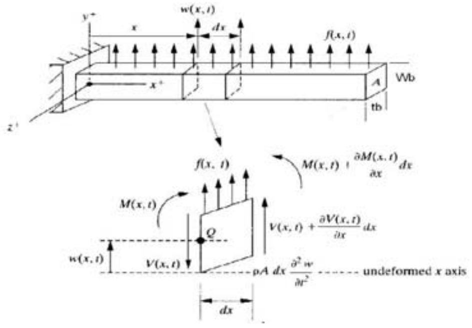

y⁺
w(x,t)
x
dx
f(x,t)
x⁺
A
Wb
tb
z⁺
f(x, t)
$M(x,t) + \frac{\partial M(x,t)}{\partial x} dx$
M(x,t)
Q
$V(x,t) + \frac{\partial V(x,t)}{\partial x} dx$
w(x,t)
V(x,t)
$\rho A\, dx \frac{\partial^2 w}{\partial t^2}$
undeformed x axis
dx

Figure 1. Simple beam in transverse vibration and a free body diagram of a small element as it is deformed by a distributed force

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

Assuming the deformation to be small enough such that the shear deformation is much smaller than w(x,t) so that the sides of the element dx do not bend, a summation of forces in the y direction yields:

$$\left(\mathrm{V(x,t)}+\frac{\partial \mathrm{V(x,t)}}{\partial \mathrm{x}}\mathrm{dx}\right)-\mathrm{V(x,t)}+\mathrm{f(x,t)dx} = \rho \mathrm{A}\frac{\partial^2 \mathrm{w(x,t)}}{\partial \mathrm{t}^2}\mathrm{dx} \qquad (2)$$

Here V(x,t) is the shear force at the left end of the element dx, V(x,t)+V(x,t)dx is the shear force at the right end of the element dx, f(x,t) is the total external force applied to the element per unit length, and the term on the right side of the equality is the inertial force of the element.

Next the moments acting on the element dx about the z axis through point Q are summed. This yield:

$$\left(\mathrm{M(x,t)}+\frac{\partial \mathrm{M(x,t)}}{\partial \mathrm{x}}\mathrm{dx}\right)-\mathrm{M(x,t)}+ \left(\mathrm{V(x,t)}+\frac{\partial \mathrm{V(x,t)}}{\partial \mathrm{x}}\mathrm{dx}\right)\mathrm{dx}+\left(\mathrm{f(x,t)dx}\right)\mathrm{dx}=0 \qquad (3)$$

It is assumed that the rotary inertia of the element dx is negligible. Simplifying this expression yields ((dx)$^2$ is assumed to be almost zero):

$$\mathrm{V(x,t)}=-\frac{\partial \mathrm{M(x,t)}}{\partial \mathrm{x}} \qquad (4)$$

Substitution of this expression for the shear force into (2) yields:

$$-\frac{\partial^2}{\partial \mathrm{x}^2}\left(\mathrm{M(x,t)}\right)\mathrm{dx}+\mathrm{f(x,t)dx}=\rho \mathrm{Adx}\frac{\partial^2 \mathrm{w(x,t)}}{\partial \mathrm{t}^2} \qquad (5)$$

Further substitution of (1) into (5) and dividing by dx yields:

$$\rho \mathrm{Adx}\frac{\partial^2 \mathrm{w(x,t)}}{\partial \mathrm{t}^2}+\frac{\partial^2}{\partial \mathrm{x}^2}\left[\mathrm{EI_{cs}}\frac{\partial^2 \mathrm{w(x,t)}}{\partial \mathrm{x}^2}\right]=\mathrm{f(x,t)} \qquad (6)$$

If no external force is applied so that f(x,t)=0 and (6) simplifies so that free vibration is governed by:

$$\frac{\partial^2 \mathrm{w(x,t)}}{\partial \mathrm{t}^2}+\mathrm{c}^2\frac{\partial^4 \mathrm{w(x,t)}}{\partial \mathrm{x}^4}=0 \quad \mathrm{c}=\sqrt{\frac{\mathrm{EI_{cs}}}{\rho \mathrm{A}}} \qquad (7)$$

Equation (7) is based on the classical undamped Euler-Bernoulli beam theory.

To solve (7), w(x,t) can take the following expanded separated form with the chosen deflection mode shapes $\varphi_i(x)$ and the modal amplitudes $q_i(t)$:

$$w(x,t)=\sum_{i=1}^{\infty} q_i(t).\varphi_i(x) \qquad (8)$$

By substituting (8) into the equation of motion (7) and after rearrangement yields:

$$\mathrm{c}^2\frac{\varphi_\mathrm{i}''''(\mathrm{x})}{\varphi_\mathrm{i}(\mathrm{x})}=-\frac{\ddot{\mathrm{q}}_\mathrm{i}(\mathrm{t})}{\mathrm{q}_\mathrm{i}(\mathrm{t})}=\omega_\mathrm{i}^2, \qquad \mathrm{c}=\sqrt{\frac{\mathrm{EI_{cs}}}{\rho \mathrm{A}}} \qquad (9)$$

where the partial derivatives have been replaced with total derivatives. (Note: $\varphi''''=d^4\varphi/dx^4, \ddot{q}=d^2q/dt^2$ ). $\omega_i$ is the natural frequency of vibration of mode i.

The spatial equation comes from rearranging (9), which yields:

$$\varphi_\mathrm{i}''''(\mathrm{x})-\frac{\omega_\mathrm{i}^2}{\mathrm{c}^2}\varphi_\mathrm{i}(\mathrm{x})=0, \qquad \mathrm{c}=\sqrt{\frac{\mathrm{EI_{cs}}}{\rho \mathrm{A}}} \qquad (10)$$

By defining:

$$\lambda_\mathrm{i}^4=\frac{\omega_\mathrm{i}^2}{\mathrm{c}^2}=\frac{\rho \mathrm{A}\omega_\mathrm{i}^2}{\mathrm{EI_{cs}}} \qquad (11)$$

And considering the boundary conditions for this problem correspond to those of a clamped-free beam [8]:

$$\varphi_\mathrm{i}(0)=0, \qquad \frac{\mathrm{d}\varphi_\mathrm{i}}{\mathrm{dx}}(0)$$
$$\frac{\mathrm{d}^2\varphi_\mathrm{i}}{\mathrm{dx}^2}(\mathrm{L})=0, \qquad \frac{\mathrm{d}^3\varphi_\mathrm{i}}{\mathrm{dx}^3}(\mathrm{L})=0 \qquad (12)$$

General solution of (10) can be obtained as:

$$\varphi_\mathrm{i}(\mathrm{x})=\cosh(\lambda_\mathrm{i}\mathrm{x})-\cos(\lambda_\mathrm{i}\mathrm{x})-\sigma_\mathrm{i}(\sinh(\lambda_\mathrm{i}\mathrm{x})-\sin(\lambda_\mathrm{i}\mathrm{x}))$$
$$\sigma_\mathrm{i}=\frac{\sinh(\lambda_\mathrm{i}\mathrm{L})-\sin(\lambda_\mathrm{i}\mathrm{L})}{\cosh(\lambda_\mathrm{i}\mathrm{L})+\cos(\lambda_\mathrm{i}\mathrm{L})} \qquad (13)$$

The quantities $\lambda_i L$ are the real roots of the equation:

$$1+\cos(\lambda_i L)\cosh(\lambda_i L)=0 \qquad (14)$$

These quantities determine the natural frequencies of the beam as follows:

$$\omega_i = \sqrt{\frac{EI_{cs}}{\rho A}}.\lambda_i^{\,2} \tag{15}$$

Table (I) shows the calculated $\lambda_i$ and $\omega_i$ for the first four modes of vibration. The temporal equation comes from rearranging (9), which yields:

$$\ddot{q}_i(t) + \omega_i^2 q_i(t) = 0 \tag{16}$$

A simple procedure for including damping is to add it to the temporal equation after separation of variables. So the modal damping can be added to (16) as follows:

$$\ddot{q}_i(t) + 2\xi_i\omega_i\dot{q}_i(t) + \omega_i^2 q_i(t) = 0 \tag{17}$$

Where $\xi_i$ is the ith modal damping ratio. The damping ratios $\xi_i$ are chosen based on experience or on experimental measurements. Empirical results in [8] show that the damping factor of the first mode is 0.01. The higher order modes were assumed to have the same damping factor. The solution, for an under damped mode becomes:

$$\begin{aligned} q_i(t) &= A_i e^{-\xi_i\omega_i t}\sin(\omega_{di}t + \Phi_i) \\ A_i &= \left[\frac{(\dot{q}_i(0) + \xi_i\omega_i q_i(0))^2 + (q_i(0)\omega_{di})^2}{\omega_{di}^2}\right]^{1/2} \\ \Phi_i &= \tan^{-1}\frac{q_i(0)\omega_{di}}{\dot{q}_i(0) + \xi_i\omega_i q_i(0)} \end{aligned} \tag{18}$$

Where $A_i$ and $\Phi_i$ are constants to be determined by the initial condition. The robot link transverse displacement, approximately obtained, as a solution of the damped Euler-Bernoulli beam theory is finally expressed by [7]:

$$\begin{aligned} w(x,t) = \sum_{i=1}^{\infty} A_i e^{-\zeta_i\omega_i t}\sin(\omega_{di}t - \Phi_i) \times \\ \{\cosh(\lambda_i x) - \cos(\lambda_i x) - \sigma_i(\sinh(\lambda_i x) - \sin(\lambda_i x))\} \end{aligned} \tag{19}$$

## B. Robot Dynamic Model

In order to obtain a set of ODE of motion to describe the dynamics of the flexible link manipulator, N differential equations must be satisfied [8]:

TABLE I. CALCULATED $\lambda_i$ AND $\omega_i$ FOR FIRST FOUR MODES

| Modes | $\lambda_i$ | $\omega_i$ (rad/s) |
|---|---|---|
| First mode | 1.8751 | 9.7474 |
| Second mode | 4.6940 | 61.0838 |
| Third mode | 7.8547 | 171.0406 |
| Fourth mode | 10.9955 | 335.1738 |

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) - \frac{\partial L}{\partial q_i} = Q_i \ , \ i = 1,\ldots,N+1 \tag{20}$$

Where L is the so called Lagrangian which is given by:

$$L = T - U \tag{21}$$

T represents the kinetic energy of the system and U its potential energy that can be found in [8]. The schematic of the system is shown in Fig. 2.

The joint angle $\theta$ and modal coordinates $q_i$ can form a new vector $\tilde{q}$ of the system's generalized coordinates as:

$$\tilde{q}_{(N+1)\times 1} = [\theta\, q_1\, q_2 \cdots q_N]^T \tag{22}$$

Substituting the expressions for kinetic and potential energy into (20) and performing the required operations, one obtains the following matrix equation, which is a set of N+1 ode that model the dynamic behavior of the system:

$$M\ddot{\tilde{q}}(t) + H\dot{\tilde{q}}(t) + K\tilde{q}(t) = fv_m \tag{23}$$

$v_m$ is the voltage applied to the motor. The mathematical derivation of the mass (M), damping (H) and stiffness (K) matrices also the expressions for energies of one link manipulator can be found in Appendix (B).The parameter $f$ in (23) is defined as:

$$f_{(N+1)\times 1} = [k_u\, 0\, 0 \cdots 0]^T \tag{24}$$

where $k_u$ is the torque constant and $v_m$ is the voltage applied to the motor. The model in (23) does not include the effects of the piezoelectric films, which will be derived in the following section.

## C. Analysis of Beam-Piezoelectric Interaction

Piezoelectric film could be used as an actuator or a sensor by applying a voltage or measuring the open circuit voltage, respectively. If the piezoelectric film is used as a sensor, the open circuit voltage of the sensor is given by [9] as:

$$V_{oc}(t) = \sum_{i=1}^{n} Q_i^{V_{oc}} q_i(t) \tag{25}$$

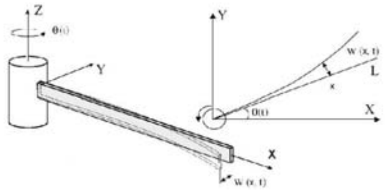


Figure 2. Schematic of the system

Where the ith modal coefficient is:

$$Q_i^{V_{oc}} = \frac{d_{31}E_cW_ct_b}{2C_c} \times \int_{a-m}^{a+m} \Xi_c(x)\frac{d^2\varphi_i(x)}{dx^2}dx \qquad (26)$$

The physical characteristics of the beam and piezoelectric films and motor for simulations are given in Appendix A. The variable $d_{31}$ is the transverse piezoelectric charge to stress ratio, $E_c$ is the Young's modulus of the film, $C_c$ is the capacitance of the piezoelectric film, $\Xi_c(x)$ is the shape function of the film, $W_c$ is the maximum width of the piezoelectric film, $t_b$ is the thickness of the beam. If the piezoelectric film is used as an actuator, its effect on the dynamic model is through the passive stiffness (negligible in this work) and the force produced by the actuator. The generalized force associated with ith mode is [9]:

$$F_i = f_i v_p \qquad (27)$$

$v_p$ is the voltage applied to the actuator. The coefficient of the generalized force is defined as:

$$f_i = -\operatorname{sgn}(z_c)\frac{d_{31}E_cW_ct_b}{2}\left(\int_{r_s^n}^{r_e^n}\Xi_c(x)\frac{d^2\varphi_i(x)}{dx^2}dx\right) \qquad (28)$$

Introducing the effect of piezoelectric films, the new system dynamic model can be described as:

$$M\ddot{\tilde{q}}(t) + H\dot{\tilde{q}}(t) + K\tilde{q}(t) = f_S v_m \qquad (29)$$

Where the new modal force coefficient matrix $f_S$, since one film were used as actuator in this work, becomes:

$$f_S = \begin{bmatrix} k_u & 0 & 0 & \dots & 0 \\ 0 & f_1 & f_2 & \dots & f_N \end{bmatrix}^T \qquad (30)$$

For modeling and control purposes it is convenient to write the model of the system in state-space form as follows:

$$\begin{aligned} \dot{x} &= Ax + Bu \\ y &= Cx \end{aligned} \qquad (31)$$

Where the matrices $A$ and $B$ are defined in terms of the stiffness, mass and damping matrices K, M and H, respectively, and the force vector $f_S$:

$$A = \begin{bmatrix} [0]_{(N+1)\times(N+1)} & I_{(N+1)\times(N+1)} \\ -[M^{-1}K]_{(N+1)\times(N+1)} & -[M^{-1}H]_{(N+1)\times(N+1)} \end{bmatrix}$$
$$B = \begin{Bmatrix} 0_{N\times1} \\ \{M^{-1}f_S\}_{N\times1} \end{Bmatrix}_{2N\times1} \qquad (32)$$

Where the state vector x includes the system's coordinates $\tilde{q}$ and $\dot{\tilde{q}}$ as:

$$x = [\tilde{q} \quad \dot{\tilde{q}}]^T_{2(N+1)\times1} \qquad (33)$$

The observation matrix C, which relates the state space to the output space, has to be obtained based on the available sensor. It includes the potentiometer encoder gain and the coefficients of the open circuit voltage for sensor that is expressed as:

$$C = \begin{bmatrix} \begin{bmatrix} k_\theta & 0 & 0 & \dots & 0 \\ 0 & Q_1^{V_{oc1}} & Q_2^{V_{oc1}} & \dots & Q_N^{V_{oc1}} \end{bmatrix} & [0] \end{bmatrix}_{2\times2(N+1)} \qquad (34)$$

Where its second row is obtained by evaluating (26) with the parameters of the sensor patch [8].

## III. Controller formulation

Model-based predictive control (MPC) is based on the principle of minimizing an objective function J that contains a vector of future errors e over a prediction horizon p, resulting in changes on m control actions every sampling instant. The error vector e is evaluated as the difference between a prediction of each output variable of the process and a set point trajectory r. If the controller focuses exclusively on set point tracking, it might choose to make large manipulated-variable adjustments. These could be impossible to achieve. They could also accelerate equipment wear or lead to control system instability [10]. Thus, the Model Predictive Controller also monitors a weighted sum of controller adjustments, calculated according to the following equation:

$$J = \begin{bmatrix} \Delta\underline{u}(0) \\ \dots \\ \Delta\underline{u}(p-1) \end{bmatrix}^T W_{\Delta u}^2 \begin{bmatrix} \Delta\underline{u}(0) \\ \dots \\ \Delta\underline{u}(p-1) \end{bmatrix} + \left(\begin{bmatrix} \hat{\underline{y}}(1) \\ \dots \\ \hat{\underline{y}}(p) \end{bmatrix} - \begin{bmatrix} \underline{r}(1) \\ \dots \\ \underline{r}(p) \end{bmatrix}\right)^T W_y^2 \left(\begin{bmatrix} \hat{\underline{y}}(1) \\ \dots \\ \hat{\underline{y}}(p) \end{bmatrix} - \begin{bmatrix} \underline{r}(1) \\ \dots \\ \underline{r}(p) \end{bmatrix}\right) \qquad (35)$$

In this work, there are two inputs to the system, one to the motor $v_m$ and one to the actuator $v_p$, and two outputs from the system, one is the motor or joint angular position which is measured by a high precision potentiometer that provides an analog voltage signal $v_\theta$, the second is the beam vibration which is measured by a piezoelectric sensor that provides an analog voltage output $v_{oc}$. So in (35), $\underline{u} = [v_m, v_p]^T$, $\underline{y} = [v_\theta, v_{oc}]^T$, $W_{\Delta u}$ and $W_y$ are weighting matrices as follows:

$$\begin{aligned} W_{\Delta u} &= \operatorname{diag}(w_{0,1}^{\Delta u}, w_{0,2}^{\Delta u}, \dots, w_{p-1,1}^{\Delta u}, w_{p-1,2}^{\Delta u}) \\ W_y &= \operatorname{diag}(w_{1,1}^{y}, w_{1,2}^{y}, \dots, w_{p,1}^{y}, w_{p,2}^{y}) \end{aligned} \qquad (36)$$

In this study, $W_y$ has unit diagonal terms so that all future errors are equally weighted while the diagonal terms in $W_{\Delta u}$ are set to the move suppression. The general MPC architecture can be illustrated in Fig. 3. The future inputs $\{\Delta \underline{u}(k \mid k), \ldots, \Delta \underline{u}(m-1+k \mid k)\}$ (optimal sequence) are calculated by minimizing the cost function subject to the following constraints on inputs to the motor and actuator: $-50 \le v_m \le 50, -200 \le v_p \le 200$.

## IV. Simulation Results

### A. Open loop

The open loop response simulation due to a step input on the motor is shown in Fig. 4. The open loop response simulation due to a step input to the piezoelectric actuator is shown in Fig 5. It is clearly seen from the open loop test that the piezoelectric actuator has very little effect on the motor's joint angle and the sensor signal is oscillating with the beam's natural frequency.

### B. Close loop

In the previous section we developed a model for the system to be controlled. In this section for simulation purposes we utilize a model as process for generating the data (see Fig. 3), considering four modes of vibration. Moreover the controller needs an approximate linear model to predict the future behaviors of the plant. We utilize a model as model predictor in Fig. 3 by considering only the first mode of vibration.

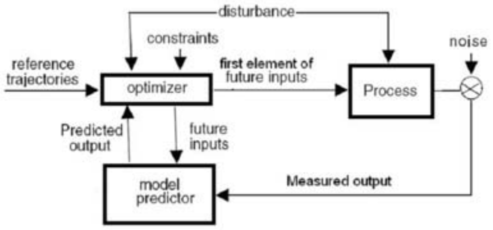


Figure 3. General MPC architecture

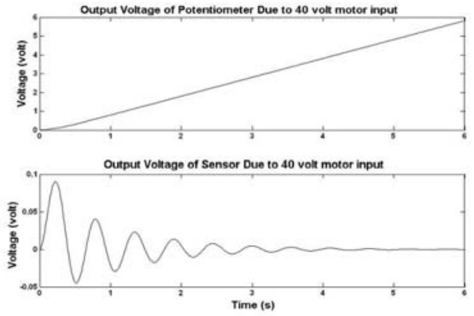


Figure 4. open loop test due to 40V motor input

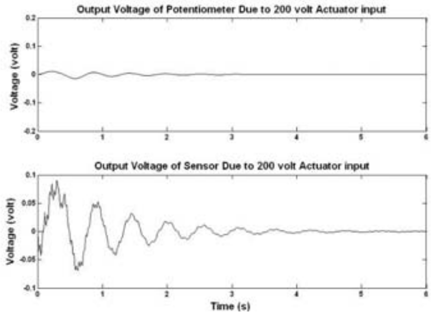


Figure 5. open loop test due to 200V actuator input

The value of prediction horizon p and control horizon m is based on the number of discrete sampling intervals required to reach within 95% of the plant output steady state. For many models, there is not much change beyond m=3~5, but *p* must be chosen as large as possible so that the value of p-m be greater than the settling time [11]. The motor is rotated and controlled to an angular set point using a multivariable predictive controller. During rotation, the beam’s vibration is suppressed until angular rotation of the motor has been completed. Fig. 6 and Fig. 7 illustrate simulation results comparison between predictive control and conventional PI control when positioning the manipulator to 57 degree (assumed prediction horizon for both output is p=70 and the control move horizon for both inputs is m=5). It can be seen the vibration is suppressed by motor and piezoelectric actuator effectively. Predictive controller signal to motor and actuator are shown in Fig. 8.

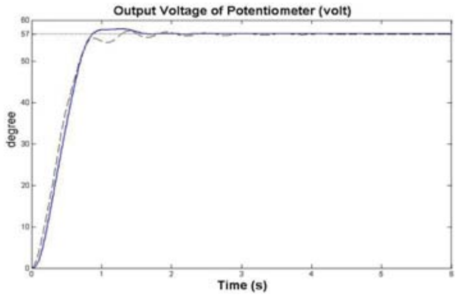


Figure 6. Joint angle comparison between MPC control (line) and PI control (dashed line)

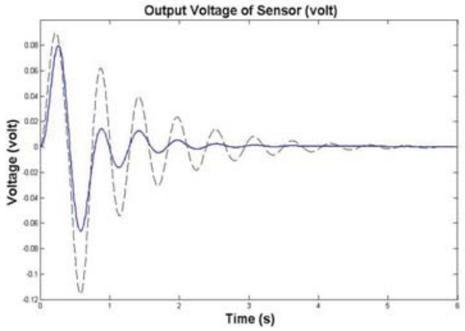


Figure 7. Displacement comparison between MPC control (line) and PI control (dashed line)

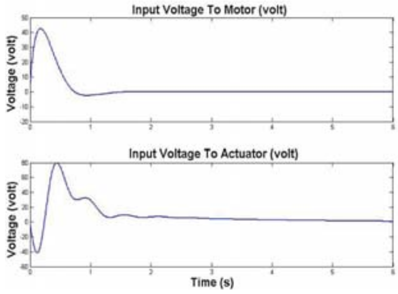


Figure 8. MPC controller signal to motor (upper) and to actuator (lower)Prepare Your Paper Before Styling

## V. CONCLUSIONS

As stated in the introduction, the aim of this paper is to present a vibration control design methodology based on predictive control strategy. This methodology is illustrated by the design of a controller for a flexible cantilever beam. The advantage of this methodology over existing methodologies is that adjustments can be made on the prediction of beam vibration that takes into account the effects due to nonlinearities in the system. This study successfully demonstrates that predictive control can be applied to suppress vibration on a flexible beam, making it an excellent candidate for future research on topics such as intelligent control using an array of sensors and actuators, as well as controller tuning specifically for other applications such as a multi-jointed flexible structures.

## APPENDIX A

This appendix contains the parameter values for the beam, motor and piezoelectric films.

TABLE II. BEAM PROPERTIES

| Material | Aluminum |
|---|---|
| Density $\rho$ (Kg/m$^3$) | 2700 |
| Young's modulus E (N/m$^2$) | $6.9\times10^{10}$ |
| Length×Width×Thickness (m) | 1×0.035×0.0019 |

TABLE III. MOTOR PROPERTIES

| | |
|---|---|
| Motor fixture inertia $J_h$ (Kg.m$^2$) | 0.14 |
| Friction coeff. $B_m$ (Nm/rad) | 0.95 |
| Motor torque constant $k_u$ (Nm/V) | 0.06 |
| Encoder gain $k_\theta$ (V/rad) | 0.3979 |

TABLE IV. PIEZOELECTRIC FILM PROPERTIES

| parameters | Sensor | Actuator |
|---|---|---|
| Charge constant $d_{31}$ (C/N) | $23\times10^{-12}$ | $175\times10^{-12}$ |
| Young's modulus $E_c$ (N/m$^2$) | $2\times10^9$ | $6.5\times10^{10}$ |
| Length×Width×Thickness (mm) | 250×35×1 | 150×35×1 |
| Capacitance $C_c$ (F) | $2.7\times10^{-6}$ | $2\times10^{-8}$ |
| Shape function $\Xi_c$ | 1 | 1 |

## APPENDIX B

$$M=\begin{bmatrix} J_h+\frac{\rho A L^3}{3} & \rho A\int_0^L \varphi_1(x)x\,dx & \rho A\int_0^L \varphi_2(x)x\,dx & \cdots & \rho A\int_0^L \varphi_N(x)x\,dx \\ \rho A\int_0^L \varphi_1(x)x\,dx & \rho A\int_0^L \varphi_1^2(x)dx & 0 & \cdots & 0 \\ \rho A\int_0^L \varphi_2(x)x\,dx & 0 & \rho A\int_0^L \varphi_2^2(x)dx & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho A\int_0^L \varphi_N(x)x\,dx & 0 & \cdots & \cdots & \rho A\int_0^L \varphi_N^2(x)dx \end{bmatrix}_{(N+1)\times(N+1)}$$

$$K=\begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & EI_{cs}\int_0^L(\varphi_1''(x))^2dx & 0 & \cdots & 0 \\ 0 & 0 & EI_{cs}\int_0^L(\varphi_2''(x)^2)dx & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & EI_{cs}\int_0^L(\varphi_N''(x))^2dx \end{bmatrix}_{(N+1)\times(N+1)}$$

$$H=\begin{bmatrix} B_m & 0 & 0 & \cdots & 0 \\ 0 & 2\zeta_1 m_{22}\omega_1 & 0 & \cdots & 0 \\ 0 & 0 & 2\zeta_2 m_{33}\omega_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 2\zeta_N m_{N+1,N+1}\omega_N \end{bmatrix}_{(N+1)\times(N+1)}$$

## REFERENCES

[1] C. K. Lee, F.C. Moon, "Modal Sensors/Actuators," J. Appl. Mech., 1990, Vol. 57, No. 2, pp. 434-441.

[2] P.T. Knotnic, S. Yurkvocich, U. Ozguner, "Acceleration Feedback for control of a Flexible Manipulator Arm," J. Robotic Syst., 1988, Vol. 5, No. 3, pp. 181-196.

[3] Y. Aoustin, C. Chevallereau, A. Glumineau, C.H Moog, "Experimental Results for the End-Effector control of a Single Flexible Robotic Arm," IEEE Trans Contr.Syst. Technol, 1994, Vol. 2, No. 4, pp. 371-381.

[4] H.T. Banks, R.C.H Del Rosario, H.T. Tran, "Proper Orthogonal Decomposition-Based Control of Transverse Beam Vibrations: Experimental Implementation," IEEE Trans Contr. Syst. Technol, Sep. 2002, Vol. 10, No. 5, pp. 717-726.

[5] Wei Chen, Markus Buehler, Gordon Parker, and Bernhard Bettig, "Optimal Sensor Design and Control of Piezoelectric Laminate Beams," IEEE Trans Contr. Syst.Technol., Vol. 12, No. 1, pp. 148-155, Jan. 2004.

[6] J. Daafouz, G. Garcia, and J. Bernussou, "Rubust Control of a Flexible Robot Arm Using the Quadratic D-Stability Approach," IEEE Trans Contr. Syst. Technol ., Vol.6, pp. 524-533, Sep. 1998.

[7] Daniel J. Inman, Engineering Vibration.Prentice-Hall, second edition 2000.

[8] R . Bravo, Vibration Control of Flexible Structures Using Smart Material, PhD. Dissertation, McMaster University,Canada,2000.

[9] A.F. Vaz, "Composite Modeling of Flexible Structures with Bonded Piezoelectric Film Actuators and Sensors," IEEE Trans on Intrumentation and Measurement, April 1998,Vol. 47,No. 2, pp.513-520.

[10] J.M. Maciejowski, Predictive Control with constraints, Prentice-Hall, first edition 2000.

[11] Camacho, Eduardo. Bordons, Carlon. Model Predictive Control, Springer, first edition 1999.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Real time Virtual Simulation of an Underactuated Pendulum-Driven Capsule System

Keattikorn Samarnggoon and Hongnian Yu

*Abstract*— **In this paper, a real time virtual simulation framework which is the foundation for studying human adaptive mechatronics (HAM) is proposed. This framework allows researchers to interact and experiment with the system in real time. Thus, motion control patterns can be identified and learned with, for example, a heuristic strategy. The prototype is developed with an underactuated pendulum-driven capsule robot model. Motion control patterns are identified and presented. The experimentation results demonstrate the proposed concept.**

***Keywords-human adaptive mechatronics, pendulum capsule robot, underactuated systems, virtual environment, real time system (key words)***

## I. INTRODUCTION

Human adaptive mechatronics (HAM) is defined as an intelligence human machine system in which the system can be self-adapted intelligently based on the current user competency level to obtain optimum performance [1–5]. To achieve the HAM requirements, there must be several mechanisms working together. The main components of a HAM system are human operators, the intelligent discrimination of operator actions, competency evaluation metrics, human machine interaction mechanisms, and the machine system.

The work presented in this paper is a part of HAM research which covers a real time virtual system for understand the functions of human operators in HAM which has many invaluable advantages. This kind of virtual simulation systems running in real time allows researchers to experiment with dynamic of the modelled system in an immediate and interactive manner. Robotic researchers usually design mechanical systems by modelling mathematical relation of system parts but there exists troublesome to find control patterns for human operating a robot. This issue can be overcome by the help of real time virtual simulation systems. Motion control patterns could be identified by trial and error (heuristic) experimentation strategies using this virtual system. Moreover, apart from the robot mechanical simulation itself, dynamics of the environment can also be integrated into the simulation, for examples, different frictions of ground areas, dynamic of fluid while controlling robot movements, and capsule bots moving on a simulated deformable surface in medical application robotics.

Virtual training is also capable with this real time simulation based on the human-in-the-loop concept of mechanical systems. Training scenarios can be implemented with little effort or at no cost. Measurement of performance improvements can also be done from the feedback within the environments. This allows users to practice as much as they want. As a result, the user learning curve could be improved drastically. Regarding training environment with virtual real time simulation, it is a novel concept called human adaptive mechatronics that could further help optimise the learning curve of a user while training by its assisting behaviours.

The main contributions of the paper are

- Proposing a real time virtual human and machine interactions framework. The proposed framework will be a basis for development and realisation of the HAM concept.
- Developing a human heuristic learning strategy for learning motion control patterns.
- Conducting the experimental tests to demonstrate the framework and the HAM concepts.

## II. RELATED WORKS

Human is considered the main component of the HAM systems because the aim of this system is a combination of an automatic control and adaptive manual control system which is operated by humans. Normally, humans are complex and unpredictable, but if they are involved in a goal oriented task, it is possible to recognise their intentions. Human has been long studied in many related fields e.g., neurophysiological, neuroscience, cognitive science, and psychophysical. In neuroscience study, Haynes and colleague successfully read human covert intention by decoding brain images from various sections simultaneously [6], [7]. The pattern recognition technique is used in decoding those human intentions by discriminate patterns from spatial information from various brain activity areas. This method of using spatial brain information is claimed to be more accurate than analysing only specific area of the human brain. The reason is that when human performing an activity, several of brain areas are working together according to its functions. Additionally, human intentions are influenced from personal experiences. This is indicated by Blakemore and Decety analysis of the evidences of brain activity [8]. The evidences show that when human perceive biological motions there exists brain activity that try to simulate these motions internally. As a consequence, this internal simulation would reflect as intentions in future actions. This basically works in the same way as training activity to improve personal experience.

Human has good abilities to learn, predict, and process information. However, these capabilities are depended on individual. A task that is performed by different persons might return different results because of individual ability. Individual ability is usually denoted by word ‘skill’ and the outcome from using skill to perform an action is called ‘performance’. Learning capability is another magnificence aspect of human being in which humans have learnt to improve their skills and as an overall result i.e., overall

This work has been supported by the European Erasmus-Mundus Sustainable eTourism project 2010-2359, the EPSRC UK-Japan Network on Human Adaptive Mechatronics Project (EP/E025250/1) and EU Erasmus Mundus Project-ELINK (EM ECW-ref.149674-EM-1- 2008-1-UK-ERAMUNDUS).

Keattikorn Samarnggoon* and Hongnian Yu are with Faculty of Computing, Engineering and Technology, Staffordshire University, UK. *He is currently a lecturer at College of Arts, Media and Technology, Chiangmai University, Thailand.

Email: {k.samarnggoon, h.yu}@staffs.ac.uk

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

performance improvement. The most important part that ruled all of these capabilities is the thinking inside the human brain. Consequently, as mentioned earlier, internal thinking would reflect out as the intentions to do a specified task. This intended output actions could be identified by pattern recognition techniques. The intention recognition is also considered as part of the HAM system.

For the intelligent machine to serve or adapt to human appropriately, it needs to know human intentions by estimating from various kinds of related information. Fortunately, sensor technologies have advanced significantly along with the matured field of pattern recognition. These two combinations are essential for online human intention recognition. Observations and measurements from sensors are the inputs to pattern recognition algorithms to identify or estimate human intention at time. There exist numbers of information to be monitored and measured which is depended on the type of tasks. For examples, patterns of force signals exert on an arm gripper are recognised to discriminate human operator actions when performs industrial weight loading operation using Hidden Markov Models [9], motion and velocity pattern profiles are the information used to classify human actions in telemanipulation tasks [10]. The identified actions are useful for switching among virtual fixture models which help in different mode of operations. Once the machine has ability to identify human intention in which step the human operator is performing. It is functionality of the next component of the HAM system to evaluate how well the performing competency.

The aim of competency evaluation is to measure how well the operator is performing a step of the task so that the next component of the HAM system can make adaptations for assisting the operator. A generic performance evaluation framework, human performance index (HPI), is proposed in [11]. The framework consists of two layers of evaluation. The first layer is the collection of performance variables that evaluate raw competency of actions. The second layer is the weighted conditional integral of those variables in the first layer for specific area of measurement e.g. speed, and accuracy. This layer is called performance criterions. Final performance conclusion, HPI, is then weighted and accumulated from the second layer values. On the other hand, this HPI measurement concept can be viewed as grading evaluation in education such as school. Evaluations such as paper works, examinations, and attendance are scored. These scores are weighted with different percentage values according to its importance. The subject's grade is calculated from these values. Grading point average (GPA) is finally calculated from weighted credits of each subject. Therefore, the HPI is viewed as the GPA while performance criterions are viewed as subjects, and raw evaluations are viewed as those scorings. In addition, this HPI framework could be used in two modes, open form and closed form. The open form is located at the second layer in which these performance criterions can be used in any applicable future closed form. The closed form is located at the final accumulation evaluations, HPI or GPA. Performance criterions such as speed and accuracy are the example of competency measurement metric. This metric is a basis for the next step of the HAM system, adaptive tuning.

Intelligent adaptation of the HAM system is tuned based on current operator competency. There are two types of adaptation i.e., passive and active adaptation. Tuning parameters inside the machine without interfering the operator is a passive adaptation [12]. An active adaptation works in the opposite way. It actively assists the operator by, as an example, pushing small amount of force to the controller grip to help achieving the aimed intention easily [13–15].

The basis system model for this paper is an underactuated modelling approach and a 6-step motion control strategy to develop a desired driving profile studied in [16].

Underactuated mechanical systems are a system that has less control inputs than degrees of freedom of the system to be controlled. This system may also occur in a full actuated system because it losses some freedom of control due to some reasons such as accident or system failure.

## III. PROPOSED REAL TIME VIRTUAL SIMULATION SYSTEM

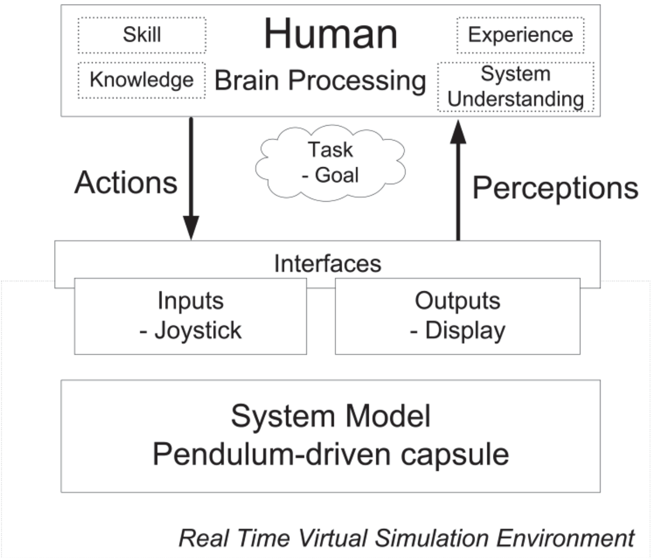


**Figure 1. Diagram of the proposed real time virtual simulation based on HAM.**

In this paper, the framework for the human machine controlling system in real time virtual simulation environment is proposed. Fig.1 shows a diagram and components of the system. The human operator interacts with a real time virtual simulation via the provided interfaces while perceiving information from the controlling system through a display monitor. It is the human operator's brain that processes information and orders the muscles to take actions to control an interface to manipulate the machine for accomplishing the desired task. Information is retrieved through various perception channels, e.g., eyes looking at meaningful data on the display screen, ears hearing the alert sound signal, and feeling of touching control interfaces. The human operator then observes, interprets, and processes this information and reacts with appropriate actions with the aimed goal in mind. Overall, these activities can be viewed as a human-in-the-loop control scheme and they are working together to be a system. Lacks of one of these components could cause the system failure.

The human block in the proposed framework diagram (Fig.1) acts as a controller that controls the underlying virtual simulation system. Loop of brain processing, perceptions, and actions that related to the human block is performed simultaneously. To control the system, the human operator first needs to know the goal of the controlling task. Then, the control strategy is planned to reach the goal. For example, the heuristic strategy is one of many strategy selections. Based on the planned strategy, series of actions are performed repeatedly. Outcome of each action may not be as planned but it can be adapted according to the situation because of adaptability of human. This process can be viewed as a learning process to control the system. It is individual skills that affect all blocks in the human related loop i.e., skill for perceptions, skill for information processing, and skill for conducting actions. These inputs (perceptions), outputs (actions), and internal flows (brain activities) work as a control system that interacts with the underlying virtual simulation environment.

There is a ‘task’ block located in conjunction between a human controller and the system (Fig.1). Task understanding is needed to be given first so that the human operator is able to plan actions ahead in mind. For example, the given task as controlling a robot to the right, an operator might think ahead about how to control to reach the given goal. Thus, it is very important to describe the task goal to the human operator.

The proposed real time virtual simulation environment needs software components to compose the system. These components are responsible to simulate the dynamical system, in this case the pendulum-driven capsule robot, to interface with the input system, to render the outputs to the display interface, and in the future functionalities; to recognise human intention and to calculate assisted tuning parameters and forces. The blocks component of this simulation environment from the software architecture point of view is shown in Fig.2.

Software architecture design for this proposed system in Fig.2 is designed centred on the following system functional requirements: 1) simulating dynamics in real time, 2) allowing the user to interact with the simulated dynamic via some controlling interfaces, 3) displaying adequate information for the user to perceive, 4) recognising and adapting the system behaviour based on the current user’s competency, and 5) logging and saving experiment data for future analysis. It starts with initial conditions and enters the main simulation loop with the aimed sampling time step. The simulation loop continues running until the software is terminated. Inside of the simulation loop, there are particular components executing to serve the whole functionalities of this virtual system. The ordinary differential equation solver, ODE Solver block, is used for solving ordinary differential equations with the implemented method and algorithm. The equations are based on the mathematical model of the mechanical system. The input system is responsible to handle an interface between the human operator and the virtual system. The input values from the device are transformed into the model’s input at every single step of simulation loops. The display output is drawn by the underlying graphic rendering system to visualize the simulating environment. Additional features such as the log system and the real time oscilloscope alike, the graphing system are essential for analysing immediate simulating values as well as logged values for later analysis. Realisation of HAM cannot be achieved without the following components; adaptation computation based on human intention and its corresponding competency, adaptation computation which is divided into passive and active tuning (Shaded blocks in Fig.2).

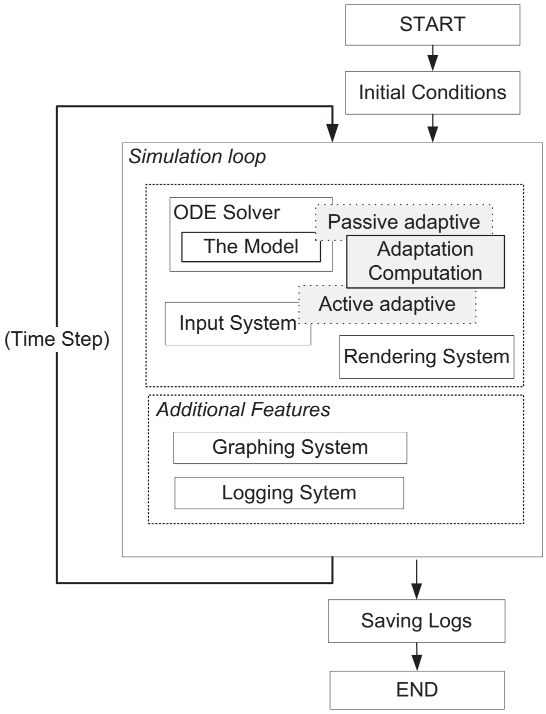


**Figure 2. The software architecture.**

## IV. MODEL OF THE DYNAMICAL SYSTEM

The schematic diagram of the underactuated pendulum-driven capsule system [16] shown in Fig.3 is adopted as a machine to the proposed virtual simulation system. M is mass of a capsule body. The mass m is on the top of the weightless link L. The link can rotate 360 degrees around the centre. One dimensional movement is defined by a position denoted by x and friction f is modelled to point in an opposite direction of the body movement base on the Coulomb’s friction model. The system is driven only by the force from the movement of the ball which is exerted by input torque τ and its moving momentum that causes forces.

The movement is possible because of both pendulum force and surface friction force.

From Fig.3, the ball position is defined in terms of cart position x at the centre as shown in equation (1). Then, the ball position equation is differentiated to get velocity and acceleration as in equations (2) and (3) respectively.

$$ball\ position = (x - Lsin\theta)\hat{\imath} + (Lcos\theta)\hat{\jmath} \quad (1)$$

$$ball\ velocity = \left(\dot{x} - L\dot{\theta}cos\theta\right)\hat{\imath} - (L\dot{\theta}sin\theta)\hat{\jmath} \quad (2)$$

$$ball\ acceleration = (\ddot{x} - L\ddot{\theta}cos\theta + L\dot{\theta}^2 sin\theta)\hat{\imath} - (L\ddot{\theta}sin\theta + L\dot{\theta}^2 cos\theta)\hat{\jmath} \quad (3)$$

Equation (3) and Newton's law of motion give forces from motion of pendulum ball in both x and y directions as follows.

$$F_{bx} = -m\ddot{x}_b \text{ , and } F_{by} - mg = m\ddot{y}_b$$

$$F_b = \begin{bmatrix} F_{bx} \\ F_{by} \end{bmatrix} = \begin{bmatrix} -m\ddot{x} + mL\ddot{\theta}cos\theta - mL\dot{\theta}^2 sin\theta \\ mg - mL\ddot{\theta}sin\theta - mL\dot{\theta}^2 cos\theta \end{bmatrix}$$

Also, the input torque to the joint is calculated as follows.

$$\tau = (-mLcos\theta)\ddot{x} + (mL^2)\ddot{\theta} - mgLsin\theta$$

$$F_{bx} - f = M\ddot{x}\ ; \quad \text{where } \ f = \mu N sgn(\dot{x})$$

$$N = Mg + F_{by}$$

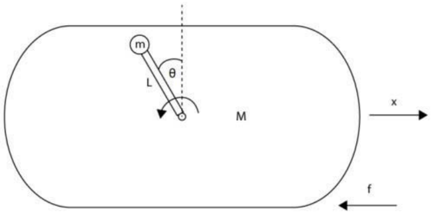


**Figure 3. Pendulum-driven capsule system.**

From above equations, we have

$$\ddot{x} = \frac{f\sigma_1 + \ddot{\theta}Lmcos\theta - \dot{\theta}Lmsin\theta}{M + m} \quad (4)$$

where $\sigma_1 = -g(M + m) + \dot{\theta}^2 Lmcos\theta + \ddot{\theta}Lmsin\theta$

$$\ddot{\theta} = \frac{Lmcos\theta\ddot{x} + \tau + gLmsin\theta}{L^2 m} \quad (5)$$

Equations (4) and (5) are the system equations with the single control input torque τ.

## V. IMPLEMENTATION OF REAL TIME SIMULATION

To conduct real time simulation, the forth order Runge Kutta numerical approximation method of ODEs [17] is used. From the system model (4) and (5), we have

$$\dot{v} = \frac{(2M + 2m)(\sigma_2 + \mu S\sigma_3) - \omega^2 Lmsin\theta}{(M + m)(2M + m - mcos2\theta - \mu Smsin2\theta)} \quad (6)$$

$$where\ \sigma_2 = \frac{cos\theta(\tau + gLmsin\theta)}{L}$$

$$and\ \sigma_3 = \frac{gLmsin^2\theta + \tau sin\theta}{L} - g(M + m) + \omega^2 Lmcos\theta$$

$$\dot{\omega} = \frac{(2M + 2m)(\tau + gLmsin\theta - \sigma_4)}{L^2 m(2M + m - mcos2\theta - \mu Smsin2\theta)} \quad (7)$$

$$where\ \sigma_4 = \frac{Lmcos\theta(\mu S(Mg + mg - \omega^2 Lmcos\theta) + \omega^2 Lmsin\theta)}{M + m}$$

$$\dot{x} = v \quad (8)$$

$$\dot{\theta} = \omega \quad (9)$$

where $S = sgn(\dot{x})$

Equations (6), (7), (8), and (9) are then solved by the forth order Runge Kutta numerical approximation algorithm.

An implementation of this real time virtual simulation system is developed using the industry leading application programming interface named Microsoft XNA and C# programming language. Sampling time is chosen at 10ms although it might change depending on the system performance but the system implementation is coded to compensate the issue by using elapsed time of each loop as a time step. The system parameters are as follows; M=0.5kg, m=0.05kg, L=0.3m, g=9.81m/s$^2$, μ=0.01 N*m/s.

The proposed real time virtual simulation system is controlled by the gaming joystick. The only system input is the amount of torque applied to the joint. The amount of torque can be varied by pushing an analogue stick in which its value is range between -1.0 and 1.0 N. In this case, the mapping is straightforward i.e. [-1.0, 1.0], value from an analogue stick is mapped to the input torque, τ, to drive the underactuated pendulum-driven capsule robot. However, it is noticed that the aimed system time step is 10ms. Therefore, the torque pushed by the joystick in real time is applied to the system at every time step of the system loop.

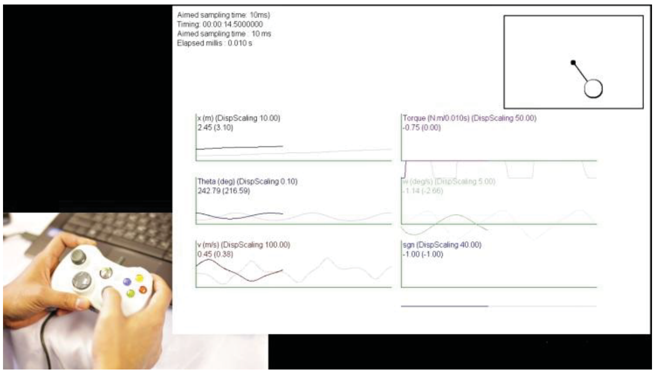
The screenshot of the simulation display is shown in Fig.4 when the system is simulated. The capsule body and its inner swinging shaft with the attached pendulum ball are displayed for the user to observe the capsule robot. Also, additional features for output information data are shown as online oscilloscope like a graphing system for both user observation and validation purposes.

Observations and manual controls are an inevitable couple in the human-in-the-loop control system. The proposed online simulation system displays necessary information on the monitor for observation while the user control amount of input torque via a joystick is shown in Fig.4. The user has an assigned task in mind while observing the pendulum movement on the screen and react to the dynamic behaviour of the system in real time to achieve desire control motions. In this case motion is in one dimensional movement i.e. moving to the left or vice versa.

Both input and output raw data during runtime experimentation of controlling are logged and saved for further analysis. Angle θ, angular velocity ω, capsule position x, capsule velocity v, and input torque τ are those variables that have been recorded. Also, an extra variable such as sign ($sgn\dot{x}$) of the friction term is logged for more clarification and validation of the implemented friction model.

## VI. Learning of Motion Control Patterns

One of the useful functionality of the real-time simulation system is apparent for heuristic strategy experimentation. In the following section, searches and results of motion control patterns for the pendulum-driven capsule system are presented. Control characteristics were experimented by the heuristic strategy. Ability to control this dynamical system is depended on the user's skill and understanding of the system. However, once understood, control characteristics can be identified and used as a pattern of control strategy.



Aimed sampling time: 10ms)
Timing: 00:00:14.5000000
Aimed sampling time : 10 ms
Elapsed millis : 0.010 s

x (m) (DispScaling 10.00)
2.45 (3.10)

Torque (N.m/0.010s) (DispScaling 50.00)
-0.75 (0.00)

Theta (deg) (DispScaling 0.10)
242.79 (216.59)

w (deg/s) (DispScaling 5.00)
-1.14 (-2.66)

v (m/s) (DispScaling 100.00)
0.45 (0.38)

sgn (DispScaling 40.00)
-1.00 (-1.00)

**Figure 4. User using the joystick to control virtual simulation system.**

The system initial values θ, ω, x, v, and τ are 180 degrees, 0 rad/s, 0 m, 0 m/s, and 0 N.m/s respectively. At the beginning the system stays still with the pendulum shaft and the ball lying straight down. When a small torque is applied, the pendulum begins to swing and the capsule start to move to the left and to the right repeatedly according to forces from the ball and the surface friction model as shown in Fig.5. The capsule is unintentionally displaced to the right by small torque after it finally comes to the steady state.

After several tries to control movement of the pendulum-driven capsule, the control strategy is developed. The system begins at the steady state and is intentionally controlled using the identified control patterns to move a capsule to the left and then to the right (Appendix 1). The identified control patterns to move a capsule by an input torque is summarised by the following strategies.

Step 1) Generate a torque by pushing the joystick to allow the pendulum to swing freely around, and then release the joystick (Fig.6).

Step 2) If one wants to move the capsule to the left, while the pendulum is freely swinging to the left side, the human operator needs to push the torque backward suddenly only in an appropriate short period of time. Moving to the right is done in the opposite way (Fig.7).

More precisely, to move to the left, the user needs to push the torque in the middle of rising or falling of angular velocity. In other words, one needs to push the torque at the edge of sine curves. These torque control strategies allow the user to control the pendulum driven capsule in the desired directions.

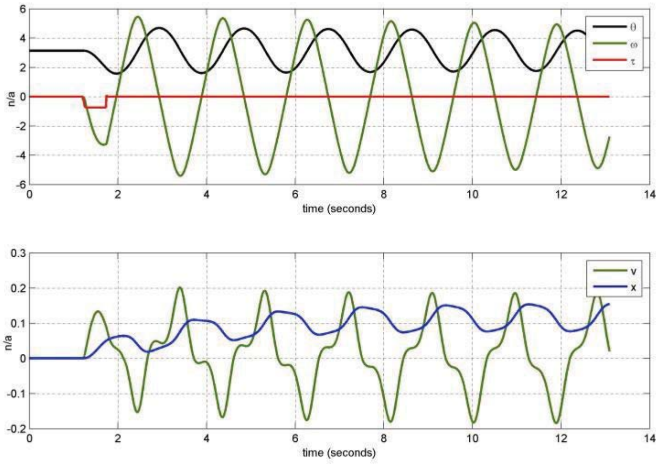


**Figure 5. Single pushed torque.**

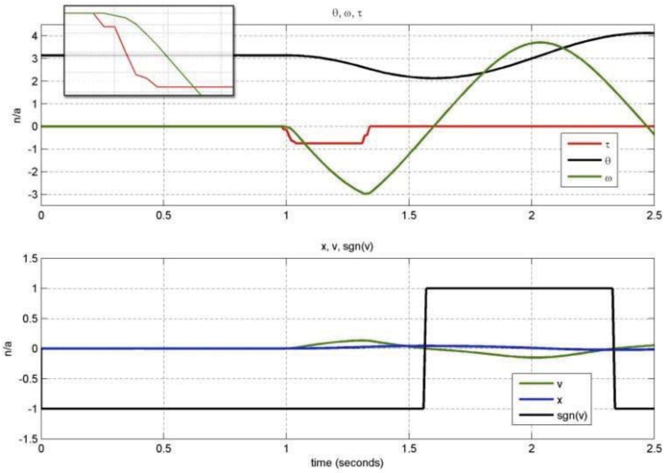


**Figure 6. Control characteristics for step 1.**

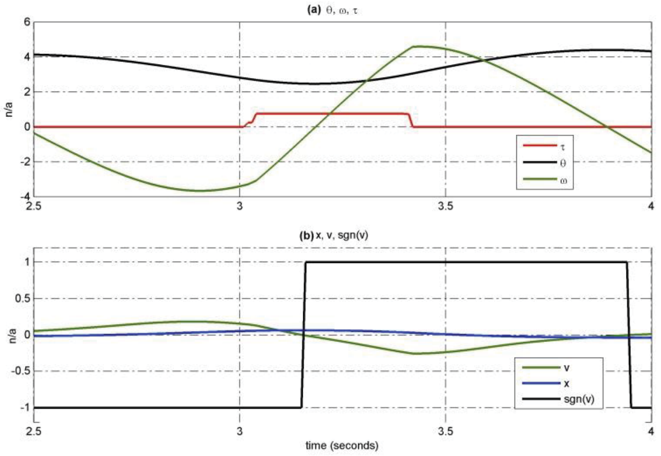


**Figure 7. Control characteristics for step 2.**

## VII. CONCLUSIONS AND FUTURE WORK

A framework of the human-in-the-loop control scheme using real time virtual simulation has been proposed. The software architecture and implementation of the underactuated pendulum-driven capsule robot system have been developed. Usefulness of real time simulation is apparent because of an interactivity nature of this type of systems. The system dynamic model can be realised experimentally. As a result, systematic motion control patterns can be identified. The system also exposes an important of human controlling ability. Different user controlling skills appear to be an important factor in the human-in-the-loop system control. The human controlling skill is depended on user's perceptions, brain processing of particular circumstances, and control actions. Overall performance of the system is another aspect compared to user skills that control the system.

The identified patterns of motion control for the joint torque seem similar to a walking cycle of human. The inverted bottom half circle of leg movements is shown in Fig.8. For example, given that desired movement is to move to the right, at first push the pendulum to swing freely from A to B and vice versa. At the moment that the pendulum ball nearly reaches point B, the torque should add in the opposite way. This will make the capsule move to the right because of both pushed torque and friction. This is working in the same way as human walking habits.

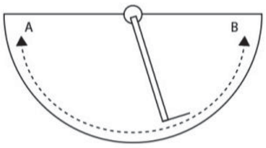


**Figure 8. Human walk cycle.**

In future works, closed loop control of an underactuated pendulum-driven capsule robot and a more complex model of double underactuated pendulum-driven robot [18] will be implemented as well as realization of an assisting control system based on human adaptive mechatronics. Also, the important adaptive mechanisms that would affect and optimise the learning curve of training will be experimented.

## REFERENCES

[1] S. Suzuki, "Human Adaptive Mechatronics," *Industrial Electronics Magazine, IEEE*, vol. 4, no. 2, pp. 28–35, 2010.

[2] H. Yu, "Overview of human adaptive mechatronics," in *Proceedings of the 9th WSEAS International Conference on Mathematics & Computers In Business and Economics*, 2008, pp. 152–157.

[3] F. Harashima and S. Suzuki, "Human adaptive mechatronics-interaction and intelligence," in *Advanced Motion Control, 2006. 9th IEEE International Workshop on*, 2006, pp. 1–8.

[4] "Guest Editorial," *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, vol. 225, no. 6, pp. 705–708, 2011.

[5] "Editorial," *International Journal of Modelling, Identification and Control*, vol. 4, no. 4, pp. 299–303, 2008.

[6] J. D. Haynes and G. Rees, "Decoding mental states from brain activity in humans," *Nature Reviews Neuroscience*, vol. 7, no. 7, pp. 523–534, 2006.

[7] J. D. Haynes, K. Sakai, G. Rees, S. Gilbert, C. Frith, and R. E. Passingham, "Reading hidden intentions in the human brain," *Current Biology*, vol. 17, no. 4, pp. 323–328, 2007.

[8] S. J. Blakemore and J. Decety, "From the perception of action to the understanding of intention," *Nature Reviews Neuroscience*, vol. 2, no. 8, pp. 561–567, 2001.

[9] V. Fernandez, C. Balaguer, D. Blanco, and M. A. Salichs, "Active human-mobile manipulator cooperation through intention recognition," in *Robotics and Automation, 2001. Proceedings 2001 ICRA. IEEE International Conference on*, 2001, vol. 3, pp. 2668–2673.

[10] W. Yu, R. Alqasemi, R. Dubey, and N. Pernalete, "Telemanipulation assistance based on motion intention recognition," in *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*, 2005, pp. 1121–1126.

[11] T. Parthornratt, R. Parkin, and M. Jackson, "Human performance index–a generic performance indicator," *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, vol. 225, no. 6, pp. 721–734, 2011.

[12] K. Tervo, "Human Adaptive Mechatronics Methods for Mobile Working Machines," Doctoral thesis, Department of Automation and Systems Technology, Aalto University, Espoo,Finland. 2010.

[13] K. Furuta, Y. Kado, S. Shiratori, and S. Suzuki, "Assisting control for pendulum-like juggling in human adaptive mechatronics," *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, vol. 225, no. 6, pp. 709–720, 2011.

[14] S. Suzuki and F. Harashima, "Assist control and its tuning method for haptic system," in *Advanced Motion Control, 2006. 9th IEEE International Workshop on*, 2006, pp. 374–379.

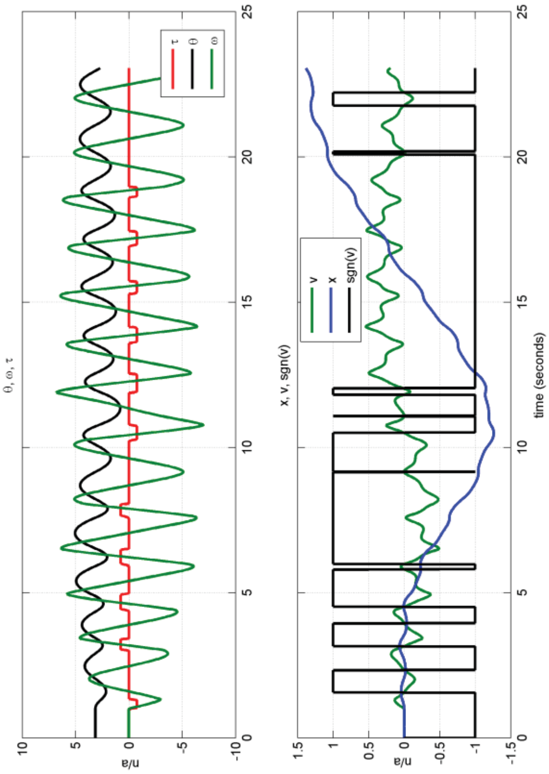
[15] S. Suzuki, K. Kurihara, K. Furuta, and F. Harashima, "Assistance control on a haptic system for human adaptive mechatronics," *Advanced Robotics*, vol. 20, no. 3, pp. 323–348, 2006.

[16] H. Yu, Y. Liu, and T. Yang, "Closed-loop tracking control of a pendulum-driven cart-pole underactuated system," *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, vol. 222, no. 2, pp. 109–125, 2008.

[17] J. C. Butcher and J. Wiley, *Numerical methods for ordinary differential equations*, vol. 2. Wiley Online Library, 2008.

[18] Y. Liu, H. Yu, and S. Cang, "Modelling and motion control of a double-pendulum driven cart," *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, vol. 226, no. 2, pp. 175–187, 2012.

## Appendix 1



**Graph of all logged variables of operation move to the left and then right.**

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# Modelling and Experimental Investigation of a Current to Pressure Converter

Tarik Saneecharaun[1], Dave Thompson[1], Marc Robertson[1]
[1]Norgren Ltd., Leeds (subsidiary of IMI plc.), UK

Pete Olley[2], Andrew Day[2]
[2]School of Engineering, Design & Technology, University of Bradford, UK

*Abstract*— **A Current-to-Pressure (I/P) converter is a device which converts a current input, typically between 4mA and 20mA, to a proportional pressure output. Such devices are often used in valve positioners and regulators to provide accurate and flexible control. This paper describes a physics based mathematical model of an I/P converter implemented using Matlab which is capable of simulating response under dynamic conditions. Detailed physical conditions such as thermal conduction and convection (and their effects on components), compressible orifice flow and inertial movement of components have been incorporated into the model using an explicit time-stepping lumped-parameter scheme. Parameters were obtained for the model using a series of novel experimental procedures. Experimental tests were carried out on a set of I/P converters over a range of operating temperatures and input current sequences. It is shown that the correlation between the simulated results (based on only measured physical characteristics of components) and the experimental test results is quantitatively accurate. It is further shown that the simulation allows the effects of significant design changes to be predicted, and that comparison between experimental and simulated results reveals areas where complex flow behaviour modifies pressure output significantly.**

***Keywords: Current-to-Pressure converter, Time-stepping, compressible flow***

## I. INTRODUCTION

Current-to-pressure converters are used to control the pressure of many applications in a wide range of industries because of their capacity for precise control and their dynamic behaviour. The dynamic modelling of a voice-coil I/P converter relies on the information obtained from actual testing done on existing devices and eventually can be used to validate the simulation model for various parameters which the system might endure during its operation.

The mathematical model has been developed based on governing equations which are determined by the flow, mechanical and electromagnetic behaviours of the system. The following sequential steps were used for the model: calculation of mass flow rate through each compartment, determination of the pressure in each compartment by considering thermal effects and finally the determination of the dynamics of mechanical elements. By means of various tests on specific components, key parameters were found which could be fed to the mathematical model and by making sensible approximation of non-critical parameters (for which information proved difficult to obtain), a working simulation was developed on Matlab to predict the dynamic behaviour of the I/P converter. The simulation is based on an explicit time-stepping lumped-parameter scheme.

There is a limited number of reported researches on the modelling and simulation of current-to-pressure converters; however the following references have proved valuable. There are different methodologies for developing and simulating pneumatic actuators which are controlled by proportional valves [1][3]. Simulation based on polytropic and bondgraph modelling are two different methods which can allow analysis of the dynamics of pneumatic actuators. The polytropic method does not take into consideration the effect of thermal heat transfer [2]. Validation of a simulation of a developed model is important. By means of experiments, it is possible to investigate how accurate the dynamics of a theoretical model is compared with pneumatic actuators [4].

A non-linear mathematical modelling which considers the 1) mass flow rate through restrictions, 2) pressure and temperature evolution in each compartment and 3) dynamics of the mechanical elements has been used in this paper [5]. Effects of non-linear flow through valves, and compressibility, are critical elements for accurate modelling [6]. To validate the results of simulation, testing was carried out on an actual I/P converter and the results were compared for quantitative analysis.

## II. MATHEMATICAL MODEL

An analysis of subsystems at a fundamental level was applied to derive the mathematical model; fluid, thermal and electro-mechanical interactions were incorporated as differential equations of motion. Considering the flow through the compartments of the I/P convertor, there are three points in the system which actively restrict the mass flow rate as illustrated in Fig.1; modelling must include the effects of compressibility as the pressures involved can exceed 5 bar.

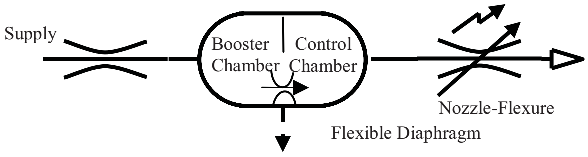


Figure 1: Simplified Schematic of I/P converter

Norgren Ltd., (Watson Smith) – A subsidiary of IMI plc.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

The governing equation for mass flow rate through a restrictor for a compressible fluid is given by (1).

$$\dot{m} = \frac{C_d P_{01}}{\sqrt{RT_{01}}} A_t \left(\frac{P_t}{P_{01}}\right)^{1/\gamma} \sqrt{\frac{2\gamma}{\gamma-1}\left(1-\left(\frac{P_t}{P_{01}}\right)^{\frac{\gamma-1}{\gamma}}\right)} \qquad (1)$$

for a Critical Pressure ratio >0.528, and

$$\dot{m} = \frac{C_d P_{01}}{\sqrt{RT_{01}}} A_t \, \gamma^{\frac{1}{2}} \left(\frac{2}{\gamma+1}\right)^{(\gamma+1)/2(\gamma-1)} \qquad (2)$$

for a Critical Pressure ratio ≤0.528,

where $\dot{m}$ is the mass flow rate, $C_d$is the discharge coefficient, $P_{01}$is the upstream pressure, $T_{01}$is the upstream temperature and $A_t$ is the cross sectional area of the throat.

The density, pressure and temperature in each compartment are linked by the ideal gas law:

$$P = \rho RT \qquad (3)$$

Where P is the Pressure, $\rho$ is the density and T is temperature.

The characteristic of the pressure, volume and density in each volume varies as a function of time, thus giving the following expression for the differential pressure:

$$\frac{dP}{dt} = \frac{d\rho}{dt} RT + \rho R \frac{dT}{dt} \qquad (4)$$

Heat transfer will have an effect on the overall system, especially considering the operating range of the I/P converter varies between -40°C and +85°C.

Heat transfer through conduction is modelled by considering the overall heat transfer coefficients, the contact area between pairs of components, and the temperature differences. This allows a calculation of heat flow rate between any given pair of components (the word ‘components’ is here used to include the external air or the flowing pressurized gas). Reasonable assumptions have been made for the heat transfer coefficients based on the types of material. Direct application of Fourier’s law of heat conduction by considering thermal conductivity proved quite difficult since many of the components within the I/P converter have an uneven shape and a varied surface finish. The rate of change of temperature from heat transfer follows from Newton’s Law:

$$\frac{dT_i}{dt} = \frac{1}{m_i C_i} \sum_j h_{ij} A_{ij} \Delta T_{ij} \qquad (5)$$

Where $m_i$ is the mass of individual component $i$, $C_i$ is the specific heat capacity of the component (or air), $h_{ij}$ is the overall heat transfer coefficient between adjacent components (or fluids) $i$ and $j$, $A_{ij}$ is the area of component $i$ in contact with $j$ and $\Delta T_{ij}$ is the temperature difference between the components.

Hence the temperature of each component, $i$, evolves in time according to

$$T_i(t+\Delta t) = T_i(t) + \frac{dT_i}{dt} \times \Delta t, \qquad (6)$$

where $\Delta t$ is the time step used in the simulation. Each volume contains mechanical elements which regulate the flow. A poppet valve controls the flow going into the booster region from a high pressure air supply and its distance of travel determines the throat area and thus the flow rate. The distance ($x$) it moves is dependent upon the balancing force created by the control volume pressure and the booster volume pressure. The mass of the poppet valve is assumed to be negligible and considering the elastic component controlling the movement, the following equation was derived:

$$\mathrm{x} = \mathrm{b}(\mathrm{P_c} - \mathrm{P_b}) + \mathrm{x_o} \qquad (7)$$

Where $x_o$ is the initial clearance of the ball valve and *b* is the opening per Pascal difference in the control and booster pressure, measured using a force gauge.

A voice coil is used to regulate the flow from the control volume which eventually controls the control pressure ($P_c$). An electromagnetic force, $F_e$ which is created by the voice coil is used to push a flexure towards a nozzle and by varying the input signal the amount of force can be controlled.

The electromagnetic force is given by:

$$F_e = d.I, \qquad (8)$$

Where *d* is the force per unit current of the solenoid (measured experimentally) and *I* is the input current.

The flexure behaviour is quite similar to a spring and obeys Hooke’s law.

$$F_x = k.\Delta x_2 \qquad (9)$$

Where, k is the stiffness of the flexure element and $x_2$ is the distance travelled by the flexure.

By considering the system in balance condition, the opening of the flexure from the nozzle is a function of the force of the pressure pushing against the flexure through a nozzle, the electromagnetic force and the mechanical (spring) force resulting from the flexure.

Neglecting spring stiffness of the flexure and aerodynamic effects, the (idealized) steady-state pressure of the control volume is given by a force balance between the magnetic force of the voice-coil pulling the flexure towards a nozzle ($dI$) and the control volume's pressure within the nozzle that pushes against this force :

$$P_c \frac{\pi D^2}{4} = dI \,, \tag{10}$$

where $P_c$ is the control volume pressure and $D$ is the diameter of the nozzle.

## III. SIMULATION

The mathematical model was implemented in Matlab to allow the simulation of the response under dynamic conditions. This was achieved by using a forward-difference method to time-step the differential equations. Various physical parameters have to be defined in the simulation model for accurate response; these physical parameters were obtained by means of experiments on crucial components.

The Matlab model has been defined such that it will generate graphs of the booster pressure ($P_b$) and the control pressure ($P_c$) with respect to time and input current. The following step inputs current have been used:

From 0-1s: 0mA
From 1-2s: 1mA
From 2-3s: 2mA
From 3-4s: 4mA
From 4-5s: 10mA
From 5-6s: 20mA

### *A. Simulation Results*

The initial values of the temperatures of each compartment and component were set at 293K. The fluid supply temperature set at 283K. Initial values of the booster volume pressure and the control volume pressure were set at1bar (absolute) and the supply pressure at 6bar (absolute). Fig.1 shows the simulation result and Fig.3 shows the simulation results which takes into consideration the volumes of the connecting pipes to the pressure sensors.

## IV. EXPERIMENTAL SET-UP AND RESULTS

Dynamic test was carried out at Norgren Ltd. where the booster pressure and control pressure were recorded. As illustrated in Fig.2, the I/P converter and a volumetric load cell from which the flow was supplied, were mounted inside an environmental chamber to maintain uniform temperature during the test. The current signal was generated from a programmable D.C. current supply. The pressure data from the pressure transducers were captured and processed by an oscilloscope.

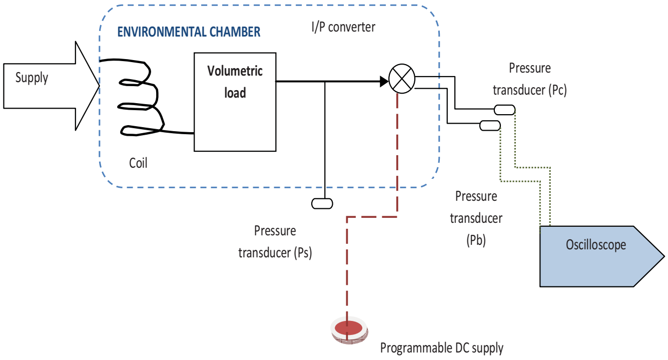


Figure 2: Experimental setup

Fig.4 shows the test results carried at a constant temperature of 293K.

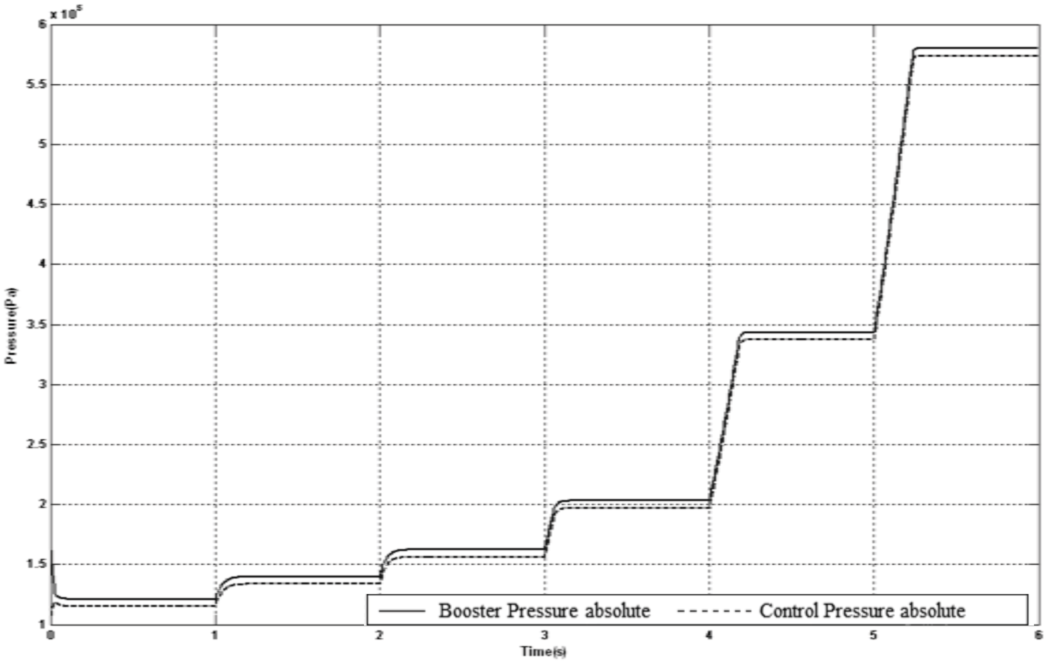


Figure 1:Simulation of output pressure as input current is increased in steps.

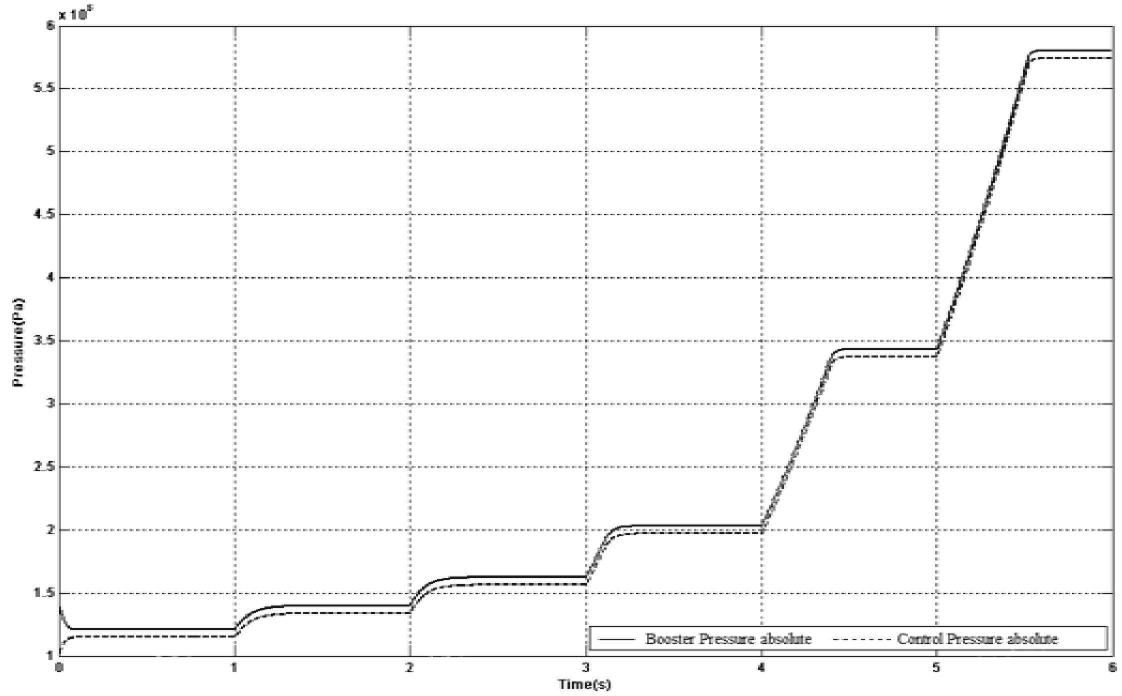


Figure 3: Simulation including volumes of connecting pipes

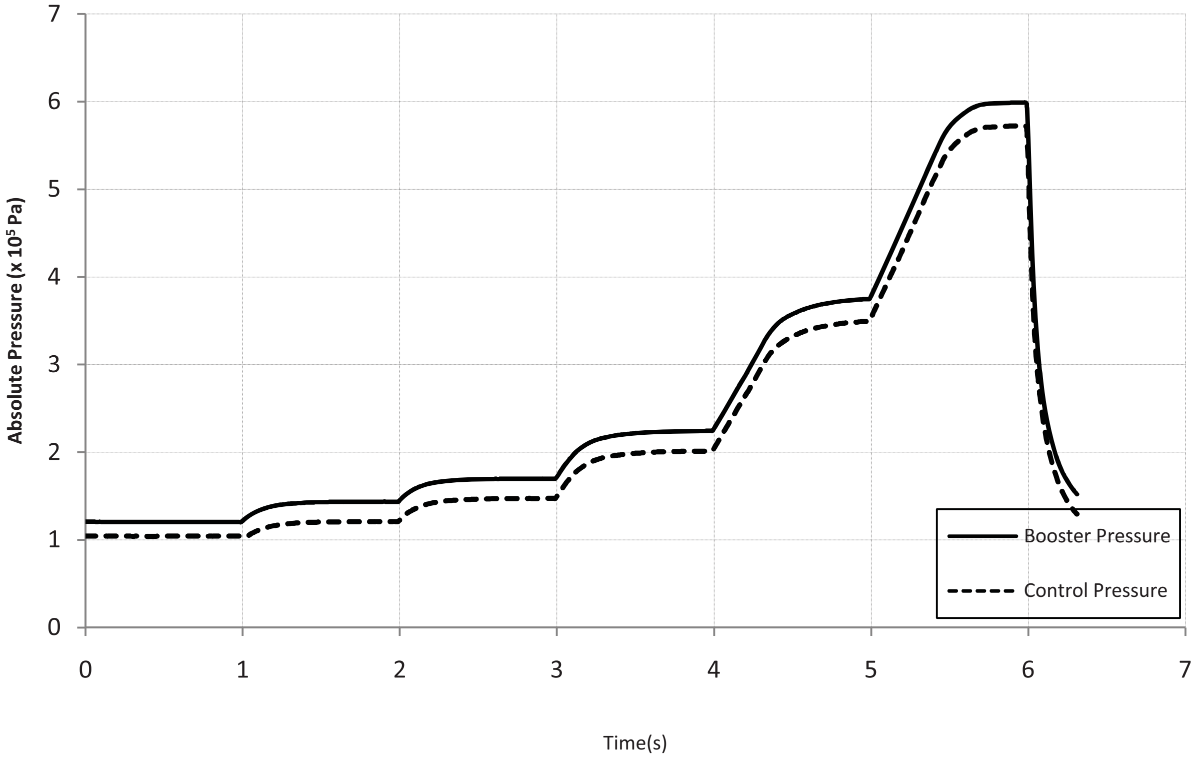


Figure 4: Test results for Pb and Pc

## V. Discussion

Table 1 and Table 2 show the test results versus the simulation results for the booster volume pressure and the control volume pressure respectively. From table 1 and table 2, it shown that the simulation results varies up to 14.3% for the booster volume pressures and up to 26.2% for the control volume pressures. Considering the normal range of input current which is *4-* to *20*-mA, the variation of the control volume pressures is within reasonable range.

TABLE 1: BOOSTER PRESSURE WITH PERCENTAGE VARIATION

| | **Booster Pressure (Bara)** | | |
|---|---|---|---|
| **I/mA** | ***Physical Test*** | ***Simulation*** | ***% variation*** |
| 0 | 1.2 | 1.372 | 14.3 |
| 1 | 1.43 | 1.581 | 10.5 |
| 2 | 1.7 | 1.818 | 7.1 |
| 4 | 2.23 | 2.148 | -3.6 |
| 10 | 3.71 | 3.527 | -4.9 |
| 20 | 5.98 | 5.879 | -1.7 |

TABLE 2: CONTROL PRESSURE WITH PERCENTAGE VARIATION

| | **Control Pressure (Bara)** | | |
|---|---|---|---|
| **I/mA** | ***Physical Test*** | ***Simulation*** | ***% variation*** |
| 0 | 1.04 | 1.312 | 26.2 |
| 1 | 1.21 | 1.521 | 25.7 |
| 2 | 1.47 | 1.756 | 19.5 |
| 4 | 2.01 | 2.087 | 3.8 |
| 10 | 3.49 | 3.465 | -0.7 |
| 20 | 5.71 | 5.812 | 1.8 |

It should be noted that from Fig. 1 and Fig. 4, that the response times for both booster volume pressures and the control volume pressures are quite different, this can be explained by the fact that the connecting pipes to the pressure sensors increase the volumes and eventually increases the response time. The additional volumes of the connecting pipes were included in the simulation run to give better accuracy of the results. This is shown in Fig. 3. It is important to highlight that the simulation will not give complete accuracy as there are assumptions and approximations made in the modelling, however comparison between simulation and experiment can lead to greater understanding of both.

## VI. CONCLUSION

This paper analyses how a mathematical model based on domains of fluid, mechanical and electromagnetic of a current-to-pressure can be developed in Matlab to simulate the dynamic response of the system. The simulation results show that the mathematical modelling of the system is effective over the operating range of input current. In addition, experiments were carried out in a controlled environment to validate the simulation model. Within acceptable levels of accuracy, the simulation has been shown to predict important behaviour of the pressures in the booster and control chambers with regards to changes made to the physical I/P converter.

## NOMENCLATURE

| Symbol | Description | Units |
|---|---|---|
| $A_i$ | Contact area | $m^2$ |
| $A_t$ | Cross sectional area of orifice | $m^2$ |
| b | Displacement per unit pressure | m/Pa |
| $C_d$ | Coefficient of discharge | |
| $C_i$ | Specific heat capacity | J/(kgK) |
| d | Electromagnetic force constant | N/A |
| $\gamma$ | Heat capacity ratio | |
| $h_i$ | Overall heat transfer coefficient | $W/(m^2 K)$ |
| I | Current | A |
| $\dot{m}$ | Mass flow rate | kg/s |
| $P_{01}$ | Upstream Pressure | Pa |
| $P_b$ | Booster volume pressure | Pa |
| $P_c$ | Control volume pressure | Pa |
| $P_t$ | Pressure at orifice | Pa |
| $\rho$ | Density | $kg/m^3$ |
| $T_{01}$ | Upstream temperature | K |
| $x_0$, x, $x_2$ | Displacement | m |

## ACKNOWLEDGEMENT

This work is supported by Knowledge Transfer Partnership (KTP) programme number 8693 and by Norgren Ltd (Watson Smith), a subsidiary of IMI plc.

## REFERENCES

[1] Hazem I. Ali, Samsul Bahari B Mohd Noor, Bashi S.M.,.Marhaban M.H, 2009. A Review of Pneumatic Actuators (Modeling and Control), Australian Journal of Basic and Applied Science, 3(2): 440-454

[2] Sorli M., Gastaldi L., E. Codina. and Heras S., 1999. Dynamic analysis of pneumatic actuators, Simulation
Practice and Theory, 7: 589-602.

[3] Hazem I. Ali, Samsul Bahari B Mohd Noor, S.M. Bashi, Mohammad Hamiruce Marhaban, 2009.
Mathematical and Intelligent Modeling of Electropneumatic Servo Actuator Systems, Australian Journal of Basic and Applied Sciences, 3(4): 3662-3670.

[4] Arcangelo, M., Nicola I.G. and Angelo G., 2005.
Experimenting and modelling the dynamics of pneumatic actuators controlled by the pulse width modulation (PWM) technique, Mechatronics, 15: 859-881

[5] French, L.G. and Cox C.S., 1988. The robust control of a modernelectropneumatic actuator, IFAC,
Automatic Control In Space.

[6] Edmond,R. and Yildirim H., 2001. A high performance pneumatic force actuator system, ASME, Journal of Dynamic Systems, Measurement and Control, 122(3): 416-425

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Research on fault diagnosis of TBM main bearing based on improved BP neural network

Tianrui Zhang, Lei Geng, Xianlei Chen, Tianbiao Yu,
Wanshan Wang
School of Mechanical Engineering & Automation
Northeastern University
Shenyang, China
tianjiangruixue@126.com

Xueting Fei
TBM Company
Northern Heavy Industries Group Co., Ltd
Shenyang, China
tianjiangruixue@hotmail.com

***Abstract*—Main bearing that plays the role of supporting and making the cutter to rotate and tunnel is the core part of TBM. Because of the harsh working conditions and complex changeful construction, the axial and radial load and environmental factors such as temperature of TBM main bearing are changing to make the fault of main bearing presenting randomness, and then the not easy identified fault may be produced. The traditional neural network model can not dynamic consider the cause of the reasons, parts and types, BP neural network fault diagnosis model based on fault reasons—signs matrix is presented in this paper. Firstly, the faults are screened through the fault reasons—symptom matrix, and the neural network structure is designed according to the screening result, then, the fault type is identified by way of model training. According to TBM main bearing fault symptoms data provided by a heavy enterprises practical engineering and MATLAB simulation validation, the feasibility and superiority of this model method are proved.**

***Keywords-TBM; BP neural network; fault diagnosis; signs matrix; MATLAB***

## I. INTRODUCTION

TBM is the main construction machinery of the underground tunneling construction, it can prevent soft foundation excavation face collapse and keep excavation face stable, at the same time, it can also finished the tunnel excavation and lining operations securely in machine[1-2]. At present, TBM is used widely in the large tunnel projects of city subway, river tunnel, and sea-crossing tunnel and so on. Because of the harsh working conditions and complex construction, the TBM fault rate is higher. Now, troubleshooting means still mainly rely on artificial maintenance, and timely fault repaired rate is low. The experiences show that a skilled technicians who to eliminate fault, need the time of total time of 70% to 90% to determine the fault reasons and parts, while only about 10%-30% for clearing the faults.

Main bearing that plays the role of supporting and making the cutter to rotate and tunnel is the core part of TBM[3]. In the Shiziyang tunnel engineering right line of China railway 12th bureau Guangshengang special passenger line, when the TBM was propelling, because of the bad lubrication and severe wear making the TBM appearing the abnormal vibration, the TBM was often at half stop condition, and the average working time was 2-5 hours. The construction progress was influenced seriously. Therefore, the significance is very great to ensure the security of TBM main bearing.

## II. ESTABLISHMENT OF THE FAILURE REASONS — SYMPTOM MATRIX

### A. *Main Bearing Fault Diagnosis Model Based on Boolean Function*

Assume that the fault symptoms of a common system are $S_1, S_2, \cdots, S_n$, and the reasons of the malfunction are $R_1, R_2, \cdots, R_m$. Fault diagnosis means that it infers one or a few possible failure causing reasons in the collection basing on a few signs in the collection of the observed fault symptoms. Therefore, we define a Boolean function $E(S_1, S_2, \cdots, S_n; R_1, R_2, \cdots, R_m)$, which is used to describe which combinations of fault symptoms appearing in which causes of the malfunctions; Define Boolean function $G(S_1, S_2, \cdots, S_n)$ to indicate the sign of a combination of the cause of the malfunction, namely the fault symptoms detected by the sensor.

Then define the Boolean function $C(R_1, R_2, \cdots, R_m)$, which is used to show that which combinations of the causes of the malfunctions can match with the existing sign combinations, namely the fault diagnosis which is looking for. The whole process of the fault diagnosis can be described the process of looking for the function $G(\bullet)$ in Formula (1):

$$C(\bullet) = G(\bullet) \bullet E(\bullet) \tag{1}$$

Failure causes and symptoms are binary Boolean relations, so Boolean Relation matrix is used to represent the function $E(\bullet)$ in the actual diagnosis, namely the failure reasons — symptom matrix.

Function $G(\bullet)$ is the fault symptoms collected through sensors and other detection means, which can be represented by Boolean vectors. So with the knowing of $G(\bullet)$ and $G(\bullet)$, we can get the diagnostic results $G(\bullet)$ through equation (1), $G(\bullet)$ is also a Boolean vector.

Fund project: 973 plan project, 2010CB736007

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

### *B. Establishment of Failure Reasons —Symptom Matrix*

The binary Boolean relation matrix between systems' failure causes and symptoms is used to describe Boolean function $E(S_1, S_2, \cdots, S_n; R_1, R_2, \cdots, R_m)$ in this paper.

An ordinary system, for example, the relationship matrix between failure causes and symptoms is:

$$\begin{bmatrix} I_{1,1} & I_{1,2} & \cdots & I_{1,j} & \cdots & I_{1,m} \\ I_{2,1} & I_{2,2} & \cdots & I_{2,j} & \cdots & I_{2,m} \\ \\ I_{i,1} & I_{i,2} & \cdots & I_{i,j} & \cdots & I_{i,m} \\ \\ I_{n,1} & I_{n,2} & \cdots & I_{n,j} & \cdots & I_{n,m} \end{bmatrix}$$

In the formula: $R_j$ is the fault type of number $j$; $S_i$ is the sign type of number $i$; $m$ is the type number of the possible cause of the malfunctions; $n$ is the type number of the detected signs.

$I_{i,j}$ is the possible state of sign $S_i$, when a fault $R_j$ occurs, and

$$I_{i,j} = \begin{cases} 1, & \left(S_i \textit{ may be caused by fault } R_j\right) \\ 0, & \left(S_i \textit{ may not be caused by fault } R_j\right) \end{cases}$$

At the same time, Let $s_j = \left(I_{1,j}, I_{2,j}, \cdots, I_{n,j}\right)$ is the sign vector produced by fault $R_j$; $r_i = \left(I_{i,1}, I_{i,2}, \cdots, I_{i,m}\right)$ is the fault vectors may exist when symptom $S_i$ occurs; $s' = \left(I_1', I_2', \cdots, I_n'\right)$ is the observation vector, If we can observe sign $S_i$, the value of $I_i'$ is 1, and 0 otherwise; $r' = (I_1, I_2, \cdots, I_m)$ is the result vector, when $I_j$ is 1, there may be a fault, on the contrary, there is no fault; $s_j$ $(j = 1, 2, \cdots, m)$ is the matrix's column vector, namely the vertical rules of a matrix. When $I_{i,j}$ is 1, it means the sign type $S_i$ may occur when the fault type $R_j$ occurs; And the sign type $S_i$ may not occur when $I_{i,j}$ is 0. $r_i$ $(i = 1, 2, \cdots, n)$ is the Matrix's row vector, namely the horizontal rules of a matrix. When $I_{i,j}$ is 1, it means the sign $S_i$ may be caused by failure $R_j$, otherwise, it isn't caused by failure $R_j$.

We collect the system's Fault signal (Fault symptoms), and then Make the assignment of each sign type whit 1 or 0, to determine the function $G(\bullet)$.

The results of the observation signs Depicted by $G(\bullet)$ is called the observation vector, which is represented by $s'$, and $s' = \left(I_1', I_2', \cdots, I_n'\right)$. If the sign $S_i$ can be observed, the value of $I_i'$ is 1 and 0, otherwise. Then we get an observation vector $s'$, namely we assign a value to the function $G(\bullet)$, we can get a result vector According to equation (1), which is represented by $r'$, $r' = (I_1, I_2, \cdots, I_m)$. When $I_j$ is 1, that may be a problem, and there is no fault on the contrary. This method is the theoretical model of the matrix screening method.

## III. BP NEURAL NETWORK MODEL OF THE MAIN BEARING

BP network is a multilayer feed forward neural network, and it is composed of input layer, hidden layer and output layer. Figure 1 shows the structure of the BP network, Neural network learning uses BP algorithm, The learning process are composed of prior to the calculation process and the error back-propagation process, In the prior to the calculation process, the input information calculates layer by layer from the input layer to the hidden layer, and transmit to the output layer, the state of each layer's neuron only affect the status of the next layer's neurons. If the output layer can not get the desired output, it transfers to the error back-propagation process, and the error signal returns along the original connection path, then makes the network system error minimization by modifying the value of every layer's neuron. Finally, the actual outputs of the network are approaching to their corresponding desired output.

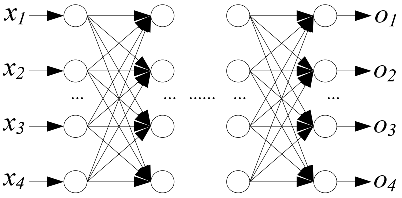


Figure 1. BP network model

To make the algorithm implementation process clearly, we can use the flow chart 2 to represent. The specific steps are as follows:

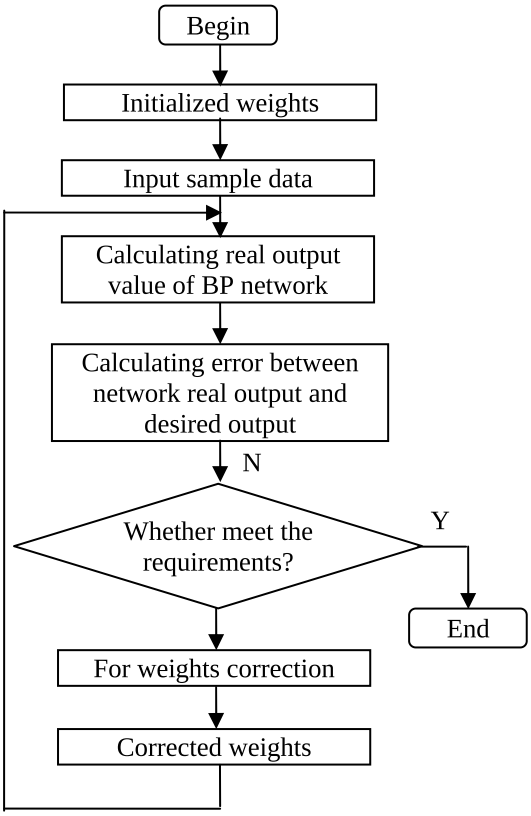

Begin
Initialized weights
Input sample data
Calculating real output value of BP network
Calculating error between network real output and desired output
N
Whether meet the requirements?
Y
End
For weights correction
Corrected weights

Figure 2. Flow chart of BP algorithm

- Initialize the weights $W$ and the threshold $\theta$, namely assign the weights $W_{ij}$ Connecting input layer units and the hidden layer, the weights connecting the hidden layer and output layer, the threshold of hidden layer $\theta_j$ and the unit threshold of output layer $\theta_k$ with a smaller value between 0 and l.
- Provide a learning sample pair (Input and expected output values), give the input vector $X_i = (x_1, x_2, \dots x_n)$ and the expected output vector $\hat{Y}_i = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_m)$ , which corresponds with the input one, import value $x_i$ to input and output layer node, and calculate forward one by one:

$$x'_j = f(\sum_{i=0}^{n} w_{ij} x_i - \theta_j) \qquad (j = 1,2,\dots,n) \qquad (2)$$

$$y_k = f(\sum_{k=1}^{n} v_{jk} x'_j - \theta_k) \qquad (j = 1,2,\dots,n) \qquad (3)$$

- Calculate the error { $\delta_k$ } between the output value { $y_k$ } of the output node and the expected value { $\hat{y}_k$ }.

$$\delta_k = y_k(1-y_k)(\hat{y}_k - y_k)$$

- Allocate error to the hidden layer nodes Reversely, namely use the connect right{ $V_{jk}$ }、the generalization error of the output layer{ $\delta_k$ }、output of the hidden layer { $x'_j$ }to calculate the error{ $\delta_J$ } of hidden layer's every unit:

$$\delta_j = x'_j(1-x'_j)\sum_{k=1}^{n} v_{jk}\delta_k \qquad (4)$$

- Use the generalization error { $\delta_k$ } of the output layer unit and the output { $x'_j$ } of hidden layer's each unit to correct the output layer's weights { $V_{jk}$ } and thresholds { $\theta_k$ }:

The weights correction of output layer and hidden layer:

$$v_{jk}(t+1) = v_{jk}(t) + \eta\delta_k x'_j \qquad (5)$$

The threshold correction of output layer:

$$\theta_k(t+1) = \theta_k(t) + \eta\delta_k \qquad (6)$$

- Use the generalization error { $\delta_j$ } of hidden layer and input { $x_i$ } of input layer's each unit to correct the connection weight { $W_{ij}$ } and threshold { $\theta_j$ }:

Correction of connection weights between the input layer and hidden layer:

$$w_{ij}(t+1) = w_{ij}(t) + \eta\delta_j x_i \qquad (7)$$

The threshold correction of the hidden layer:

$$\theta_j(t+1) = \theta_j + \eta\delta_j \qquad (8)$$

- Repeat steps two select different training samples, implement the above iterative process constantly, until meet the requirements, makes the error is small enough or zero, stop learning.

## IV. Realization of the Fault Diagnosis for Main Bearing

There are three steps in the process of TBM fault diagnosis. Firstly, collect different vibration signal by the data acquisition card. Then, eliminate the noise of vibration signal by wavelet, and pick up the Eigen value. Finally, determine the fault type and remove the faults by fault diagnosis.

Applying TRFDS based on BP neural network to diagnose faults includes the following steps:

*1) Using vibration data about main bearing from TBM produced by one Heavy equipment company in Shenyang, we can establish the faults cause—symptoms matrix.* Due to the limited space, it can only take the 6*6 sub matrix. The values of fault symptoms function $G(\bullet)$ by the data collecting are $s' = (1 \;\; 0 \;\; 1 \;\; 0 \;\; 0 \;\; 0)$. Then by formula (1):

$$\begin{aligned} r' &= C(\bullet) = G(\bullet) \bullet E(\bullet) = \\ &(1 \;\; 0 \;\; 1 \;\; 0 \;\; 0 \;\; 0)\begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \\ &= (1 \;\; 1 \;\; 1 \;\; 0 \;\; 1 \;\; 1) \end{aligned}$$

Then we can know that, $r' = (1 \;\; 1 \;\; 1 \;\; 0 \;\; 1 \;\; 1)$ shows when the value of the function $G(\bullet)$ is $s' = (1 \;\; 0 \;\; 1 \;\; 0 \;\; 0 \;\; 0)$, the possible fault types are in the table 1.

*2) To design neural network structure.* According to the faults cause—symptoms matrix, we can determine the node number of input and output layer in neural network. If the vibration signal was divided into N layers, the input layer node number is N; if there are M kinds of equipment fault, the output layer node number is M; the hidden layer node number is chosen by experience. And the transfer function is single polarity S function. According to successful experience of neural networks in fault diagnosis, we can determine the following two preliminary conclusions:

*a) Most problems of the fault pattern recognition could be resolved by the three-layer network.*

*b) There is the approximate relationship between the neurons number of hidden layer and the input layer in three-layer network* $N_2 = 2N_1 + 1$.

*3) Training the neural network.* The extracted eigenvector is input neural network as training samples, and it assume that there are $M$ kinds of equipment fault, then the output of the network is $\hat{Y}_i = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_m)$ , if the equipment is in $j$ state, the output of the network is $\hat{y}_j = 1$ , and the rest is 0, finally the output of network is $(0, 0, 0, ..., 1, ..., 0)$ . Using sample data to train the network until the simulation error is less than the target error, and save the network weights and bias after training.

TABLE I. THE VIBRATE FAULTS CAUSE—SYMPTOMS MATRIX OF MAIN BEARING OF TBM

| Fault symptoms | Fault causes | | | | | |
|---|---|---|---|---|---|---|
| | Unbalance | Touch grinding | Not centring | Rotor damaged | Surge | Bearing pedestal loosened |
| Vibration value whether stable | 1 | 1 | 1 | 0 | 1 | 1 |
| Vibration value whether mutations | 0 | 0 | 0 | 0 | 0 | 1 |
| Axial vibration whether obvious | 0 | 0 | 0 | 0 | 1 | 0 |
| cutter pressure fluctuations | 1 | 0 | 0 | 1 | 1 | 0 |
| Vibration value with the oil temperature changes | 0 | 1 | 0 | 0 | 0 | 0 |
| Vibration value with engine speed change | 0 | 0 | 0 | 1 | 1 | 1 |

*4) Identifying the fault type.* The eigenvector of vibration signals are put into neural network and trained BP neural network is used to solve the problems, then to identify the fault type according to the output of network.

The six typical faults, including unbalance, touch grinding, rotor damaged, surge, bearing pedestal loosened, are the neural network output; six spectrum peak energy value of frequency spectrum from vibration signal spectrums are used as characteristic variable to form the training samples. Shown as table 2 and table 3.

TABLE II. MAPPING TABLE OF OUTPUT NODES AND FAULT TYPES

| Output nodes | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Fault types | unbalance | touch grinding | not centring | rotor damaged | surge | bearing pedestal loosened |

TABLE III. MAPPING TABLE OF INPUT NODES AND FREQUENCY RANGE

| Input nodes | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Frequency range | 0-0.5f | 0.5-0.99f | 1f | 2f | 3-5f | High frequency range |

To avoid network saturation, the value of spectrum peak energy value cannot be used as input vector directly, so the data must be normalized before network training. Each input is normalized as following:

$$\overline{E}=\frac{E}{\sum_{i=1}^{n}\left|E_i\right|} \qquad i=1,2,...,6$$

In this way, the input values of network are all among [0, 1].

The membership degree vector of faults is obtained by membership function through measured data. For example, like the fuzzy relation of the ‘large vibration’ and ‘fault severity’, according to the experience, the fault is obvious when vibration is small. So it is appropriate to use Half Causy distribute function in fault diagnosis, the mathematic representation is:

$$\mu(x)=\begin{cases} 0 & (0\le x\le a) \\ \dfrac{k(x-a)^2}{1+k(x-a)^2} & (a\le x\le\infty) \end{cases} \tag{9}$$

In the problems we researched, $a=0$ , that is effective for positive region. The formula (9) can be simplified as follows:

$$\mu(x)=\frac{kx^2}{1+kx^2}=\frac{x^2}{1/k+x^2}$$

In the formula, the influence of $k$ to $\mu(x)$ can be obtained by formula (10): when $k$ is the maximum, that is $k=1$ , the influence of $k$ to $\mu(x)$ is very small; but $k$ is too small, the influence of $k$ to $\mu(x)$ is still very small. They all do not reflect the membership degree, so $k$ depends on the value of $x$ .

$$k=\frac{1}{\left(\sum_{i=1}^{n}\frac{x_i}{n}\right)} \tag{10}$$

In the formula, $n$ is the number of not equal to 0.

TABLE IV. TRAINING SAMPLE AND TARGET OUTPUT

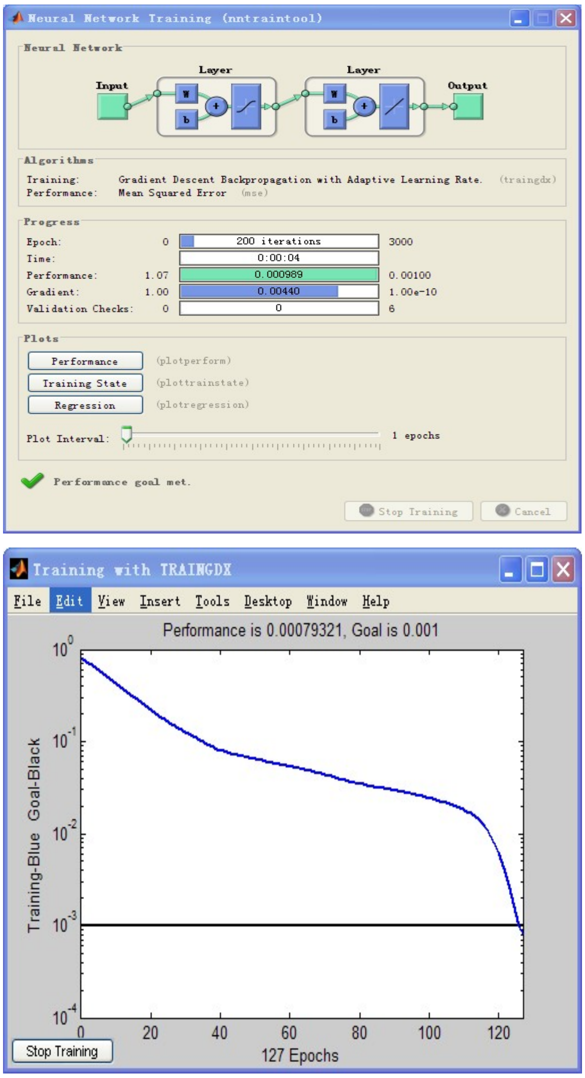
| Input sample | | | | | | Expected output | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.00 | 0.90 | 0.05 | 0.05 | 0.00 | 0.00 | 1 | 0 | 0 | 0 | 0 | 0 |
| 0.10 | 0.10 | 0.10 | 0.20 | 0.10 | 0.30 | 0 | 1 | 0 | 0 | 0 | 0 |
| 0.00 | 0.10 | 0.70 | 0.10 | 0.10 | 0.00 | 0 | 0 | 1 | 0 | 0 | 0 |
| 0.60 | 0.06 | 0.16 | 0.18 | 0.00 | 0.00 | 0 | 0 | 0 | 1 | 0 | 0 |
| 0.20 | 0.10 | 0.60 | 0.05 | 0.05 | 0.00 | 0 | 0 | 0 | 0 | 1 | 0 |
| 0.90 | 0.00 | 0.00 | 0.00 | 0.10 | 0.00 | 0 | 0 | 0 | 0 | 0 | 1 |

It chose the three layer network, 6 input units, 13 middle layers, and 6 output units. Besides, the error coefficient is set as 0.0001. Simulation result of the MATLAB is displayed as the figure 3; the actual output is displayed as table 5.

After the wavelet neural network has finished the study for the input and output sample, we can use it to the fault diagnosis on the bearings. Assume that a new vibration signal $[0.00, 0.15, 0.65, 0.10, 0.10, 0.00]$ which is put into neural network as Eigenvectors, then we can get the results [-0.0315, 0.1574, 0.6221, 0.0329, 0.2245, -0.0626] by nonlinear mapping in neural network. From the output results, it is obvious that the maximum is 0.6221 , which correspond the third fault type. So the fault signal is in line with the type of not centring fault.

TABLE V. ACTUAL OUTPUT OF TRAINING SAMPLE

| Actual output | | | | | |
|---|---|---|---|---|---|
| 1.007145 | 0.008596 | −0.00033 | 0.03914 | −0.04347 | −0.00267 |
| 0.002015 | 0.997113 | 0.030207 | 0.039964 | −0.06028 | 0.064406 |
| −0.00694 | −0.00297 | 1.002852 | −0.00296 | 0.000423 | 0.009778 |
| 0.008455 | 0.012074 | −0.01246 | 1.002135 | 0.001375 | −0.01221 |
| −0.02881 | −0.02597 | 0.003585 | −0.04818 | 1.062456 | 0.032021 |
| 0.025081 | 0.021992 | 0.001093 | 0.057425 | −0.06466 | 0.971879 |



Training: Gradient Descent Backpropagation with Adaptive Learning Rate. (traingdx)
Performance: Mean Squared Error (mse)

Epoch: 0 200 iterations 3000
Time: 0:00:04
Performance: 1.07 0.000989 0.00100
Gradient: 1.00 0.00440 1.00e-10
Validation Checks: 0 0 6

Performance (plotperform)
Training State (plottrainstate)
Regression (plotregression)

Plot Interval: 1 epochs

✓ Performance goal met.

Performance is 0.00079321, Goal is 0.001

Figure 3. Simulation result of the MATLAB

## V. CONCLUSIONS

Real-time fault diagnosis of TBM main bearing can improve the efficiency of the project construction, and ensure the construction period. This paper put forward BP neural network fault diagnosis model based on the fault reasons—symptom matrix. Through practical engineering data validation and MATLAB simulation, it confirmed that this fault diagnosis model of TBM main bearing is more feasible and advantaged than the common BP neural network model.

1) The fault reason of TBM main bearing is dynamic changing, so the scope of the fault types can be further narrowing through the fault reasons—symptom matrix screening, and make the BP neural network to lock the fault type rapidly in the training process. Then the problem that the traditional algorithm can't dynamic diagnosis main bearing fault is solved.

2) The training process of neural networks can be real-time tested by MATLAB simulation, to ensure that the training results are accurate feasible.

3) The algorithm has shorten the fault diagnosis time and improved the working efficiency of TBM.

## REFERENCES

[1] Yang B S, Lim D S, Tan A C C., "VIBEX: anexpert system for vibration fault diagnosis of rotating machinery using decision tree and decision table," Expert Systems with Applications, Vol. 28, pp. 735-742, 2005.

[2] Tian X, Lin J, Zuo M, et al. "Vibration signature database for rotating equipment fault diagnosis," Alberta, Edmonton: Department of Mechanical Engineering , University of Alberta , 2002.

[3] Kim K O, Zuo M J. "Two fault classification methods for large systems w hen available data are limited," Reliability Engineering and System Safety, Vol.92, pp.585-592, 2007.

[4] Babnik T, Cubina R. Two approaches to power transformer fault classification based on protection signals [J]. Int. J. Electr Power Energy Syst., 2002 (24): 459-468.

[5] Cen Nan, Faisal Khan, M. Tariq Iqbal. "Real-time fault diagnosis using knowledge-based expert system," Process safety and environmental protection, Vol. 86, pp.55–71, 2008.

[6] Qian Feng, "Research on fault diagnosis method of TBM based on the data mining technology," Shanghai University, 2007.

[7] Chen minjie, liu xiaobo, he shizhong. "Research on fault diagnosis and countermeasures of main bearing lubrication," Lubrication and sealing, Vol. 35, pp.113-117, May, 2010.

[8] Zhao Jiong, zheng sheng, tang qiang, hou xiaomeng, wangwei. "Research on remote online monitoring and diagnosis design of TBM," Electromechanical integration, pp.44-48, October, 2010.

# A General Regression Neural Network Model for Gearbox Fault Detection using Motor Operating Parameters

Mabrouka Baqqar[1], Tie Wang[2], Mahmud Ahmed[1], Fengshou Gu[1,] Joan Lu[1], Andrew Ball[1]
[1]School of Computing and Engineering, University of Huddersfield, Huddersfield, HD1 3DH, UK
[2]Department of Vehicle Engineering, Taiyuan University of Technology, Shanxi, 030024, P.R. China
M.Baqqar@hud.ac.uk

*Abstract*. **Condition monitoring of a gearbox is a very important activity because of the importance of gearboxes in power transmission in many industrial processes. Thus there has always been a constant pressure to improve measuring techniques and analytical tools for early detection of faults in gearboxes. This study forces on developing gearbox monitoring methods based on operating parameters which are available in machine control processes rather than using additional measurements such as vibration and acoustics used in many studies. To utilise these parameters for gearbox monitoring, this paper examines a model based approach in which a data model has been developed using a General Regression Neural Network (GRNN) to captures the nonlinear connections between the electrical current of driving motor and control parameters such as load settings and temperatures based on a two stage helical gearbox power transmission system. Using the model a direct comparison can be made between the measured and predicted values to find abnormal gearbox conditions of different gear tooth breakages based on a threshold setup in developing the model.**

***Keywords: Condition Monitoring, Gearbox, Static dataset, Fault detection, General Regression Neural Network.***

## I. INTRODUCTION

Condition monitoring (CM) is a technique for acquiring different datasets and analyzing them to assess the health and condition of equipment. Thus potential problems can be detected and diagnosed at an early stage in their development, providing the opportunity to take suitable recovery measures before they become so severe as to cause machine breakdown. To obtain accurate results CM collects large amounts of data with wide diversity including operating parameters, high density dynamic signals and special event datasets to produce historical trends which are presented to engineers and stored in databases. This gives rise to the problem that the volume of data is very large and the relationship between measurements is very complicated. Consequently, the CM data is not always understood properly [1] and the extraction of useful and meaningful information from the data is extremely challenging. In addition, because machine and sensor technologies are growing in complexity, combined with the recent progress in information technology (IT), data acquisition systems (DQS) can produce an overwhelming amount of data which is continuously increasing and contains features representing hundreds of attributes.

Among the different methods for condition monitoring of rotating machinery, artificial neural networks (ANN), in the recent decades have become an outstanding method exploiting their non-linear pattern classification properties, offering advantages for automatic detection and identification of gearbox failure conditions, whereas they do not require an in-depth knowledge of the behaviour of the system.

Vibration signals which have been widely used in the condition monitoring and fault diagnosis systems of rotating machinery [2-4] can be exploited as the detection medium in this case due to straightforwardness of menstruation and the rich contents of the signal incorporating system-critical information. However for fault detection and identification matters, the frequency ranges of the vibration signals are often wide; and according to the Shannon's sampling theorem, a high sampling rate is required, and consequently, large-sized samples are needed for the bearing fault detection purposes. Therefore due to existence of superfluous data and their large dimensionality, there is a requirement to pre-processing to extract an appropriate and economised feature vector which is essentially used to train a well-educated ANN.

In the literature, there are many signal processing tools for data analysis and diagnostic feature development. These include time domain averaging, power spectrum, cepestrum, demodulation, adaptive noise cancellation, time-series analysis, high-order statistics, time–frequency distribution, wavelet, etc., [5-7] and show good results in detecting gearbox faults. However, these techniques often need an additional vibration measurement system, which leads to high cost of the monitoring system.

This paper examines the performance of a model based condition monitoring approach by using just operating parameters for fault detection in a two stage gearbox. It has the potential to achieve cost effective monitoring system because the operating parameters are available in many systems. A model for current prediction is developed using a GRNN, which captures the complicated connections between measured variables and allows a direct comparison between the measured and predicted values to achieve gearbox fault detection.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

## II. MODEL-BASED CONDITION MONITORING

The aim of model-based fault detection and diagnosis is to create a model based on known and accepted mathematical and scientific principles verified and fine-tuned by past experience to generate accurate predictions of faults and defects likely to occur in target systems. Such models may be quantitative, qualitative or a combined system model. The model-based method is often referred to as an analytical method and has the enormous advantage that it is much less costly than constructing a real-life system for testing (possibly to destruction). Typically, the model of the target system is a continuous-variable dynamic system, with input(s) *u(k)* and output(s) *y(k)* in the presence of an unknown fault [8].

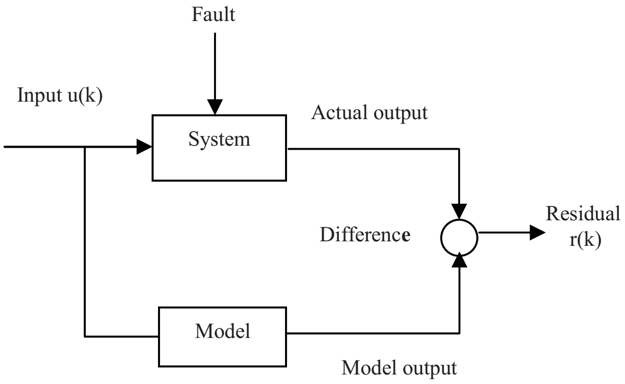


Figure 1: Model based fault detection

The model based fault detection method can easily find the fault in a system as shown in Figure 1. Residual *r(k)* in the figure is the difference between the outputs of the model and the actual system. The aim of the model is to generate a residual which can be used to indicate whether a fault is present and to identify that fault. However, the model can also be used "in reverse" information representing the behaviour of the system can be input to the model which produces an output that predicts what change in system components and/or features have taken place to produce that behaviour. The model can then predict likely causes of the change and even suggest other symptoms to search for to aid diagnosis.

A frequency division duplex (FDD) system includes three stages (procedures) with different functions: system modelling, residual generation and fault diagnosis. Firstly, a precise mathematical model is required to accurately predict system performance as model-based methods require such a model of the supervised process [11]. For most systems, such models are often very difficult to obtain. The robustness of the FDD system is often achieved by designing algorithms where the effects of model uncertainties and un-modelled dynamic disturbances on residuals are minimised and sensitivity to faults is maximised [9, 10]. Secondly, a set of residuals is generated to represent the deviation between actual and nominal features. Finally, the residuals are evaluated to relate to certain faults and to locate the fault if it is present. The model implementation and residual generation compose the model-based fault detection system.

## III. GENERAL REGRESSION NEURAL NETWORKS

Artificial intelligence and neural nets are widely used for fault detection and diagnostic. General Regression Neural Networks (GRNN) is one of the type neural networks that can be used for fault detection and diagnostic. (GRNN) works as a multi-layer feed-forward network. It is the most common network today [12]. Due to their powerful nonlinear function approximation and adaptive learning capabilities, neural networks have drawn great attention in the arena of fault diagnosis [13]. GRNN is based on localized basis function NN which uses the probability density functions. The term general regressions imply that the regression surface is not restricted to be linear. In many previous applications of the GRNN, the sigma (sigma) which is referred to as the smoothing factor in the GRNN algorithm is usually fixed and thus not applicable in a dynamic environment [14].

The main task for regression is getting relations between input variables X and output variables Y based on data including representative set of elements for analysed field. If **X** is vector containing known inputs, it is possible to define the following scalar function

$$D_i^{\,2} = (\boldsymbol{X} - X_i)^T (\boldsymbol{X} - X_i) \qquad (1)$$

This parameter provides the information about difference between two vectors. The estimate of output vector **Y** can be calculated by using this factor by:

$$\hat{Y}(\boldsymbol{X}) = \frac{\sum_{i=1}^{n} Y_i \exp\left(-\frac{D_i^{\,2}}{2\sigma^2}\right)}{\sum_{i=1}^{n} \exp\left(-\frac{D_i^{\,2}}{2\sigma^2}\right)} \qquad (2)$$

The major algorithm of the GRNN model is expressed by Equations (1) and (2). The estimate $\hat{Y}(X)$ is a weighted average of all the observed samples, $Y_i$, where each sample is weighted in an exponential manner according to the Euclidean distance, $D_i$, from each $X_i$. This appropriate weighting is explained by the inversely proportional relationship between the expression $\exp\left(-\frac{D_i^{\,2}}{2\sigma^2}\right)$ and $D_i$. That is, as $D_i$ increases, $\exp\left(-\frac{D_i^{\,2}}{2\sigma^2}\right)$ decreases and vice-versa. An optimal value for the smoothing parameter, $\sigma$- is the width of sample probability for each sample $X_i$, $Y_i$. Larger values of $\sigma$ improve smoothness of the regression surface. It must be greater than 0 and can usually range from .01 to 1 with good results [15].

## IV. GEAR FAULT SIMULATION

A tooth breakage is one of common faults in gearbox. Different levels of breakages on the pinion gear are examined in this part of research. Three levels of fault severity: 25%, 50% and 75% of a tooth are removed from three pinion gears respectively. Figure 2 illustrates the details of the faults for Gear 07 with 25% tooth breakage and Gear 08 with 50% tooth breakage. Although the defects look very large they not influence the transmission significantly because of high overlap ratio of the gear set.

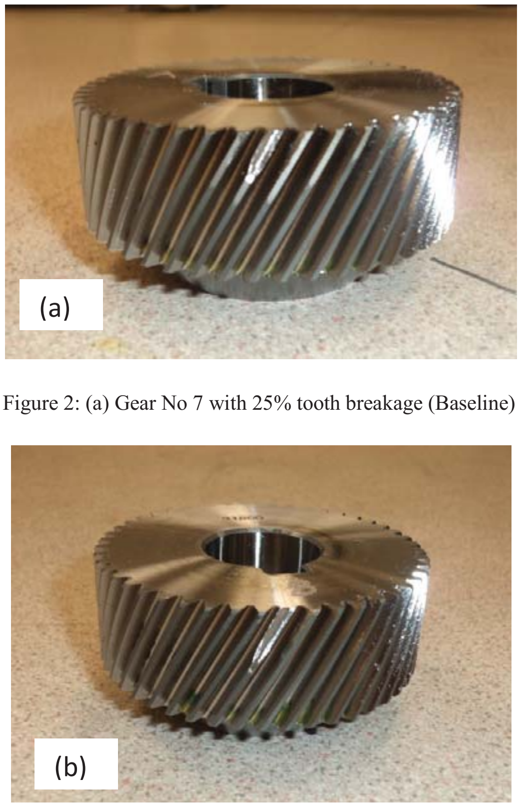


Figure 2: (a) Gear No 7 with 25% tooth breakage (Baseline)

(b) Gear No 8 with 50% tooth breakage

## V. GRNN MODEL DEVELOPMENT

GRNN was proposed by Donald Specht [14]. It uses non-iterative process and hence fast learning capability. In addition, it requires only a few training samples and very flexible to add new information with very small amount work of retraining. For these benefits, many condition monitoring applications applied GRNN to classify different fault cases. For example, GRNN is used to diagnosis different engine faults based on features extracted wavelet packet transform analyses of acoustic signals, showing GRNN is effective to classify the faults induced to the test engine[16]. In addition GRNN detectors of rotor faults of induction motor load, showing good results for rotor fault classification [17].

### *A. Data characteristics*

The data were collected for the three gear sets: Gear07, Gear08 and Gear09 using a same gearbox case. Gear07, Gear08 and Gear09 were induced with 25%, 50% and 75% tooth breakage respectively. As there was not a healthy gear for more tests, Gear07 with the smallest gear fault are taken as the baseline for model development.

To evaluate the neural network, only three variables: AC current, load set points and gearbox temperate are explored for full understanding of the principle behind. Figure 3 shows eight data sets collected from eight independent tests respectively based on Gear07. It can be seen that each data set shows a gradual increase in the current with increase in load and temperature of the gearbox. The rate of current increase with load settings is very high and in a nonlinear behavior, which indicate a complicated correlation between the current and load setting and it is not easy to model it with a simple method.

In addition, the temperature also shows considerable influences on the current. As can be seen in Figure 3, a slight inverse influence on the current can be observed. However, the decrease rate becomes smaller at higher temperature, which again indicates a more complicated model is required to describe the connections between electrical current, load settings and temperature influences.

Figure 4 shows more details of the temperature influence. It can be seen that the current decreases with the increase in temperature at each load setting. It may be due to that the damping effect of lubrication decreases with temperature. Nevertheless, the correlation also shows a nonlinear way.

As this temperature influence is very clear, it will certainly impact the model development. Fault detection must include this influence for obtaining more accurate results.

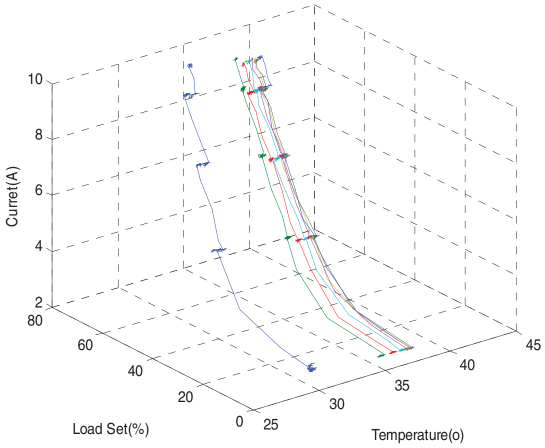


Figure 3: Data characteristics of current with temperature and load of gear in Gear07

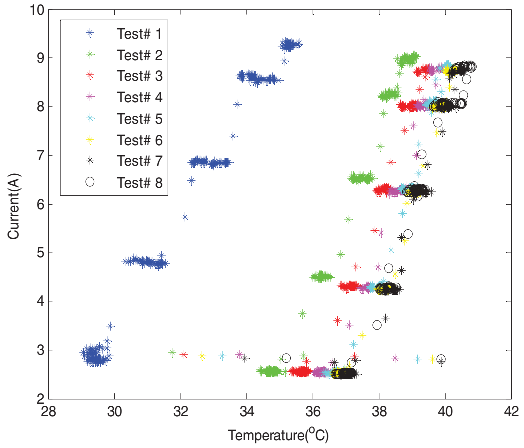


Figure 4: Data characteristics of current with temperature of gear in Gear07

### *B. Model development*

To capture the relations between the three measurements and hence to perform the model based detection discussed in Section II, a GRNN model is developed using MATLAB software based on the baseline datasets from Gear07. The model has two inputs: temperature and load set points and one output: AC current.

To train the GRNN model, the datasets from Gera07 is used as the baseline for model development. In total there are 2088 data samples from 8 tests of different runs. The 2088 data points are divided into two equal subsets of 1044 points: one for GRNN model training and the other for model verification.

After several tuning cycles, it is found that when GRNN spread parameter is 0.06, the network produces a balanced prediction in generalisation and accuracy for the first subset of data. As shown in Figure 5, the measured values are all on the model surface where the training data set is distributed. On the other hand, the model produces very small output for these which are not in the training set, which means that if there is deviation of the inputs the output will be small and the difference between measured output and predicted output will be large.

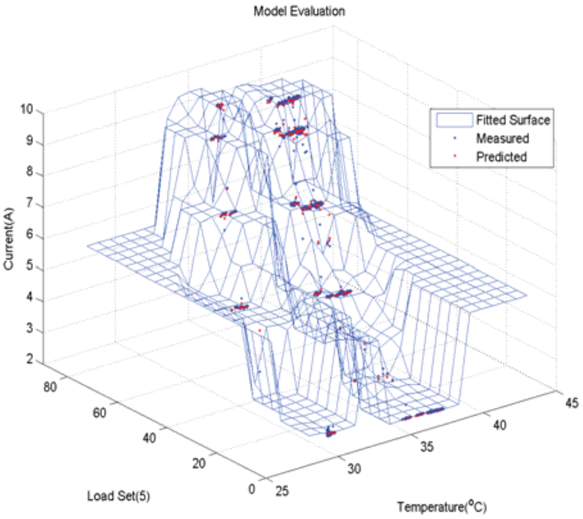


Figure 5: GRNN model inspection in the input space

### C. *Model evaluation and detection threshold*

To confirm the model performance, the 2nd dataset is employed as the input and output of the model developed from the 1st set. To measure the quality of the model in fitting to the second data and to detect abnormalities from new datasets, a threshold is developed based on the 1st dataset by comparison between the actual current and the predicted current. In particular, a threshold $D_{th}$ is defined as 3 times of the root mean squared value between the real measurement and the model prediction:

$$D_{th} = 3\sqrt{\frac{1}{N}\sum_{i=1}^{N}(I_{mi} - I_{pi})^2} \quad (1)$$

Where:
N = The number of sample.
$I_{mi}$= The actual value determined from measurements.
$I_{pi}$ = the predicted value using the GRNN.

Figure 6 shows model verification results which are calculated using the model using the 2nd part of data from Gear07. It can be seen that most of the errors are within the threshold and means that the model is fit the data very well.

On the other hand there are several data points exceeding the threshold. These data points are regarded as the outliers arisen from the load transient periods when the temperature measurements have delayed responses to current increases.

In general the model is sufficiently accurate for implementing fault detection for new data sets from other 2 gear sets.

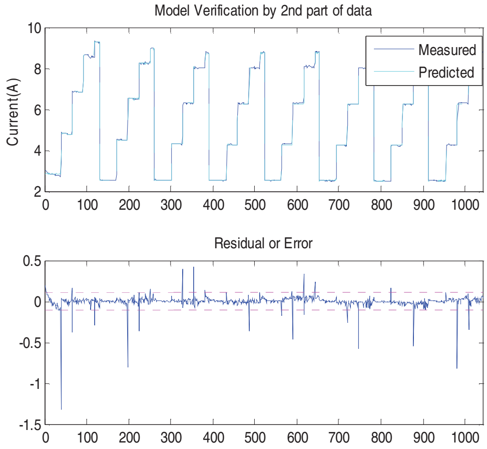


Figure 6: Model verification by 2nd part of data from Gear07

## VI. Detection results and discussions

### A. *Fault Detection on Gear08*

Figure 7(a) illustrates measured and predicted current for Gear 08 with 50% tooth breakage. It can be seen that the predicted current is very close to the measured one. However, many measurements have observed to have large difference from the predicted one.

To examine the difference only the residual data is predicted in Figure 7(b) and the details of the data points exceeding the threshold can be seen more clearly. Compared with Figure 6, many successive data points exceed the thresholds and indicate there is a fault in Gear08.

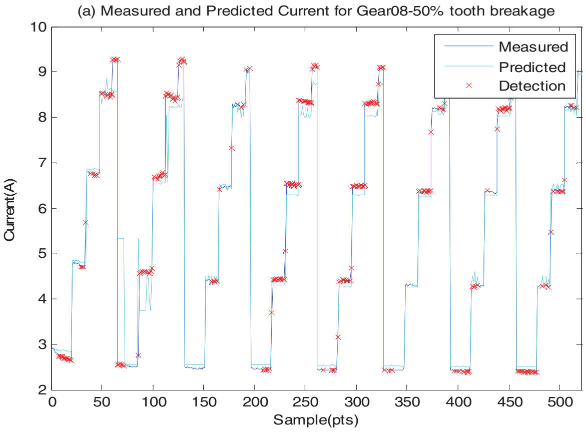
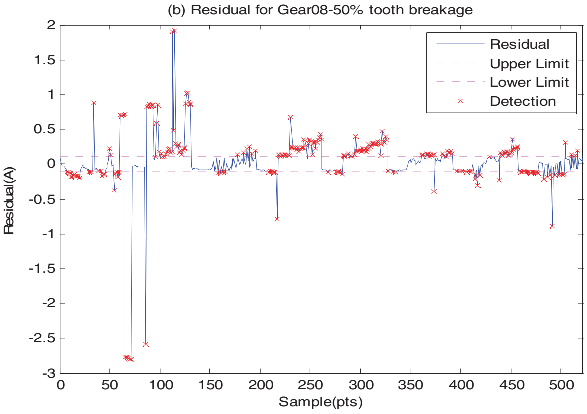


Figure 7:( a) Measured and predicted current for Gear08 under 50% tooth breakage



(b) Residual for Gear08 under 50% tooth breakage

### B. *Fault Detection on Gear09*

Figure 8(a) illustrates measured and predicted current for Gear09 on which 75% tooth breakage was induced. It can be seen clearly that the predicted currents have large difference from the measured one. To examine the difference only the residual data is predicted in Figure 8(b) and the details of the data points exceeding the threshold can be seen more clearly. Compared with Figure 6, many successive data points exceed the thresholds, which indicate that there is a fault in Gear09.

Compared with Figure 7(b), the overall amplitudes of the errors are much higher and shows that this gear have a much severer fault than Gear08.

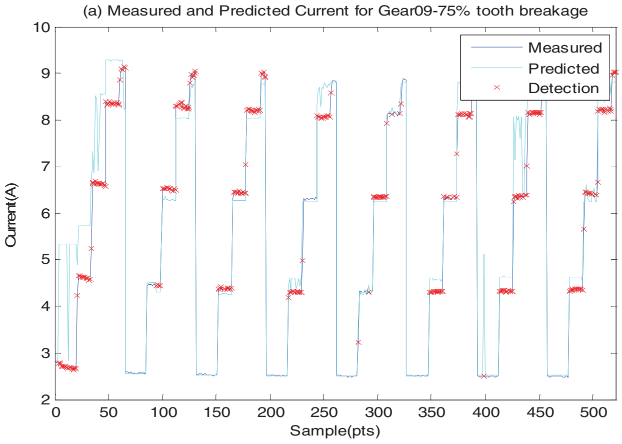
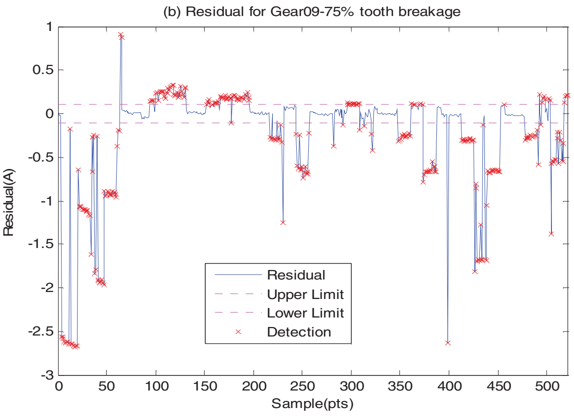


Figure 8: (a) Measured and predicted current for Gear09 under 75% tooth breakage



(b) Residual for Gear09 under75% tooth breakage

## VII. Conclusion

A GRNN model based approach is presented in this paper to detect and diagnose different faults in a gearbox using motor operating dataset. The model developed using a baseline data captures the nonlinear connections between AC current, load setting and gearbox temperature. Test results show that the GRNN model based method is accurate estimators of the complex gearbox process and allows the generation of differences from baseline and between different gear faults. Therefore, it demonstrates the effectiveness of the proposed method for detecting and diagnosing tooth faults in a two stage gearbox just using motor operating parameters.

## References

[1] S. McArthur; S. Strachan; and G. Jahn; "The design of a multi_agent system for transformer condition monitoring," IEEE Transactions on Power System, vol. 19, no. 4, pp. 1845_1852, 2004.

[2] L. Bouillaut, M. Sidahmed, "Helicopter gearbox vibrations: cyclo-stationary analysis or bilinear approach" ISSPA, Kuala Lumpur, Malaysia, 13–16 August, 2001.

[3] Q.W. Wilson, F. Ismail, M.F. Golnaraghi "Assessment of gear damage monitoring techniques using vibration measurements" Mechanical Systems and Signal Processing, 15 (5) (2001), pp. 905–922

[4] P.T. Monsen, E.S. Manolakos, M. Dzwonczyk, "Helicopter gearbox fault detection and diagnosis using analogy neural network , in: Signals, Systems and Computers", 27th Asilomar Conference, 1–3 November, 1993, vol. 1, pp. 381–385.

[5] N. Baydar, A. Ball "A comparative study of acoustic and vibration signals in detection of gear failures using Winger–Ville distribution", Mechanical Systems and Signal Processing, 15 (6) (2001), pp. 1091–1107.

[6] J. Lin, M.J. Zuo "Gearbox fault diagnosis using adaptive wavelet filter",Mechanical Systems and Signal Processing, 17 (6) (2003), pp. 1259–1269.

[7] D.M. Yang, A.F. Stronach, P. MacConnell, J. Penman "Third-order spectral techniques for the diagnosis of motor bearing condition using artificial neural networks" Mechanical Systems and Signal Processing, 16 (2–3) (2002), pp. 391–411

[8] ermann, R. & Ball´e, P. (1997). "Trends in the application of model-based fault detection and diagnosis of technical processes", Control Engineering Practice, 5(5), 707–719.

[9] Patton, R. & Chen, J. (1997)." Observer-based fault detection and isolation: robustness and applications", Control Engineering Practice, 5(5), 671–682.

[10] Gustafsson, F. (2000) "Adaptive Filtering and Change Detection", John Wiley, New York, NY, USA.

[11] Gertler, J. (1997)." Fault detection and isolation using parity relations", Control Engineering Practice, 5(5), 653–661.

[12] M. Karpenko, N. Sepehri," Neural network classifiers applied to condition monitoring of a pneumatic process valve actuator", International Journal on Engineering Applications of Artificial Intelligence 15, (2002), 273-283.

[13] P. Subbaraj,B. Kannapiran, "Artificial Neural Network Approach for Fault Detection in Pneumatic Valve in Cooler Water Spray System"International Journal of Computer Applications (0975 – 8887),Vol. 9, No.7, November 2010.

[14] Specht, D.F. 1991. "A General Regression Neural Network", IEEE Transactions on Neural Networks, 2(6): pp. 568-576.

[15] Ward Systems Group, Inc. 1996. Neuroshell/ Neurowindows Release 3.0 Manual, 3$^{rd}$ Edition. Maryland.

[16] Jian-Da Wu,Chiu-Houg Liu.2009, "An expert system for fault diagnosis in internal combustion engines using wavelet packet transform andneuralnetwork" , Expert Systems with Applications,April 2009,pp 4278-4286.

[17] Marcin Kaminski, Czeslaw T Kowalski, Teresa Orlowska-Kowalska", General Regression Neural Networks as rotor fault detectors of the induction motor", IEEE International Conference on Industrial Technology Publisher: Pages: 1239-1244, 2010.

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# Data-Driven Fault Detection of Vertical Rail Vehicle Suspension Systems

Xiukun Wei
State Key Laboratory of
Rail Traffic Control and Safety
Beijing Jiaotong University,
Beijing 100044, China.
Email: wwxxkk@gmail.com

Limin Jia
State Key Laboratory of
Rail Traffic Control and Safety
Beijing Jiaotong University,
Beijing 100044, China.
Email: jialm@vip.sina.com

Hai Liu
School of Traffic and Transportation,
Beijing Jiaotong University,
Beijing 100044, China.
Email: 10121140@bjtu.edu.cn

***Abstract*—This paper concerns data driven fault detection of vertical rail vehicle suspension systems issue. The underlying vehicle system are equipped with only accelerator sensors in the four corners of the carbody, the front and trail bogie, respectively. The faults considered are the vertical damper fault and vertical spring fault. Both PCA-based and CVA-based fault detection methods are studied in this paper. When there is a detectable fault, the detector sends an alarm signal if the residual evaluation is larger than a predefined threshold. By using the professional multi-body simulation tool, SIMPACK, the effectiveness of the proposed approach is demonstrated by simulation results for several fault scenarios.**

## I. Introduction

Suspension systems for rail vehicles are to support the carbody and bogie, to isolate the forces generated by the track unevenness at the wheels, and to control the attitude of the carbody with respect to the track surface for providing ride comfort. The railway vehicle suspension system is very important parts of railway vehicle and the reliability of the suspension system is directly related to the vehicle safety.

On line fault detection and condition monitoring for the suspension system of rail vehicles offer a number of benefits to railway systems/operations. Detection of component faults at their early stages will prevent further deterioration in vehicle performance and enhance vehicle safety. Timely repair or replacement of the faulty components will lead to increased operational reliability and availability. The need for scheduled maintenance and associated costs can be significantly reduced, because maintenance in the future may be carried out on demand. So it is necessary to timely detect the fault of vehicles suspension after it occurs. Some studies on the condition monitoring of railway vehicle and suspension systems are reported in [1],[2],[3], [6], [9], [8]and the references therein.

The fault detection issue of the rail vehicle suspension systems are paid some attentions in the recent years. In [10][11], the authors derived a fault detection approach for the rail vehicle suspension systems based on Kalman filter. The fault isolation is handled by using the similarity measurements. The interaction multi-model (IMM) approach and parameter estimation reported in [4],[6] are convinced to be appropriate alternatives. However, when IMM method is applied, the models needed is increasing very dramatically when more components in the primary suspension and secondary suspension are considered. Parameter estimation method [6] applies the particles filter for the parameter estimation. The parameter changes are identified online to indicate the health condition of the components. Nevertheless, the computation burden of this approach cannot afforded by the current available monitoring unit. It is true that there is great potential for the improvement in the performance of condition monitoring if the a priori knowledge or information captured by the models is fully used. However, in many cases, the parameters of the vehicle suspension system are not available. In the mean time, due to nonlinearities of the components and the complexities of the suspension systems, a precise model cannot be obtained. Due to these limitations of the model based fault detection approaches, there is increasing interest in using the multivariate statistical approaches to monitoring system health conditions[7][12][5]. As the knowledge of the authors, there is no any report which solves the fault detection problem based on data driven methods.

In this paper, we consider data driven fault detection of rail vehicle systems problem. The underlying vehicle are only equipped with accelerator sensors in the corners of the carbody, the front and trail bogie, which provide the measured signals for the condition monitoring for the vehicle suspension system. The Dynamical Principle Component Analysis (DPCA) and Canonical Variate Analysis (CVA) based approaches are applied for the fault detection problem, respectively. The considered faults are the vertical damper faults and the vertical spring faults of both primary and secondary suspensions. A subway vehicle model is used for the Matlab-Simpack hybrid simulation study in this paper.

This paper is organized in the following. The railway vehicle system, its dynamics and the sensor configuration for the data collection are introduced in Section 2. In the third section, the DPCA and CVA fault detection methods are briefly reviewed. The SIMPACK-Matlab co-simulation results are provided in Section 4. Finally, conclusions are given in Section 5.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

## II. THE VERTICAL RAIL VEHICLE SUSPENSION SYSTEMS

### A. The Vehicle Suspension System

A subway vehicle model is used for the study in this paper. To describe the dynamic behaviour more accuracy, three degree-of-freedom (bounce, pitch and roll) is considered for both carbody and bogies. The vertical suspension system of this subway vehicle is depicted in Fig. 1. The equations describing the dynamic behaviour of the railway vehicle are developed from the application of Newton's laws of motion to the individual masses. The dynamical equations of the suspension system for a vehicle moving on a straight track are derived in the following.

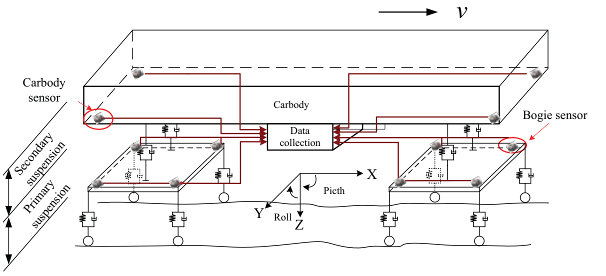


Fig. 1. The vertical suspension system of the rail vehicle

For the carbody, the three DOF equations are described as:

$$M\ddot{z} + 4C_{2z}\dot{z} - 2C_{2z}\dot{z}_{FB} - 2C_{2z}\dot{z}_{RB} + 4K_{2z}z$$
$$-2K_{2z}z_{FB} - 2K_{2z}z_{RB} = 0 \tag{1}$$

$$J_{\phi}\ddot{\phi} + 4C_{2z}l^2\dot{\phi} - 2C_{2z}l\dot{z}_{FB} + 2C_{2z}l\dot{z}_{RB} +$$
$$4K_{2z}l^2\phi - 2K_{2z}lz_{FB} + 2K_{2z}lz_{RB} = 0 \tag{2}$$

$$J_{\theta}\ddot{\theta} + 4C_{2z}b^2\dot{\theta} - 2C_{2z}b^2\dot{\theta}_{FB} - 2C_{2z}b^2\theta_{RB}$$
$$+(4K_{2z}b^2 + 2K_{\theta})\theta - (2K_{2z}b^2 + K_{\theta})\theta_{FB}$$
$$-(2K_{2z}b^2 + K_{\theta})\theta_{RB} = 0 \tag{3}$$

Where $z, z_{FB}$ and $z_{RB}$ are the vertical displacement of the carbody, the leading bogie and the trailing bogie, respectively. $\phi$ represents the pitch angle of the center of gravity(c.g.) while $\theta$ is the roll angle of the c.g. for the masses. Their subscripts have the same meaning with the subscripts of $z$. Please refer to Table I for the parameters.

For the leading bogie, the three DOF equations are described as:

$$M_B\ddot{z}_{FB} - 2C_{2z}\dot{z} - 2C_{2z}l\dot{\phi} + (4C_{1z} + 2C_{2z})\dot{z}_{FB}$$
$$-C_{1z}\dot{z}_{W1R} - C_{1z}\dot{z}_{W1L} - C_{1z}\dot{z}_{W2R} - C_{1z}\dot{z}_{W2L} -$$
$$2K_{2z}z - 2K_{2z}l\phi + (4K_{1z} + 2K_{2z})z_{FB} - K_{1z}z_{W1R}$$
$$-K_{1z}z_{W1L} - K_{1z}z_{W2R} - K_{1z}z_{W2L} = 0 \tag{4}$$

TABLE I
VEHICLE PARAMETERS

| | Description | Unit |
|---|---|---|
| $M$ | Carbody mass | $kg$ |
| $M_B$ | Bogie mass | $kg$ |
| $J_{\phi}$ | Carbody pitch inertia | $kgm^2$ |
| $J_{\theta}$ | Carbody roll inertia | $kgm^2$ |
| $J_{B\phi}$ | Bogie pitch inertia | $kgm^2$ |
| $J_{B\theta}$ | Bogie roll inertia | $kgm^2$ |
| $l_b$ | Half of the distance between two wheelsets in a bogie | $m$ |
| $w_b$ | Half of the distance between two air spring in lateral | $m$ |
| $l_c$ | Half of the carbody length | $m$ |
| $w_c$ | Half of the carbody width | $m$ |
| $K_{2z}$ | Spring constants of air spring | $kN/m$ |
| $K_{1z}$ | Spring constants of primary spring | $kN/m$ |
| $C_{2z}$ | Damping constants of secondary damper | $kNs/m$ |
| $C_{1z}$ | Damping constants of primary damper | $kNs/m$ |
| $K_{\theta}$ | Spring constants of the anti-roll spring | $kN/m$ |

$$J_{B\phi}\ddot{\phi}_{FB} + 4C_{1z}l_1^2\dot{\phi}_{FB} - C_{1z}l_1\dot{z}_{W1R} - C_{1z}l_1\dot{z}_{W1L} +$$
$$C_{1z}l_1\dot{z}_{W2R} + C_{1z}l_1\dot{z}_{W2L} + 4K_{1z}l_1^2\phi_{FB} - K_{1z}l_1z_{W1R}$$
$$-K_{1z}l_1z_{W1L} + K_{1z}l_1z_{W2R} + K_{1z}l_1z_{W2L} = 0 \tag{5}$$

$$J_{B\theta}\ddot{\theta}_{FB} - 2C_{2z}b^2\dot{\theta} + (2C_{2z}b^2 + 4C_{1z}b_1^2)\dot{\theta}_{FB} +$$
$$C_{1z}b_1\dot{z}_{W1R} - C_{1z}b_1\dot{z}_{W1L} + C_{1z}b_1\dot{z}_{W2R} - C_{1z}b_1\dot{z}_{W2L}$$
$$-(2K_{2z}b^2 + K_{\theta})\theta + (2K_{2z}b^2 + 4K_{1z}b_1^2 + K_{\theta})\theta_{FB} +$$
$$K_{1z}b_1z_{W1R} - K_{1z}b_1z_{W1L} + K_{1z}b_1z_{W2R} - K_{1z}b_1z_{W2L}$$
$$= 0 \tag{6}$$

Where $z_{W1R}$ represents the vertical displacement of the right wheel in the leading wheelset under the leading bogie while $z_{W2L}$ means the vertical displacement of the left wheel in the trailing wheelset under the leading bogie. The meanings of other symbols can be easy understood in this logic. In consideration of the fact that the wheelsets are rolling against the track, the vertical displacement of the wheel can be seen as the the unevenness of the track. In a similar way, the model of the trailing bogie can be derived.

The state-space form of the vehicle dynamical model can be derived as:

$$\dot{x} = Ax + B_d d \tag{7}$$
$$y = Cx + D_d d \tag{8}$$

where

$$x = \begin{bmatrix} \dot{z} & \dot{\phi} & \dot{\theta} & z & \phi & \theta & \dot{z}_{FB} & \dot{\phi}_{FB} & \dot{\theta}_{FB} & z_{FB} \\ \phi_{FB} & \theta_{FB} & \dot{z}_{RB} & \dot{\phi}_{RB} & \dot{\theta}_{RB} & z_{RB} & \phi_{RB} \\ \theta_{RB} \end{bmatrix}^T$$

$$d = \begin{bmatrix} \dot{z}_{W1R} & \dot{z}_{W1L} & \dot{z}_{W2R} & \dot{z}_{W2L} & z_{W1R} & z_{W1L} \\ z_{W2R} & z_{W2L} & \dot{z}_{W3R} & \dot{z}_{W3L} & \dot{z}_{W4R} \\ \dot{z}_{W4L} & z_{W3R} & z_{W3L} & z_{W4R} & z_{W4L} \end{bmatrix}^T$$

$$y = \begin{bmatrix} z & \phi & \theta & z_{FB} & \phi_{FB} & \theta_{FB} & z_{RB} & \phi_{RB} \\ \theta_{RB} \end{bmatrix}^T$$

Matrixes $A, B_d, C$ and $D_d$ can be derived from the above differential equations, $d$ is the vertical track velocity and displacement due to track vertical irregularities.

### *B. The Model Based on Acceleration Measurements*

The developed model (7) and (8) represent the dynamical property of the vehicle suspension system. In this model, the vertical displacement, pitch angle displacement and roll angle displacement of the cardody and the bogie are needed to be measured for the fault detection purpose. However, displacement sensor and angle sensors have the problems such as reliability, maintenance and installation. Compare to these sensors, accelerator sensors have the advantages of cheap and reliable. It does not need to be maintained for a long time period. Due to all the reasons stated above, only accelerator sensors are adopted for obtaining the measurement signals. The sensor configuration is designed as shown in Fig. 1. In the following, the new output equation will be derived.

Carbody sensors are equipped in the four corners on the floorboard, and the bogie sensors are equipped in the four corners on the upside of the bogie. By applying double integral to the acceleration signals, displacement signals are acquired. That is

$$z = \int\int a dt dt \tag{9}$$

where $a$ is the acceleration value and $z$ is the displacement. However, these displacement signals obtained is different from the outputs in our model, for example, the angle displacements. So, the model outputs needed to be modified.

As depicted in Fig. 2, we set four index points on the rectangle to discuss the relationship between the vertical displacements at the four corners with the angle displacement. These four index points are the center of the rectangle, the middle pint of the front edge, the middle pint of the right edge and the front corner on the right. The vertical displacement of the four index points are $z, z_F, z_R$ and $z_{FR}$, respectively. Then one obtains

$$z + z_{FR} = z_F + z_R \tag{10}$$

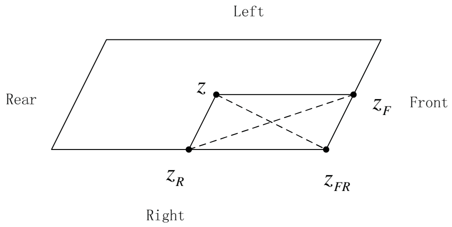


Fig. 2. Relationship among the displacement of the four index points

Consider the pitch motion, we have the equation

$$z_F = z + l\sin(\phi) \approx z + l\phi \tag{11}$$

for a small pitch angle.

Consider the roll motion, we have a similar equation:

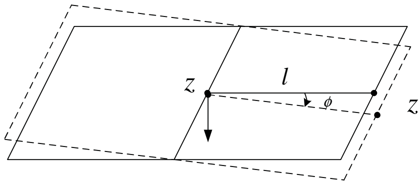


Fig. 3. The pitch motion

$$z_R \approx z - w\theta \tag{12}$$

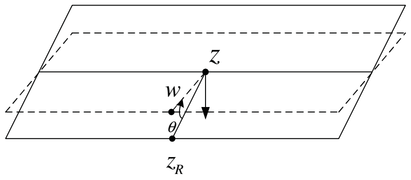


Fig. 4. The roll motion

Combing with (10),(11),(12) together, we have the following transformation relationship:

$$z_{FR} \approx z + l\phi - w\theta \tag{13}$$

In the same procedure, one obtains:

$$z_{FL} \approx z + l\phi + w\theta \tag{14}$$

$$z_{RR} \approx z - l\phi - w\theta \tag{15}$$

$$z_{RL} \approx z - l\phi + w\theta \tag{16}$$

Then the new state space is transformed into

$$\dot{x} = Ax + B_d d \tag{17}$$

$$y_{cor} = TCx + TD_d d \tag{18}$$

where the transformation matrix

$$T = \begin{bmatrix} 1 & l_c & -w_c & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & l_c & w_c & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -l_c & -w_c & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -l_c & w_c & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & l_b & -w_b & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & l_b & w_b & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -l_b & -w_b & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -l_b & w_b & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & l_b & -w_b \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & l_b & w_b \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -l_b & -w_b \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -l_b & w_b \end{bmatrix}$$

where

$$y_{cor} = \begin{bmatrix} z_{FR} & z_{FL} & z_{RR} & z_{RL} & z_{FB\_FR} z_{FB\_FL} \\ z_{FB\_RR} & z_{FB\_RL} & z_{RB\_FR} & z_{RB\_FL} \\ z_{RB\_RR} & z_{RB\_RL} \end{bmatrix}^T$$

$z_{FR}$ and $z_{FL}$ represent the vertical displacement of the front corners on the right and left, respectively. $z_{FB_{FR}}$ represent the vertical displacement of the front corner on the right of the leading bogie.

*Remark 2.1:* Data is the very critical for the data driven fault detection methods. The data should contain rich information of the system dynamics and the fault information when a fault occurs in the system. In this paper, only acceleration sensors are used for the vehicle suspension fault detection systems. From the above observation, the dynamics of rail vehicle suspension systems are contained in the displacements of the the carbody, the front and rail bogies, which can be obtained by the acceleration sensors. The sensor configuration presented before can provided enough information for the fault detection.

## III. DPCA AND CVA FAULT DETECTION PRINCIPLE

Thanks to its simplicity and efficiency in processing huge amount of process data, PCA is recognized as a powerful tool of statistical process monitoring and widely used in the process industry for purpose of fault detection and diagnosis[12][7]. In the meantime, it is noticed that the CVA based condition monitoring methods are outperformed PCA based approaches [5]. In this paper, both methods are studied for the considered vehicle suspension fault detection systems. The main steps for these two methods are briefly reviewed in this section.

### *A. A Brief Description of PCA*

The standard PCA-based fault detection consists of three steps and can be formulated as follows:

- Data collection and pre-processing: Consider a data matrix $X \in \mathbf{R}^{N \times m}$ consisting of $N$ samples and $m$ sensors collected from process. Matrix $X$ is then scaled to zero mean, and often to unit variance. Let the scaled data be

$$\mathbf{Y} = \left[ \begin{array}{c} y_1^T \\ \vdots \\ y_N^T \end{array} \right] \in \mathbf{R}^{N \times m} \tag{19}$$

with $y_i \in R^m$, $i = 1, \cdots, N$ , denoting the ith scaled vector.

- Decomposition of covariance matrix: The covariance matrix is formed as

$$\sum\nolimits_0 \approx \frac{1}{N}\mathbf{Y}^T\mathbf{Y}.$$

By means of SVD or EVD (eigen value decomposition), the covariance matrix is decomposed as follows:

$$\frac{1}{N-1}\mathbf{Y}^T\mathbf{Y} = \mathbf{P}\Lambda\mathbf{P}^T, \Lambda = \left[ \begin{array}{cc} \Lambda_{pc} & 0 \\ 0 & \Lambda_{res} \end{array} \right] \tag{20}$$

with $\sigma_i^2$ , $i = 1, \cdots, m$, is the $i^{th}$ singular value of the covariance matrix and

$$\begin{array}{l} \Lambda_{pc} = diag(\sigma_1^2, \cdots, \sigma_l^2), \Lambda_{res} = diag(\sigma_{l+1}^2, \cdots, \sigma_m^2) \\ \sigma_1^2 \geq \cdots \geq \sigma_l^2 \gg \sigma_{l+1}^2 \geq \cdots \geq \sigma_m^2, \mathbf{P}\mathbf{P}^T = \mathbf{I}_{m \times m} \\ \left[ \begin{array}{c} \mathbf{P}_{pc}^T \\ \mathbf{P}_{res}^T \end{array} \right] \left[ \begin{array}{cc} \mathbf{P}_{pc} & \mathbf{P}_{res} \end{array} \right] = \left[ \begin{array}{cc} \mathbf{I}_{l \times l} & 0 \\ 0 & \mathbf{I}_{(m-l)\times(m-l)} \end{array} \right] \end{array}$$

- On-line fault detection: When a new scaled measurement $x \in \mathbf{R}^m$ is available, the SPE and Hotelling's $T^2$ indices can be computed as

$$\mathbf{T}^2 = y^T\mathbf{P}_{pc}\Lambda_{pc}^{-1}\mathbf{P}_{pc}^T y \tag{21}$$

$$SPE = y^T\mathbf{P}_{res}\mathbf{P}_{res}^T y \tag{22}$$

The fault detection logic is $SPE \leq J_{th,SPE}$ and $T^2 \leq J_{th,T^2} \Rightarrow$ fault-free. Otherwise $\Rightarrow$ faulty.

where $J_{th,T^2}$ and $J_{th,SPE}$ are the thresholds for SPE and $T^2$ respectively.

The PCA methods can be extended to take into account the serial correlations, by augmenting each observation vector with the previous $l$ observations and stacking the data matrix in the following manner,

$$\mathbf{Y}(l) = \left[ \begin{array}{cccc} y_t^T & y_{t-1}^T & \cdots & y_{t-l}^T \\ y_{t-1}^T & y_{t-2}^T & \cdots & y_{t-l-1}^T \\ \cdots & \cdots & \ddots & \vdots \\ y_{t+l-n}^T & y_{t+l-n-1}^T & \cdots & y_{t-n}^T \end{array} \right] \tag{23}$$

where $y_t^T$ is the $m$-dimensional observation vector in the training set at time instance $t$. By performing PCA on the data matrix in Eq. 23, a multivariate autoregressive (AR) (ARX model if the process inputs are included) is extracted directly from the data. This approach of applying PCA to Eq. 23 is referred to here as dynamic PCA (DPCA).

### *B. CVA for Dynamic Processes*

The stacked past and future vectors, **p** and **f**, are represented as follows:

$$\begin{array}{rcl} \mathbf{p}_{t-1}^{t-l} & = & [\mathbf{y}_{t-1}^T \cdots \mathbf{y}_{t-l}^T]^T \qquad (24) \\ \mathbf{f}_t^{t+l+1} & = & [\mathbf{y}_t^T \cdots \mathbf{y}_{t+l-1}^T]^T \qquad (25) \end{array}$$

where **y** denotes output vectors, subscript t is the present time index for **y**, and $l$ is the number of the lag or the lead. In CVA, the stacked future and past vectors are normalized using

$$\mathbf{d}_{f,t} = \sum\nolimits_{ff}^{-1/2} \mathbf{f}_t^{t+l+1}, \mathbf{d}_{p,t-1} = \sum\nolimits_{pp}^{-1/2} \mathbf{p}_{t-1}^{t-l} \tag{26}$$

where $\sum_{ff} = E(\mathbf{f}_t^{t+l+1}\mathbf{f}_t^{t+l+1^T})$ and $\sum_{pp} = E(\mathbf{p}_{t-1}^{t-l}\mathbf{p}_{t-1}^{t-l^T})$ ( $E(\cdot)$ denotes the expectation operator). The normalized vectors $\mathbf{d}_{f,t}$ and $\mathbf{d}_{p,t-1}$ are defined as the scaled stacked future and past vectors at time $t$, respectively. The conditional expectation of the scaled future vector takes the form:

$$\hat{E}(\mathbf{d}_{f,t}|\mathbf{d}_{p,t-1}) = \sum\nolimits_{ff}^{-1/2}\sum\nolimits_{fp}\sum\nolimits_{pp}^{-1/2}\mathbf{d}_{p,t-1} \tag{27}$$

where $\hat{E}(\alpha|\beta)$ denotes the expectation of $\alpha$ under condition $\beta$. The physical mean of $\sum_{fp}$, defined as $E(\mathbf{f}_t^{t+l+1}\mathbf{p}_{t-1}^{t-l^T})$, is the well-known Hankel matrix. Thus, $\sum_{ff}^{-1/2}\sum_{fp}\sum_{pp}^{-1/2}$ indicates the scaled Hankel matrix. The scaled Hankel matrix can be factorized using singular value decomposition (SVD),

$$\begin{array}{rcl} \widehat{\mathbf{d}}_{f,t} & = & \sum\nolimits_{ff}^{-1/2}\sum\nolimits_{fp}\sum\nolimits_{pp}^{-1/2}\mathbf{d}_{p,t-1} \\ & = & \mathbf{USV}^T\mathbf{d}_{p,t-1} \approx \mathbf{U}_k\mathbf{S}_k\mathbf{V}_k^T\mathbf{d}_{p,t-1} \qquad (28) \end{array}$$

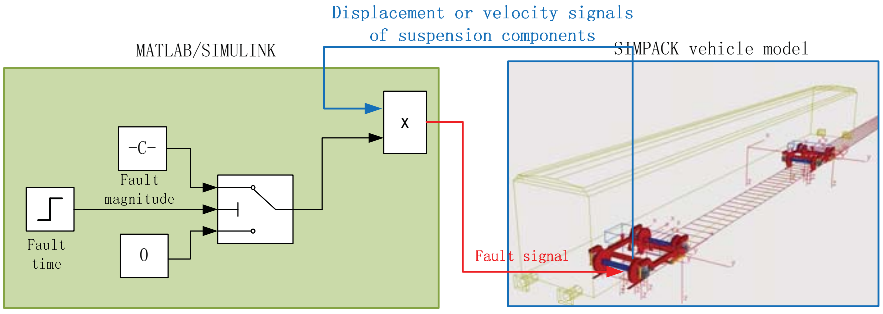


Fig. 5. Co-simulation between SIMPACK and MATLAB/SIMULINK

where $\widehat{\mathbf{d}}_{f,t}$ denotes $\hat{E}(\mathbf{d}_{f,t}|\mathbf{d}_{p,t-1})$ and $k$ represents the state order. $\mathbf{U}_k$ and $\mathbf{V}_k$ consist of the first $k$ column vectors of $U$ and $V$, respectively, and the diagonal matrix $\mathbf{S}_k$ is the $k \times k$ principal submatrix of $\mathbf{S}$. Then, the past and future CVs at time $t$ are given by

$$\begin{aligned} \mathbf{z}_t &= \mathbf{U}_k^T \mathbf{d}_{f,t} = \mathbf{U}_k^T \sum_{ff}^{-1/2} \mathbf{f}_t^{t+l+1} = \mathbf{L}_k \mathbf{f}_t^{t+l+1} \\ \mathbf{m}_t &= \mathbf{V}_k^T \mathbf{d}_{p,t-1} = \mathbf{V}_k^T \sum_{pp}^{-1/2} \mathbf{p}_{t-1}^{t-l} = \mathbf{J}_k \mathbf{p}_{t-1}^{t-l} \end{aligned} \tag{29}$$

respectively, and these CVs satisfy

$$\mathbf{z}_k = \mathbf{S}_k \mathbf{m}_t \tag{30}$$

Similar to the $T^2$ and SPE indices, the following indices are used in this paper:

$$T_s^2 = \mathbf{y}^T \mathbf{J}_k \mathbf{J}_k^T \mathbf{y} \tag{31}$$

$$SPE_s = r_t^T r_t \tag{32}$$

where $r_t = (I - \mathbf{J}_k \mathbf{J}_k)\mathbf{y}$

## IV. SIMULATION

### A. Co-simulation to Generate the Suspension Faults

In this section, SIMPACK vehicle model is used to generate the acceleration signals and different faults in both primary and secondary suspension are simulated. Here we discuss how to simulate the suspension faults in SIMPACK. The interface between MATLAB/SIMULINK and SIMPACK make it possible to simulate suspension fault with different fault magnitudes at any time during the operation.

As we all known, the damper will generate a force, whose value equals the damper coefficient times piston's velocity, to reject the moving of the piston. And we generate the damper fault signal in the way it works. As shown in Fig. 5, for example, sensors are equipped at the position of a secondary damper to measure it's moving velocity. Suggest that the damper lose half of its value at the 15th second. Then a external force is exerted on the piston to reduce the performance of the damper as shown in Fig. 5. The direction of the external force is opposite to the force generated by the damper, and the value of force equals the fault magnitude times piston's velocity. In a similar way, spring fault can be simulated.

TABLE II
FAULT SCENARIOS

| scenario | fault | fault pattern | faulty time (s) |
|---|---|---|---|
| fault 0 | no fault | | |
| fault 1 | $C_2$ | failure | 15 |
| fault 2 | $K_1$ | lost 90% | 20 |
| fault 3 | $K_2$ | lost 25% | 30 |
| fault 4 | $K_2$ | lost 90% | 30 |

### B. Fault Detection Results

The simulation results for the railway vehicle suspension fault detection are shown in this section. The track irregularities used in the simulations are the German 6th grade track spectrum. The faults considered in the simulation study are listed in Table II.

The no fault data are used for training the models. For the DPCA approach, the time lags is selected as 20. 120 principle components are selected to be retained in the model. For the CVA approach, the measured data is compressed by PCA method and only 6 variables are kept in the new data. This avoids the problem that $\Sigma_{ff}^{-1/2}$ and $\Sigma_{pp}^{-1/2}$ are not real. The time lags and leads in the CVA approach are selected as 20. In the CVA model, 18 states are selected.

The fault detection results are shown in Fig. 6-9. In all the figures, the red solid line indicates the no fault case. The blue dashed line represents the case of $C_2$ failure. The $K_1$ lost 90% of its value is indicated by the green dash-dotted line. For the fourth scenario, $K_1$ lost 25% of its value, is represented by the black dotted line. Finally, the $K_1$ lost 90% of its value fault is represented by the cyan dots.

Fig. 6 shows the fault detection results of PCA approach using the $T^2$ index. Fault 1, 2 and 4 are detected successfully. The $T^2$ index of the weak fault 3 is not obvious. The SPE index of the PCA approach is shown in Fig. 7. The performance is similar to that of the $T^2$ index. Fig. 8 shows the results of the CVA approach by using $T_s^2$ index. It can be seen that the CVA approach is very sensitive to fault 2. The fault 1 is also detected successfully. The detection results of the fault 3 and 4 are not very satisfied. Similar performance is observed by using the $SPE_s$ index.

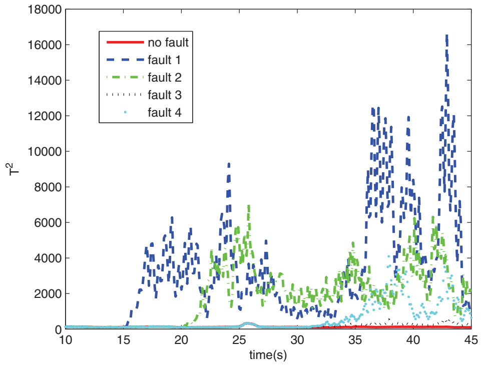


Fig. 6. Fault detection results based on PCA and $T^2$ index

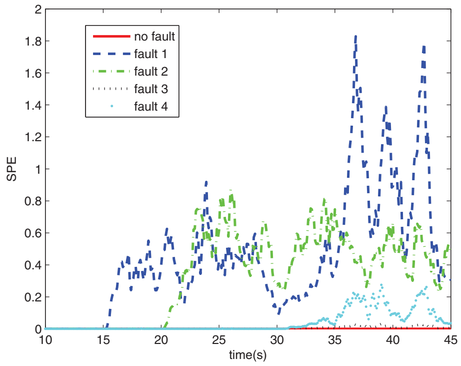


Fig. 7. Fault detection results based on PCA and $SPE$ index

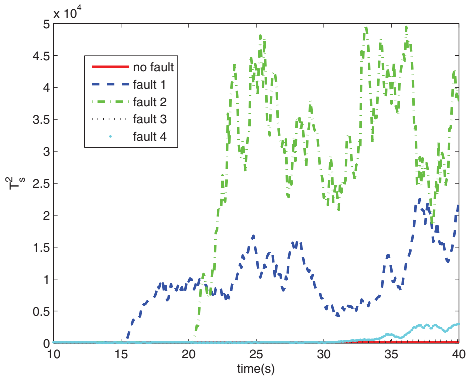


Fig. 8. Fault detection results based on CVA and $T_s^2$ index

## V. Conclusion

In this paper, the data driven fault detection problem for the railway vehicle suspension systems are studied. A new fault detection system is proposed where only the acceleration signals are used for the detection purpose. Both PCA and CVA based approaches are studied. For large magnitude faults, both approaches achieves very good detection performance. However, for weak faults, for instance, the $K_2$ lost of $25\%$ of its value, both methods do not detect the faults obviously. This means that both methods cannot detect early stage fault. New strategy for improving the sensitivity of the data driven fault detection methods are necessary. The final comment is that the components nonlinearities are not considered in the simulation study. They should be investigated in the future work.

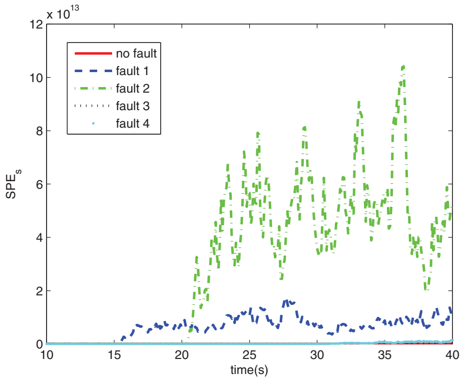


Fig. 9. Fault detection results based on PCA and $SPE_s$ index

## Acknowledgment

This work is partly supported by Chinese 863 program (Contract No. 2011AA110503-6), State Key Laboratory of Rail Traffic Control and Safety (Contract No.RCS2011ZZ005) and Ph.D. Programs Foundation of Ministry of Education of China (grant number: 20110009120037).

## References

[1] S. Bruni, R. Goodall, T. X. Mei, and H. Tsunashima, "Control and monitoring for railway vehicle dynamics," *Vehicle System Dynamics*, vol. 45, no. 7-8, pp. 765–771, 2007.

[2] R. Goodall and T. Mei, "Advanced control and monitoring for railway vehicle suspensions," in *International Symposium on Speed-up and Service Technology for Railway and Maglev Systems(STECH'06)*, Chengdu,China, 1 2006, pp. 10–16.

[3] R. Goodall1 and C. Roberts, "Concepts and techniques for railway condition monitoring," in *IET International Conference on Railway Condition Monitoring*, 2006, pp. 90–95.

[4] Y. HAYASHI, H. TSUNASHIMA, and Y. MARUMO, "Fault dectecion of railway vehicle suspension systems using multiple-model approach," *Mechanical System for Transportation and Logistics*, vol. 1, no. 1, pp. 88–98, September 2008.

[5] C. Lee, S. W. Choi, and L. In-Beum, "Variable reconstruction and sensor fault identification using canonical variate analysis," *Journal of Process Control*, vol. 16, pp. 747–761, 2006.

[6] P. Li, R. Goodall, P. Weston, C. S. Ling, C. Goodman, and C. Roberts, "Estimation of railway vehicle suspension parameters for condition monitoring," in *Control Engineering Practice*, 2006, pp. 43–55.

[7] S. J. Qin, "Data-driven fault detection and diagnosis for complex industrial processes," in *The 7th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes*, Barcelona,Spain, 2009, pp. 1115–1125.

[8] X. Wei, L. Jia, and H. Liu, "Fault diagnosis filter design for railway vehicle suspension systems based on lmi optimization," *An international Interdisciplinary Journal*, vol. accepted, 2012.

[9] X. Wei, S. Lin, and H. liu, "Distributed fault detection observer for rail vehicle suspension systems," in *Chinese Control and Decision Conference*, accepted, 2012.

[10] X. Wei, H. liu, and Y. Qin, "Fault diagnosis of rail vehicle suspension systems by using GLRT." Chinese Control and Decision Conference, 2011.

[11] ——, "Fault isolation of rail vehicle suspension systems by using similarity measure," in *International Conference on Intelligent Railway Tranportation*, Bei Jing, China, 2011, pp. 391–396.

[12] S. Yin, S. Ding, A. Naik, P. Deng, and A. Haghani, "On pca-based fault diagnosis techniques," in *2010 Conference on control and Fault Tolerant Systems*, Nice,France, 2010, pp. 179–184.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Robust adaptive fault estimation for a commercial aircraft oscillatory fault scenario

Xiaoyu Sun, Ron J Patton
Department of Engineering, University of Hull
Hull, UK
X.Sun@2009.hull.ac.uk; r.j.patton@hull.ac.uk

Philippe Goupil
Flight Control System Department, Airbus
Toulouse, France
philippe.goupil@airbus.com

*Abstract*— **A linear time invariant model-based robust fast adaptive fault estimator with unknown input decoupling is proposed to estimate aircraft elevator oscillatory faults. Since the robust fast adaptive fault estimator depends on system output error dynamics which are de-coupled from the unknown inputs (modeling uncertainty), the fault estimation signal generated by the designed fault estimator is robust to the estimated unknown inputs. To obtain a fast fault estimation speed, an adaptive fault estimator involves both proportional and integral components. A Lyapunov stability analysis of the robust fast adaptive fault estimator is given and the fault estimator dynamic response is achieved by pole assignment in subregions realized by LMIs. The proposed robust fast adaptive fault estimator is implemented on a high-fidelity nonlinear aircraft model to detect and estimate elevator actuator oscillatory faults.**

***Keywords- adaptive fault estimator; unknown input ; fault estiamation ; linear matrix inequalities; Oscillatory Fault Case.***

## I. INTRODUCTION

The traditional approach to detecting and isolating faults in a flight control system makes use of hardware redundancy by a replication of hardware [1] (sensors, actuators or even flight control comp3ters). However, there is a growing interest in methods which do not require additional hardware redundancy, for easing the development of the future more sustainable aircraft (Cleaner, Quieter, Smarter and More Affordable). Highlighting the link between aircraft sustainability and fault detection, it can be demonstrated that improving the diagnosis performance in flight control systems allows the designers to optimize the aircraft structural design (resulting in weight saving), which in turn helps improve aircraft performance and to decrease its environmental footprint. Concretely, if the minimum detectable fault amplitude and/or the detection time can be decreased, the aircraft structural design will be improved and the aircraft will be made lighter [2].

As an alternative to hardware redundancy the model-based approach, often referred to as Fault Detection and Diagnosis (FDD) or Fault Detection and Isolation (FDI) makes use of analytical redundancy by generating redundant estimates of measured signals [3]. Although fault information generation via model-based FDD method for actuators (or sensors) generally increases the flight control system computational load, they can increase aircraft sustainability by improving fault diagnosis performance which leads to the possibility of optimizing aircraft structural design. All of which can help to achieve the challenges related to the "greening" of the aircraft. Model-based FDD has often been considered for fault detection, fault location and even diagnosis of fault severity in aircraft flight control systems [4, 5].

Many approaches to robust model-based FDD have been proposed in the past decades [4-8]. The major challenge is that the fault information signal should be robust to unknown inputs (UIs), used to represent a structured form of modelling uncertainty. To achieve the FDD robustness, different methods have been studied, e.g. the use of optimization methods [4], the unknown input observer (UIO) [6], the sliding mode observer [7] and geometric design approaches [8].

The method proposed in this paper is applied to a non-linear simulation of a generic aircraft provided by AIRBUS for a benchmark study within the ADDSAFE FP7 project [9, 10]. The benchmark is considered highly representative of the flight physics and aircraft handling qualities. One of the often considered fault scenarios is the oscillatory fault case (OFC) {sometimes referred to as the "oscillatory failure case"} which can be caused, for example by electronic system component faults. The moving flight surface of an aircraft can sometimes experience oscillation which may be generated in the servo-loop control, i.e. between the flight control computer (FCC) and the actual control surface itself. The spurious sinusoidal signals can propagate through the FCC and hence the control surface, as shown in Fig.1 [11]. As the fault is a local phenomenon within a single actuator, it only has an impact on one control surface. This OFC scenario has been studied by [7, 11] to detect the OFC fault.

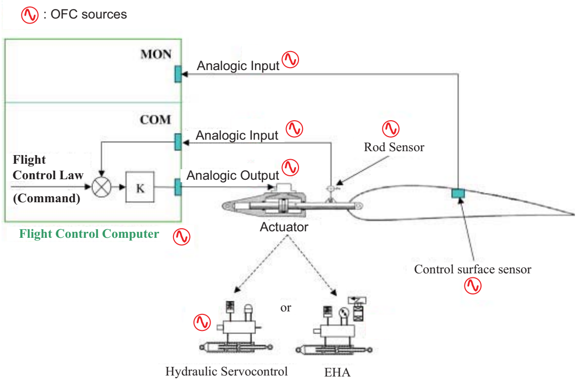


Figure 1. OFC source location in the control loop

A well-known method for estimating fault signals uses a combination of proportional and integral action within a full

The authors acknowledge funding from the EU FP7 project FP7-233815, Advanced Fault Diagnosis for Sustainable Flight Guidance and Control (ADDSAFE). Xiaoyu Sun has also been supported via a PhD scholarship in the Department of Engineering, Hull University.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

order identity observer [12]. However, these authors did not consider the robustness of the fault estimation to modeling uncertainty. The current work provides an extension to the work of [12] by using a UIO to take into account the effects of so-called UI signals. The estimator design problem is divided into two stages of (i) UI distribution matrix estimation followed by (ii) the actual fault estimation, with inclusion of proportional (not only integral) action to enhance to the fault estimation speed. The proposed approach is termed a Robust Fast Adaptive Fault Estimator (RFAFE) based on a combination of the UIO proposed in [6] and the Fast Adaptive Fault Estimator of [12]. The RFAFE is applied to the problem of estimating the oscillatory fault signal acting on an elevator of the ADDSAFE benchmark system. The benefit of the proposed RFAFE is that by making the output error of the observer insensitive to modelling uncertainty the fault estimation robustness is improved.

The robustness of the fault estimation is defined to be the degree of comparison between the sensitivity of the estimation to the fault compared with the sensitivity to modeling uncertainty. The fault estimation must be accurate with relative insensitivity to modelling uncertainty.

The nonlinear aircraft model is not available for publication due to confidential issues. However, the results in this paper have been generated by applying the new RFAFE fault estimation strategy to the fully non-linear aircraft system dynamics via the ADDSAFE project. Following a procedure in [6] the structure of the modelling uncertainty inherent between the nonlinear and linear time invariant (LTI) aircraft models are considered as UI terms in the linear models used for the development of the fault estimator. In stage (i) of the RFAFE design, the influences of the UIs are estimated by estimating the "directions" (i.e. distributions) of these terms into the state space model as described in [3, 13]. In stage (ii) the fault estimator is then applied directly to the fault residual signal. This study focuses on the problem of detecting OFC fault activity in one elevator actuator (referred to as the "left" actuator).

The structure of the paper is as follows: In Section II, the proposed RFAFE design is formulated. In Section III, the LTI longitudinal aircraft model dynamics including elevator model dynamics are constructed. The RFAFE method is applied to the ADDSAFE benchmark system in Section IV to estimate various OFC faults. The conclusion is given in Section V.

## II. ROUBST ADAPTIVE FAULT ESTIMATION THEORY

### *A. Fast adaptive fault estimation with UI decoupling*

A LTI system considering actuator faults (all sensors are assumed to be fault-free) and with modeling uncertainty, represented by the UI term $Ed(t)$ is represented as:

$$\left.\begin{aligned} \dot{x}(t) &= A\,x(t) + Bu(t) + Ed(t) + F_a f_a\,(t) \\ y(t) &= C\,x(t) \end{aligned}\right\} \tag{1}$$

where $x \in \Re^n$ denotes the time-varying system state vector, $u \in \Re^r$ and $y \in \Re^m$ denote the input and measurement vectors, respectively and $d \in \Re^p$ is a vector of UIs. $f_a \in \Re^l$ represents a vector of time-varying actuator faults. $A, B, C$ are known system matrices with appropriate dimensions. The matrix $E \in \Re^{n\times q}$ represents the distribution matrix for the UIs. The columns of the matrix $F_a \in \Re^{n\times l}$ denote the independent fault directions. It is thus considered that both $E$ and $F_a$ act as system inputs.

Following [3], a functional observer is constructed as:

$$\left.\begin{aligned} \dot{z}(t) &= Nz(t) + TBu(t) + Ky(t) + TF_a\hat{f}_a(t) \\ \hat{x}(t) &= z(t) + Hy(t) \end{aligned}\right\} \tag{2}$$

where $\hat{x} \in \Re^n$ is the estimated state vector and $z \in \Re^n$ is the observer state vector, and $N$ , $T$, $K$ and $H$ are design matrices.

**Definition 1**: Observer (2) is defined as a *robust fast adaptive fault estimator (RFAFE)* for the system (1), if its state and fault estimation errors $e_x = x - \hat{x}$ and $e_f = f_a - \hat{f}_a$ approach zero asymptomatically, in the presence of the system UIs and faults.

Assuming that $E$ is known, the estimation error dynamics are governed by:

$$\begin{aligned} \dot{e}_x(t) = (A - HCA - K_1C)e_x(t) \\ +[N - (A - HCA - K_1C)]z(t) \\ +[K_2 - (A - HCA - K_1C)H]y(t) \\ +[T - (I - HC)]Bu(t) \\ +(HC - I)Ed(t) + TF_a e_f(t) \end{aligned} \tag{3}$$

where

$$K = K_1 + K_2 \tag{4}$$

If the following relations are satisfied:

$$(HC - I)E = 0 \tag{5}$$

$$T = I - HC \tag{6}$$

$$N = A - HCA - K_1C = A_1 - K_1C \tag{7}$$

$$K_2 = NH \tag{8}$$

The state estimation error is then refined as:

$$\dot{e}_x(t) = Ne_x(t) + TF_a e_f(t) \tag{9}$$

$$e_y(t) = ce_x(t) \tag{10}$$

$$r(t) = y(t) - C\hat{x}(t) = Ce_x(t) = e_y(t) \tag{11}$$

Furthermore, if all eigenvalues of $N$ are stable, $r(t)$ will approach zero asymptotically, i.e. $\hat{x} \rightarrow x$ and $\hat{f}_a \rightarrow f_a$ . The observer (2) is an UI decoupling *fast adaptive fault estimator* for the system (1) when conditions (5) – (8) are satisfied. Therefore, this RFAFE design involves the solution of (4) to (8) whilst placing all the eigenvalues of the system matrix $N$ to be stable. Meanwhile, $N$ , $T$ , $K$ and $H$ in (2) are designed to achieve the required fault estimation performance.

A particular solution to (5) can be calculated as follows:

$$H = E(CE)^{+} \tag{12}$$

where: $(CE)^{+} = [(CE)^{T}CE)]^{-1}(CE)^{T}$ denotes the Moore-Penrose pseudo-inverse.

**Theorem 1**. The necessary and sufficient conditions for the existence of RFAFE of system (1) are [3, 12]:

(i) $rank(CE) = rank(E)$
(ii) $rank(CF_a) = rank(F_a)$
(iii) $rank([E\ F_a] = q + l$
(iv) $(C, A_1)$ is a detectable pair

**Remark 1**: Condition (i) denotes that the maximum number of independent UIs cannot be larger than the maximum number of independent measurements, i.e. the necessary condition for UI decoupling in the state estimation error dynamics is $rank(E) \leq m$. If this condition is not satisfied, a rank approximation via a matrix $E^{*}$ can be derived using Singular Value Decomposition (SVD) [14]. The details in [3] are addressed in Section B.

**Remark 2**: (ii) expresses that the maximum number of independent faults cannot be larger than the maximum number of independent measurements, i.e. the necessary condition for fault estimation in the states error dynamics is $rank(F_a) \leq m$.

**Remark3**: (iii) means that the UIs and faults are separable.

**Remark 4**: (iv) is equivalent to the following two equations.

*(1)* $rank\begin{bmatrix} sI - A & E \\ C & 0 \end{bmatrix} = n + rank(E)$

*(2)* $rank\begin{bmatrix} sI - A & B \\ C & 0 \end{bmatrix} = n + rank(B)$

Lemma 1 is used to verify the RFAFE existence conditions:

**Lemma 1** [15]: Given a scalar $\theta > 0$ and a symmetric positive definite matrix $P$, the following inequality holds:

$$2x^{T}y \leq (1/\theta)(x^{T}Px) + \theta y^{T}P^{-1}y \quad x, y \in \Re^{n} \tag{13}$$

Assume that $\dot{f}_a \neq 0$, e.g. a sinusoidal perturbation (as required for the OFC fault case). The derivative of $e_f$ is represented as:

$$\dot{e}_f = \dot{f}_a - \dot{\hat{f}}_a \tag{14}$$

The system error dynamics can be guaranteed by Theorem 2.

**Theorem 2**: With the assumption of Theorem 1, given the scalar $\alpha, \theta > 0$, if there exist symmetric positive definite matrices $P \in \Re^{n\times n}$, $Q \in \Re^{n\times n}$, $G \in \Re^{l\times l}$, and matrices $Y \in \Re^{n\times m}$, $N \in \Re^{l\times m}$ such that the following conditions hold.

$$\begin{bmatrix} PN + N^{T}P & -(1/\alpha)(N^{T}PTF_a) \\ * & -2(1/\alpha)(TF_a)^{T}PTF_a + (1/\alpha\theta)G \end{bmatrix} < 0 \tag{15}$$

$$(TF_a)^{T}P = MC \tag{16}$$

$*$ denotes the elements of a symmetric matrix, the UI decoupling fast adaptive fault estimator can be defined as:

$$\dot{\hat{f}}(t) = \Gamma M(\dot{r}(t) + \alpha r(t)) \tag{17}$$

(17) can be realized when $r(t)$ and $e_f$ are uniformly bounded functions. $\Gamma \in \Re^{l\times l}$ is a symmetric positive definite learning rate matrix.

**Proof**: Consider the following Lyapunov function:

$$V(t) = e_x^{T}(t)Pe_x(t) + (1/\alpha)\, e_f^{T}(t)\Gamma^{-1}e_f(t) \tag{18}$$

Substituting (9) and (17) into (18), the derivative of $V(t)$ with respect to time is derived as:

$$\begin{aligned} \dot{V}(t) &= \dot{e}_x^{T}(t)Pe_x(t) + e_x^{T}(t)P\dot{e}_x^{T}(t) + 2(1/\alpha)\, e_f^{T}\Gamma^{-1}\dot{e}_f(t) \\ &= e_x^{T}(t)(PN + N^{T}P)e_x(t) + 2e_x^{T}(t)PTF_a e_f(t) \\ &\quad -2(1/\alpha)\, e_f^{T}(t)M(\dot{r}(t) + \sigma r(t)) \\ &\quad -2(1/\alpha)\, e_f^{T}(t)\Gamma^{-1}\dot{f}(t) \end{aligned} \tag{19}$$

Using (16), the term $-2(1/\alpha)\, e_f^{T}(t)M\big(\dot{r}(t) + \sigma r(t)\big)$ on the left hand side of (19) can be rewritten as:

$$\begin{aligned} &-2(1/\alpha)\, e_f^{T}(t)M\big(\dot{r}(t) + \sigma r(t)\big) \\ &\quad = -2(1/\alpha)\, e_f^{T}(t)(TF_a)^{T}P\big(\dot{e}_x(t) + \sigma e_x(t)\big) \end{aligned} \tag{20}$$

Substituting (9) and (20) into (19), $\dot{V}(t)$ can be formulated as:

$$\begin{aligned} \dot{V}(t) &= \dot{e}_x^{T}(t)(PN + N^{T}P)e_x(t) \\ &\quad -2(1/\alpha)\, e_f^{T}(t)(TF_a)^{T}PNe_x(t) \\ &\quad -2(1/\alpha)e_f^{T}(t)(TF_a)^{T}PTF_a e_f(t) \\ &\quad -2(1/\alpha)\, e_f^{T}(t)\Gamma^{-1}\dot{f}(t) \end{aligned} \tag{21}$$

By using Lemma 1, the following inequality can be obtained:

$$\begin{aligned} -2(1/\alpha)\, e_f^{T}(t)\Gamma^{-1}\dot{f}(t) &\leq (1/\alpha\theta)\, e_f^{T}(t)Ge_f(t) \\ &\quad +(\theta/\alpha)f_1^{2}\lambda_{max}(\Gamma^{-1}G^{-1}\Gamma^{-1}) \end{aligned} \tag{22}$$

Substituting (22) into (21), $\dot{V}(t)$ can be reformulated as:

$$\dot{V}(t) = \varphi^{T}(t)\, \Xi\, \varphi(t) + \delta \tag{23}$$

where

$$\varphi(t) = \begin{bmatrix} e_x(t) \\ e_f(t) \end{bmatrix}, \quad \delta = (\theta/\alpha)f_1^{2}\lambda_{max}(\Gamma^{-1}G^{-1}\Gamma^{-1}),$$

$$\Xi = \begin{bmatrix} PN + N^{T}P & -(1/\alpha)(N^{T}PTF_a) \\ * & -2(1/\alpha)(TF_a)^{T}PTF_a + (1/\alpha\theta)G \end{bmatrix}$$

$TF_a$ is full column rank, under the condition of $\Xi < 0$, and $\epsilon = \lambda_{min}(-\Xi)$, then:

$$\dot{V}(t) < -\epsilon\|\varphi(t)\|^2 + \delta \tag{24}$$

for

$$\delta < \epsilon\|\varphi(t)\|^2 \tag{25}$$

Then, it follows that:

$$\dot{V}(t) < 0 \tag{26}$$

In terms of Lyapunov stability theory, (26) indicates that $e_x(t)$ and $e_f(t)$ converge to a small set of $\delta$. This ends the proof.

If the fault signal is defined as:

$$f_a(t) = \begin{cases} 0 & t \in (0,\ t_f) \\ f_a(t) & t \in (t_f,\ \infty) \end{cases} \tag{27}$$

The fault estimation can be derived by (17) and given as:

$$\hat{f}(t) = \Gamma M(r(t) + \alpha \int_{tf}^{t} r(t)\, d\tau) \tag{28}$$

From (28), it can be seen that the fault estimation includes both proportional and integral parts. The proportional part enhances the fault estimator dynamic performance giving improved fault estimation speed.

**Remark 5**: Although inequality (15) can be solved easily via the Matlab LMI tool box, the simultaneous solution of (15) and (16) is difficult to achieve using functions in the LMI tool box. However, the problem can be solved by reformulating (16) into (29), which leads to the solution of optimization problem:

$$\begin{bmatrix} -\gamma I & (TF_a)^T P - MC \\ * & -\gamma I \end{bmatrix} < 0 \tag{29}$$

The RFAFE derivation is complete with proof of stability.

Apart from guaranteeing the observer stability, the observer dynamic response plays an important role in obtaining a qualified observer performance achieved by forcing the poles to lie within suitable complex plane subregions comprising either vertical strips, disks, conic sectors etc. (or their combinations) using LMIs optimization [16]. Here, disk and vertical strip LMI regions are employed as a further refinement to improve the fault estimator dynamics with LMIs defined as:

**Definition 2**: $N$ is defined as in (7). Let $\mathcal{D}$ be an LMI subregion with characteristic function in the left hand side of the complex plane as a disk of radius $r$ and centre $(-q, 0)$. Then there exists a symmetric matrix $P$ such that:

$$\begin{pmatrix} -rP & qP + N^T P \\ * & -rP \end{pmatrix} < 0 \qquad P > 0 \tag{30}$$

Then, $N$ is called $\mathcal{D}$–stable.

**Definition 3**: $N$ is defined as in (7). Let $\mathcal{D}$ be a subregion which presents a $\zeta$-stability region in the left-half plane. $\mathcal{D}$ is an LMI region with characteristic function, so that there exists a symmetric matrix $P$ such that:

$$N^T P + PN + 2\zeta P < 0 \qquad P > 0 \tag{31}$$

Then, $N$ is called $\mathcal{D}$ –stable. Fig. 2 shows the RFAFE poles assignment within an subregion $\mathcal{D}$ of an intersection between a specified disk and vertical strip by solving (30) & (31).

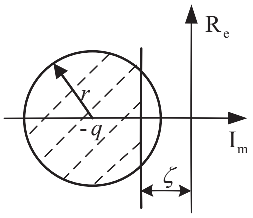


Figure 2. $\mathcal{D}$ subregion (hatched)

Consequently, a complete RFAFE with UI decoupling can be designed by solving (15), (29), (30), (31) and conditions (5)–(8) with the satisfaction of Theorem 2.

### *B. UI distribution matrix estimation*

An augmented state observer is utilized to estimate the UI distribution matrix $E$. Assume that $d_1(t) = Ed(t)$ is a slowly time-varying vector, then the system model can be formulated in augmented form as [3, 13]:

$$\begin{bmatrix} \dot{x}(t) \\ \dot{d}(t) \end{bmatrix} = \begin{bmatrix} A & I \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x(t) \\ d_1(t) \end{bmatrix} + \begin{bmatrix} B \\ 0 \end{bmatrix} u(t)$$

$$y(t) = \begin{bmatrix} C_a & 0 \end{bmatrix} \begin{bmatrix} x(t) \\ d_1(t) \end{bmatrix} \tag{32}$$

If the system inputs and outputs $\{u(t), y(t)\}$ are available, an observer based on the model presented by (32) can be used to estimate the $d_1(t)$ directly. The distribution matrix $E$ is calculated as the ratio of the elements of $\hat{d}_1(t)$. The necessary condition for observability is given in Theorem 3:.

**Theorem 3** The system (32) is observable if and only if the following conditions are satisfied [6].

(i) $rank(C) = n$
(ii) $rank(C, A)$ is a observable pair

**Remark 6**: Normally, the condition (i) could limit the application of this estimation approach. For a modern aircraft the states are available for measurement and this is also the case in the simulated non-linear system, so that the above observability and rank conditions are satisfied. Furthermore, the $E$ estimation is a state space problem, and hence independent of any measurement restrictions.

For the RFEFA design, the necessary rank condition $rank(E) \leq m$ has been given in Theorem 1. If this condition is not satisfied, a sub-optimal matrix $E^*$can be computed as follows via an SVD expansion of $E$ [3, 14]:

$$E = S\Sigma T^T \tag{33}$$

where:

$$\Sigma = \begin{bmatrix} diag\{\sigma_{1,\cdots,}\sigma_k\} & 0 \\ 0 & 0 \end{bmatrix} \tag{34}$$

$S$ and $T$ are orthogonal matrices, $k$ is the rank, and $\sigma_{1,\dots,}\sigma_k$ are the singular values of $E$ , respectively. A low rank approximation for $E$ by minimizing $\|E - E^*\|_F^2$ is given by:

$$E^* = S\hat{\Sigma}T^T \tag{35}$$

where

$$\hat{\Sigma} = \begin{bmatrix} diag\{0,\cdots,0,\sigma_{k-q,\cdots,}\sigma_k\} & 0 \\ 0 & 0 \end{bmatrix} \tag{36}$$

$q = rank(E^*) \leq m$ to satisfy Theorem 1 (for $E^*$ instead of $E$).

## III. LTI LONGITUDINAL AIRCRAFT MODEL DYNAMICS

The proposed RFAFE is implemented on a global longitudinal LTI model derived from the ADDSAFE benchmark system. Two parts constitute the global aircraft LTI model: one is the aircraft body axis LTI model derived from an LPV realization of the benchmark system obtained by choosing the trimming parameters given in TABLE I. The other part comprises the locally linear aircraft actuator models representing the right and left elevators on the aircraft tail surface. The linearized aircraft actuator models are generated for the same trimming parameters as the aircraft body axis LTI model.

TABLE I. TRIMMING POINTS FOR LONGITUDINAL AIRCRAFT LTI MODEL

| Trimming parameter | value |
|---|---|
| MASS (Net mass in Kg) | 200000 |
| XG (Centre gravity of the aircraft in % / 100) | 0.30 |
| ZP (Altitude in feet) | 20000 |
| VC (Calibrate aircraft speed in kts) | 290 |

The LTI local elevator model is represented by a first order system dynamic. For the longitudinal motion, the left and right elevator dynamics combined together have the structure:

$$\left.\begin{aligned} \dot{x}_a(t) &= A_a\, x_a(t) + B_a u_c(t) \\ y_a(t) &= C_a\, x_a(t) + D_a u_c(t) \end{aligned}\right\} \tag{37}$$

where $x_a(t) \in \Re^{2\times1}$ is the augmented state vector for both the left and right elevators. $u_c(t) \in \Re^{2\times1}$ is the vector of elevator control inputs (the actuator input signals fed from the FCC), $y_a$ is the actuator output. $A_a$, $B_a$, $C_a$,$D_a$ are corresponding system matrices with proper dimensions.

The LTI state space representation of the aircraft body axis dynamics can be expressed as:

$$\left.\begin{aligned} \dot{x}_b(t) &= A_b\, x_b(t) + B_u u_b(t) \\ y_b(t) &= \ C_b x_b(t) + D_u u_b(t) \end{aligned}\right\} \tag{38}$$

where $x_b = [V_{tas}, \ \alpha, \ q, \theta]$ and $y_b = [V_{tas}, \ \alpha, \ q, \theta]$ are the aircraft body axis states and outputs, respectively. $u_b$ is equal to $y_a$. $V_{tas}$ is the true air speed in m s$^{-1}$, $\alpha$ is the angle of attack in deg, $q$ is the pitch rate in deg s$^{-1}$, $\theta$ is pitch angle in deg. $A_b$, $B_b$,$C_b$, $D_b$ are the corresponding system matrices.

The complete LTI longitudinal motion model is formulated as:

$$\begin{bmatrix} \dot{x}_b(t) \\ \dot{x}_a(t) \end{bmatrix} = \begin{bmatrix} A_b & B_u C_a \\ 0 & A_a \end{bmatrix} \begin{bmatrix} x_b(t) \\ x_a(t) \end{bmatrix} + \begin{bmatrix} B_u D_a \\ B_a \end{bmatrix} u_c(t)$$

$$\begin{bmatrix} y_b(t) \\ y_a(t) \end{bmatrix} = \begin{bmatrix} C_b & D_u C_a \\ 0 & C_a \end{bmatrix} \begin{bmatrix} x_b(t) \\ x_a(t) \end{bmatrix} + \begin{bmatrix} D_u D_a \\ D_a \end{bmatrix} u_c(t) \tag{39}$$

(39) can be rewritten as:

$$\left.\begin{aligned} \dot{x}(t) &= A\, x(t) + Bu(t) \\ y(t) &= C\, x(t) + Du(t) \end{aligned}\right\} \tag{40}$$

LTI system (40) with UIs and faults can be presented as:

$$\left.\begin{aligned} \dot{x}(t) &= A\, x(t) + Bu(t) + Ed(t) + F_a f_a\, (t) \\ y(t) &= C\, x(t) + Du(t) \end{aligned}\right\} \tag{41}$$

## IV. SIMULATION RESULTS

In this paper, the RFAFE design is implemented on a generic AIRBUS aircraft model to estimate the left elevator OFC fault. Two types of OFC are classified, the “liquid” and “solid” faults. The liquid fault is considered as an additive fault which adds to the control command inside the control loop. The solid fault is considered as a ‘disconnected’ fault which substitutes the control command completely inside the control loop. Both of these two OFC faults lead to the control surface performing with a spurious control command. In this project, the OFC faults are simulated as sinusoidal signals within a range of magnitudes and frequencies. The estimated OFC fault signals are normalized into the entire interval [0, 1] according to the elevator control surface deflection range of operation. In this simulation result section, the OFC signals 0.016 and 0.33 (in normalized units) are estimated to (a) demonstrate the effect that the OFC has on the elevator operation and (b) the effectiveness of the RFAFE design.

The first step of the RFAFE design is to estimate the UI distribution matrix as an off-line analysis. The modelling uncertainties between the nominal nonlinear aircraft model and the LTI aircraft longitudinal model are considered as UIs. The off-line design of the ASO for $E$ estimation is made by running the ADDSAFE benchmark model. Six single fault-free cases (cruise phase, triggering of angle of attack protection, nose-up (abrupt longitudinal maneuver), triggering of pitch protection, coordinated turn and a “yaw-angle-mode” which roughly corresponds to an enhanced auto-pilot hold mode) are used to implement the estimation of the matrix $E$. The lower rank technique via the SVD approach is applied to post-process the modelling uncertainty data, so that condition (i) in Theorem 1 is satisfied. The second step of the RFAFE design is to construct the UI decoupling fast adaptive estimator in terms of the matrix $E$ estimated in step 1. A set of conditions should be satisfied first and a group of LMIs should be solved as discussed in Section II. All the designs described in this paper use one aircraft longitudinal LTI model that corresponds to the operating point in TABLE I. The left elevator fault direction $F_a$ is the first column of $B = [B_G, B_D]$, i.e. $F_a = B_G$.

Figs. 3&4 show the left elevator control surface position in two fault cases compared with the fault-free case, respectively. The control surface deflection is apparent, i.e. the OFC fault leads to unwanted control surface oscillation.

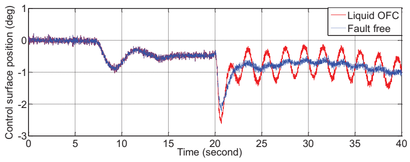


Figure 3. Left elevator control surface position (liquid OFC&fault-free cases)

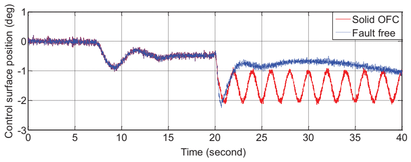


Figure 4. Left elevator control surface position (solid OFC & fault-free cases)

Fig. 3 corresponds to the liquid OFC. A sinusoidal signal is added to the normal control surface position. Fig. 4 shows the control surface movement trajectory is totally substituted by a sinusoidal signal for the solid OFC “disconnection” behaviour.

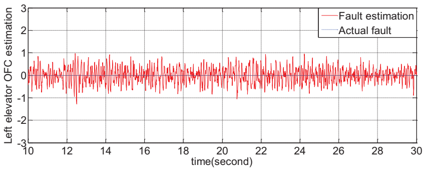


Figure 5. Left elevator fault estimation for the fault-free case

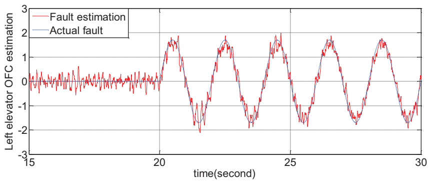


Figure 6. Left elevator fault estimation for the liquid OFC fault

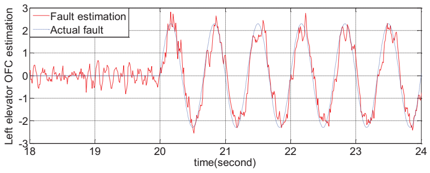


Figure 7. Left elevator fault estimation for solid OFC

Fig. 5 shows that in the fault-free case (left elevator) the estimates are in a noise level. In Figs. 6&7, the estimation of the so-called liquid OFC (0.016 OFC) and solid OFC (0.033 OFC) are shown, respectively. For each fault scenario, the faults occur at 20s and the fault estimation signals track the actual fault signals in magnitude and frequency.

The RFAFE design learning rate should be tuned to a suitable value to achieve accurate and fast fault estimation. It can also be seen that the fault estimation signal is not significantly affected by the modelling uncertainties, but is influenced by high frequency sensor noise. Hence, the fault estimation signal obtained is considered robust to modelling uncertainties.

## V. CONCLUSION

In this paper, the RFAFE approach to fast fault estimation has been applied to an aircraft actuator OFC problem, taking into account modeling uncertainties through UI estimation. The UIs reflect the modelling mismatch between the linear and non-linear aircraft systems. The UI estimation considering a range of flight conditions provides a structured approach to robustness which leads to a robust fault estimation. The UI estimation with UI decoupling is utilized in the RFAFE design. The results show that the fault estimation signals track the actual fault signals accurately under both liquid and solid OFC faults in different magnitudes and frequencies, demonstrating the effectiveness and efficiency of the RFAFE method.

## REFERENCES

[1] P. Goupil. AIRBUS State of the Art and Practices on FDI and FTC in Flight Control System. Control Engineering Practice 19 (2011), pp. 524-539 DOI information: 10.1016/j.conengprac.2010.12.009

[2] Goupil, P., Zolghadri, A., Gheorghe, A., Cieslak, J., Dayre, R. and Le-Berre H. Airbus efforts towards advanced fault diagnosis for flight control system actuators. 5th International Conference on Recent Advances in Aerospace Actuation Systems and Components (R3ASC'12), Toulouse, France, 13-14 June 2012.

[3] Chen, J. and R.J. Patton, *Robust model-based fault diagnosis for dynamic systems*. 1999, Norwell: Kluwer Academic Publishers

[4] Varga, A. Integrated algorithm for solving $H_2$-optimal fault detection and isolation problems. Proc.of Control and Fault-Tolerant Systems (SysTol). 2010.

[5] Chen, L. and R.J. Patton. Polytope LPV Fault Estimation for Non-Linear Flight Control. in Proceedings of the 18th IFAC World Congress. 2011.

[6] Chen, J., R.J. Patton, and H.Y. Zhang, Design of unknown input observers and robust fault detection filters. International Journal of Control, 1996. **63**(1): p. 85-105.

[7] Alwi, H. and C. Edwards. Oscillatory fault case detection for aircraft using an adaptive sliding mode differentiator scheme. in American Control Conference (ACC). 2011.

[8] Vanek, B., et al. Robust Model Matching for Geometric Fault Detection Filters: A Commercial Aircraft Example. 2011.

[9] DEIMOS Space, S.L.U, http://addsafe.deimos-space.com, 2011.

[10] Goupil, P. and A. Marcos. Advanced Diagnosis for Sustainable Flight Guidance and Control: the European ADDSAFE project. SAE 2011 AeroTech Congress & Exhibition October 18-21, 2011, Toulouse, France.

[11] Goupil, P., Oscillatory fault case detection in the A380 electrical flight control system by analytical redundancy. Control Engineering Practice, 2010. **18**(9): p. 1110-1119.

[12] Zhang, K., B. Jiang, and V. Cocquempot, *Adaptive observer-based fast fault estimation.* International Journal of Control Automation and Systems, 2008. **6**(3): p. 320.

[13] Patton, R.J. and J. Chen, Optimal unknown input distribution matrix selection in robust fault diagnosis. Automatica, 1993. **29**(4): p. 837-841.

[14] Golub, G.H. and C.F. Van Loan, *Matrix computations*. 1996, Baltimore and London: The Johns Hopkins University Press.

[15] Jiang, B., J.L. Wang, and Y.C. Soh, *An adaptive technique for robust diagnosis of faults with independent effects on system outputs.* International Journal of Control, 2002. **75**(11): p. 792-802.

[16] Chilali, M. and P. Gahinet, *H-infinity design with pole placement constraints: An LMI approach.* Automatic Control, IEEE Transactions on, 1996. **41**(3): p. 358-367.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# A variable structure observer for unknown input estimation in sampled systems

L. Orihuela, S. K. Spurgeon, X. G. Yan and F. R. Rubio

*Abstract*— **This paper considers the design of a variable structure observer for unknown input estimation and/or fault reconstruction in systems where the process measurements are sampled. It is well known that the principle of the equivalent injection signal from the sliding mode domain can be used for reconstruction of unknown inputs but much of the associated theory is predicated on output sampling of infinite frequency. Sample rate may be a physical constraint of the process and the reconstruction properties of such continuous time sliding mode observers degrade under this constraint. This paper explores how a recently developed ultimately bounded stable variable structure discrete time observer can be used for unknown input estimation. The main novelty of the approach is that the design of the observer is written as an optimization problem with linear constraints with the output sampling incorporated explicitly in the model used for observer design. The design methodology is shown to have advantages in terms of reconstruction accuracy when the performance is compared to that of a classical sliding mode observer on a case study.**

## I. INTRODUCTION

Variable structure systems were perhaps originally best known for their potential as a robust control method [1], [2], [3]. They are characterised by a suite of feedback control laws and a decision rule. The decision rule, termed the switching function, has as its input some measure of the current system behaviour and produces as an output the particular feedback controller which should be used at that instant in time. In sliding mode control, variable structure control systems are designed to drive and then constrain the system state to lie within a neighbourhood of the switching function. The paradigm has several advantages: the dynamic behaviour of the system may be specified by the choice of switching function and the system is completely insensitive to an important class of uncertainties. A disadvantage of the methodology has been the fundamentally discontinuous control signal which, in theoretical terms, must switch with infinite frequency to provide total rejection of uncertainty. Control implementation via approximate, smooth strategies is widely reported, but in such cases total invariance is routinely lost.

In contrast, the application of sliding mode methods to the observer problem is less mature and has some fundamentally different properties [4]. The discontinuous injection signals which were perceived as problematic for many control applications have no disadvantages for software based observer frameworks. The ability to generate a sliding motion on the error between the measured plant output and the output of the observer ensures that a sliding mode observer produces a set of state estimates that are precisely commensurate with the actual output of the plant. Further, analysis of the average value of the applied observer injection signal, the so-called equivalent injection signal, contains useful information about the mismatch between the model used to define the observer and the actual plant [5]. This property has been employed for general unknown input estimation as well as for fault reconstruction [6], [7], [8]. The results obtained to date most frequently require that an ideal sliding motion is attained in finite time and the effects of sampling on the physical measurements used to drive the observer are typically not considered within the observer design frameworks. However, in the presence of a sampled output, the ideal sliding mode cannot be achieved. Indeed, the error dynamics in the observer may become unstable if the sampling frequency is reduced significantly. The effect of output sampling on the performance of a sliding mode observer designed using classical continuous variable structure control theory has been discussed by several authors, see for example [9], where the fast sampling required for fault reconstruction via such a sliding mode observer on a motor experiment is reported.

In practice, the sample rate is not always a parameter that can be selected by the designer and in this case consideration must be given to developing design methods that incorporate the sampling characteristics if good estimates are to be obtained for the unknown inputs. Recent work has considered the development of a sliding mode observer in the presence of sampled output information and its application to fault reconstruction by using the delayed continuous-time representation of the sampled-data system, for which a set of Linear Matrix Inequalities (LMIs) provide conditions for ultimate boundedness of the solution [10]. An alternative approach is to consider a discrete time observer design methodology. Compared to the continuous time case, the literature in this area is sparse. Several contributions develop sliding mode observers for systems with a single output [11], [12], [13]. More recently, [14] studied sliding mode observers for a class of discrete-time multi-output systems, but the design of the observer is largely heuristic, and it is not possible to ensure *a priori* the stability of the observer. A variable structure observer design framework for discrete-time multi-output systems which uses Linear Matrix Inequalities to constructively exploit the degrees of freedom

L. Orihuela and F. R. Rubio would like to acknowledge CICYT (Grant DPI2010-19154) for funding this work.

L. Orihuela and F. R. Rubio are with with Dpto. de Ingeniería de Sistemas y Automática. Escuela Superior de Ingenieros. University of Seville, Spain orihuela,rubio@us.es

S. K. Spurgeon and X. G. Yan are with the School of Engineering and Digital Arts. University of Kent, United Kingdom S.K.Spurgeon, X.Yan@kent.ac.uk

within the design has been recently developed [15].

This paper extends this later framework to incorporate the estimation of faults and/or unknown inputs as a design requirement and assesses the degree to which the methodology overcomes problems of sampled data implementation frameworks for variable structure based signal reconstruction.

The paper is structured as follows. The problem is stated in Section II. The variable structure observer is proposed in Section III. The estimation of unknown inputs using the discrete-time observer is studied in Section IV. Some examples are presented in Section V. The paper ends with conclusions and a discussion of future research directions.

## II. PROBLEM STATEMENT

Consider the multi-output, discrete-time linear system described by

$$\begin{aligned} x(k+1) &= Ax(k)+Dw(k), && (1)\\ y(k) &= Cx(k), && (2) \end{aligned}$$

where $x \in \mathcal{R}^n$ is the state, $y \in \mathcal{R}^p$ is the output and $w \in \mathcal{R}^m$ is the unknown input. Matrices $A, D, C$ are known with appropriate dimensions. It is assumed that the pair $(A, C)$ is observable. Note that no control input affects the system, as the observer problem is the focus of this paper. The inclusion of the control signal can be trivially dealt with, as both the observer and the system are subject to the same input and thus the control signal has no affect on the dynamics of the error between the system and observer.

Assume that $rank(CD) = m$ and the invariant zeros of the triple $(A, D, C)$ lie inside the unit circle. Then, there exists a linear change of coordinates $T_o$ (see [3]) such that the system can be written as:

$$\begin{aligned} x_1(k+1) &= A_{11}x_1(k)+A_{12}z(k)+D_1w(k), && (3)\\ z(k+1) &= A_{21}x_1(k)+A_{22}z(k)+D_2w(k), && (4)\\ y(k) &= z(k), && (5) \end{aligned}$$

where $z(k) \in \mathcal{R}^p$, $x_1 \in \mathcal{R}^{n-p}$ and $D_2 \in \mathcal{R}^{p\times m}$. Matrix $A_{11}$ is stable and $D_2$ has full column rank.

Unlike [10] and similar approaches, the disturbance process or unknown input $w(k)$ affects both dynamics. The classical nomenclature used in sliding mode theory, namely, unmatched and matched disturbances has been adopted (see [3]). It is assumed to be bounded by

$$\|w(k)\| \leq \xi,\ \forall k,$$

where $\xi$ is a known positive scalar.

The following section is devoted to the design of an ultimately bounded variable structure observer for the system (1), by driving the observation error $e(k)$ to the vicinity of the equilibrium point $e(k) = 0$ in finite time and maintaining it in the neighbourhood thereafter. Due to the presence of the disturbances, asymptotic stability is not possible. However, the proposed observer reduces the ultimate bound on the response when compared to classical observation strategies. In Section IV, the properties of this observer will be used in order to estimate the unknown input $w(k)$.

## III. VARIABLE STRUCTURE OBSERVER

The proposed variable structure observer for the multi-output system (3)-(5) is defined as

$$\begin{aligned} \hat{x}_1(k+1) &= A_{11}\hat{x}_1(k)+A_{12}y(k) && (6)\\ \hat{z}(k+1) &= A_{21}\hat{x}_1(k)+A_{22}\hat{z}(k) && (7)\\ & \quad -(A_{22}-A_{22}^s)e_z(k)+\nu(k) \end{aligned}$$

where $e_z(k) = \hat{z}(k) - y(k)$ and $A_{22}^s \in \mathcal{R}^{p\times p}$ is a design matrix. The variable structure term $\nu(k)$ is defined by:

$$\nu(k) = Bf_{sat}(e_z(k),\Delta) = B\begin{bmatrix} sat\left(\frac{e_{z1}(k)}{\Delta}\right)\\ sat\left(\frac{e_{z2}(k)}{\Delta}\right)\\ \vdots\\ sat\left(\frac{e_{zp}(k)}{\Delta}\right) \end{bmatrix} \quad (8)$$

where $\Delta$ is a positive scalar and $B \in \mathcal{R}^{p\times p}$ is also a design matrix. The function $sat(.)$ is defined as:

$$sat\left(\frac{e_{zi}(k)}{\Delta}\right) = \begin{cases} sgn(e_{zi}(k)), & |e_{zi}(k)| > \Delta\\ \frac{e_{zi}(k)}{\Delta}, & |e_{zi}(k)| \leq \Delta \end{cases}$$

It can be viewed as a set of $p$ unidimensional switching functions. Each one switches whenever the associated component of the output observation error $e_{zi}$, $i = 1..., p$ crosses the boundary of the region.

Let the state estimation errors be $e_1(k) = \hat{x}_1(k) - x_1(k)$ and $e_z(k) = \hat{z}(k) - z(k)$. It follows that the error dynamics are

$$\begin{aligned} e_1(k{+}1) &= A_{11}e_1(k){-}D_1w(k), && (9)\\ e_z(k{+}1) &= A_{21}e_1(k){+}A_{22}^se_z(k){-}D_2w(k){+}\nu(k). && (10) \end{aligned}$$

where $y(k) = z(k)$ in (5) is used to obtain the equations above.

In order to study the stability of the observer, a Lyapunov framework is used. Defining $\Delta$ as the ultimate boundedness, the objective of this section is to design the variable structure observer in such a way that $\Delta$ is minimized ensuring that the forward increment of the Lyapunov function is negative for all $k$ such that $|e_z(k)| \geq \Delta$. Consider the following Lyapunov function:

$$V(k) = e_1^T(k)P_1e_1(k) + e_z^T(k)P_2e_z(k), \quad (11)$$

where $P_1, P_2$ are positive definite matrices of appropriate dimensions.

Due to the presence of the saturation, $\nu(k)$ is a nonlinear function. A linear representation of the saturation is introduced which will be useful when designing the observer via linear matrix inequalities. This idea was presented in [15].

Denote by $\mathfrak{E}_\beta$ the set of states such that

$$\begin{aligned} \mathfrak{E}_\beta = \{&(e_1(k), e_z(k)) \mid e_1(k) \in \mathcal{R}^{n-p}, e_z(k) \in \mathcal{R}^p,\\ &V(k) = e_1^T(k)P_1e_1(k) + e_z^T(k)P_2e_z(k) \leq \beta^{-1}\}, \end{aligned}$$

for a positive scalar $\beta$. The following lemma gives the linear representation of the nonlinear dynamics.

**Lemma 1. [15]** Given $\beta > 0$, assume that there exists a matrix $H_z \in \mathcal{R}^{p\times p}$, such that $|h_{zi} e_z| \leq 1$, for all $e_z \in \mathfrak{E}_\beta$, where $h_{zi}$ denotes the $i$-th row of $H_z$. Then, for $(e_1, e_z) \in \mathfrak{E}_\beta$, the observation error dynamic system (9)-(10) with switching function (8) admits the following representation:

$$\begin{aligned} e_1(k+1) &= A_{11}e_1(k) - D_1 w(k), \\ e_z(k+1) &= A_{21}e_1(k) + A_{22}^s e_z(k) - D_2 w(k) \\ &+ B\sum_{j=1}^{2^p} \lambda_j(k) A_j e_z(k), \end{aligned}$$

where:

$$\begin{aligned} A_j &= F_j K + F_j^- H_z, \quad j = 1, .., 2^p, \\ \sum_{j=1}^{2^p} \lambda_j(k) &= 1, \quad \lambda_j(k) \geq 0, \forall k > 0, \\ K &= diag\{1/\Delta, ..., 1/\Delta\}, \end{aligned}$$

with $F_j$ a diagonal matrix with diagonal elements that are either 1 or 0, and $F_j^- \triangleq I_m - F_j, \forall j$.

From Lemma 1, it can be concluded that the linear representation for the saturation is valid only if the region $\mathfrak{E}_\beta$ (defined by the parameter $\beta$) and a corresponding matrix $H_z$ can be found such that the error remains in $\mathfrak{E}_\beta$ for all $k$. Hence, this must be an additional constraint in the design of the observer.

The following theorem presents the main result of this section.

**Theorem 1.** Given the disturbance bound $\xi$ and the size $\beta$ of the set $\mathfrak{E}_\beta$, if positive definite matrices $P_1, P_2$, matrices $A_{22}^s, B, H_z$ of appropriate dimensions and scalar $\tau > 0$ solve the following optimization problem

$$\min_{A_{22}^s, B, P_1, P_2, H_z, \tau} \Delta \tag{12}$$

subject to (13)-(14), then the observation error dynamic system (9)-(10) is ultimately bounded stable. The minimum boundedness is $\Delta$.

**Proof.** It is first proved that conditions (13)-(14) imply that, given a positive $\Delta$, the forward increment of the Lyapunov function (11) decreases for all $|e_z(k)| \geq \Delta$.

The inequalities (14) guarantee that $|h_{zi} e_z| \leq 1$, $(i = 1, ..., p)$ for all $(e_1, e_z) \in \mathfrak{E}_\beta$. This results from the fact that any error belonging to $\mathfrak{E}_\beta$ satisfies

$$\begin{aligned} \beta e_1^T(k) P_1 e_1(k) + \beta e_z^T(k) P_2 e_z(k) \leq 1 \\ \Rightarrow \beta e_z^T(k) P_2 e_z(k) \leq 1 \end{aligned}$$

Then, for $(e_1, e_z) \in \mathfrak{E}_\beta$, the following inequalities

$$2 \geq 1 + \beta e_z^T(k) P_2 e_z(k) \geq 2|h_{zi} e_z|$$

imply that $|h_{zi} e_z| \leq 1$ for $i = 1, ..., p$. The latter inequality, which can be written as

$$\begin{bmatrix} 1 & \pm e_z^T \end{bmatrix} \begin{bmatrix} 1 & h_{zi} \\ * & \beta P_2 \end{bmatrix} \begin{bmatrix} 1 \\ \pm e_z \end{bmatrix} \geq 0,$$

is satisfied by (14).

As the assumption of Lemma 1 is verified, the polytopic description of the system given in that lemma holds. Using this linear representation, the Lyapunov function (11) at $k+1$ for vertex $j$ of the polytope is[1]

$$\begin{aligned} & V_j(k+1) = \\ & = (A_{11}e_1 - D_1 w)^T P_1 (A_{11}e_1 - D_1 w) \\ & + \big(A_{21}e_1 + M_{2j}e_z - D_2 w\big)^T P_2 \big(A_{21}e_1 + M_{2j}e_z - D_2 w\big), \end{aligned}$$

where

$$M_{2j} = A_{22}^s + BF_j K + BF_j^- H_z.$$

The forward increment of the Lyapunov function will be

$$\begin{aligned} \Delta V_j(k) = {} & V_j(k+1) - V_j(k) \\ & = -e_1^T Q_1 e_1 - 2e_1^T A_{11}^T P_1 D_1 w + w^T D_1^T P_1 D_1 w \\ & + e_1^T A_{21}^T P_2 A_{21} e_1 + 2e_1^T A_{21}^T P_2 M_{2j} e_z \\ & - 2e_1^T A_{21}^T P_2 D_2 w + e_z^T (M_{2_j}^T P_2 M_{2j} - P_2) e_z \\ & - 2e_z^T M_{2_j}^T P_2 D_2 w + w^T D_2^T P_2 D_2 w \end{aligned} \tag{15}$$

where $-Q_1 \triangleq A_{11}^T P_1 A_{11} - P_1$. The positive term $\tau w^T(k) w(k)$ can be bounded by:

$$\tau w^T(k) w(k) \leq \tau\xi^2 \leq \frac{\tau\xi^2}{\Delta^2} e_z^T(k) e_z(k),$$

taking into account that $\|e_z(k)\| \geq \Delta$. Therefore,

$$\frac{\tau\xi^2}{\Delta^2} e_z^T(k) e_z(k) - \tau w^T(k) w(k) \geq 0 \tag{16}$$

From (15) and (16), it follows that

$$\begin{aligned} \Delta V_j(k) & = V_j(k+1) - V_j(k) \\ & = -e_1^T \left(A_{11}^T P_1 A_{11} - P_1 + A_{21}^T P_2 A_{21}\right) e_1 + \\ & 2e_1^T A_{21}^T P_2 M_{2j} e_z - 2e_1^T \left(A_{11}^T P_1 D_1 + A_{21}^T P_2 D_2\right) w \\ & + e_z^T \left(M_{2_j}^T P_2 M_{2j} - P_2\right) e_z - 2e_z^T M_{2_j}^T P_2 D_2 w \\ & + w^T \left(D_1^T P_1 D_1 + D_2^T P_2 D_2\right) w \\ & + \frac{\tau\xi^2}{\Delta^2} e_z^T(k) e_z(k) - \tau w^T(k) w(k) \end{aligned}$$

Then, the increment of the Lyapunov function can be written in the following quadratic manner:

$$\Delta V_j(k) \leq \zeta^T(k) \Xi_j \zeta(k),$$

where the stacked state vector is

$$\zeta(k) = \begin{bmatrix} e_1(k) \\ e_z(k) \\ w(k) \end{bmatrix}$$

and the symmetric matrix $\Xi_j$ is given in equation (13).

The inequalities (13) imply that matrices $\Xi_j$ are negative definite for all the vertices of the polytope, and then, the forward increment of the Lyapunov function will be negative for all $\zeta(k) \neq 0$ (see [16]), which ensures the asymptotic stability of the system.

[1] The time script $k$ has been removed for ease of exposition.

$$\begin{bmatrix} A_{11}^T P_1 A_{11} - P_1 + A_{21}^T P_2 A_{21} & A_{21}^T P_2 M_{2j} & -A_{11}^T P_1 D_1 - A_{21}^T P_2 D_2 \\ * & M_{2j}^T P_2 M_{2j} - P_2 + \frac{\tau \xi^2}{\Delta^2} I & -M_{2j}^T P_2 D_2 \\ * & * & -\tau I + D_1^T P_1 D_1 + D_2^T P_2 D_2 \end{bmatrix} < 0, j = 1, ..., 2^p, \quad (13)$$

$$\begin{bmatrix} 1 & h_{zi} \\ * & \beta P_2 \end{bmatrix} \geq 0, i = 1, ..., p. \quad (14)$$

where

$$M_{2j} = A_{22}^s + BF_j K + BF_j^- H_z.$$

Finally, it must be ensured that the state of the system remains in $\mathfrak{E}_\beta$. To demonstrate this, the fact that the set $\mathfrak{E}_\beta$ is an invariant set is utilised so that from any initial condition $e_1(0), e_z(0)$ in $\mathfrak{E}_\beta$, any error $e_1(k), e_z(k)$ will belong to $\mathfrak{E}_\beta$, for all $k \geq 0$. The reason is clear as $\Delta V(k)$ is negative definite, then $V(k) \leq V(0) \leq \beta^{-1}$.

Finally, the optimization problem is introduced to minimize the size of the ultimate boundedness $\Delta$. This concludes the proof. □

Theorem 1 does not give any insights into the design of the observer. There are many unknown matrices that must be designed: some related to the observer dynamics $A_{22}^s, B$ and some are needed for stability considerations such as $P_1, P_2, H_z$. There are also constants $\Delta, \beta, \tau$ to be selected. In the following subsection some modifications on the conditions of Theorem 1 are introduced in such a way that some Linear Matrix Inequalities (LMIs) are obtained, which can be efficiently solved using appropriate software tools.

### A. OBSERVER DESIGN VIA LMI

Assume that the matrices $A_{22}^s$ and $H_z$ have been well designed. Imposing a particular choice of $A_{22}^s$, define the dynamics of the observation error when there are no disturbances (see eq. (10)). The following lemma can be used to design the observer.

**Lemma 2.** Given matrices $A_{22}^s, H_z$ and scalars $\beta, \xi$, if positive definite matrices $P_1, P_2$, matrix $W$ of appropriate dimensions and scalar $\tau > 0$ solve the following optimization problem

$$\min_{P_1, P_2, W, \tau} \Delta$$

subject to

$$\begin{bmatrix} -P_1 & 0 & 0 & A_{11}^T P_1 & A_{21}^T P_2 \\ * & -P_2 + \tau \frac{\xi^2}{\Delta^2} I & 0 & 0 & \Theta_j \\ * & * & -\tau I & -D_1^T P_1 & -D_2^T P_2 \\ * & * & * & -P_1 & 0 \\ * & * & * & * & -P_2 \end{bmatrix} < 0,$$

$$j = 1, ..., 2^p, \quad (17)$$

$$\begin{bmatrix} 1 & h_{zi} \\ * & \beta P_2 \end{bmatrix} \geq 0, \quad i = 1, ..., p. \quad (18)$$

with

$$\Theta_j \;\; = \;\; A_{22}^{sT} P_2 + K^T F_j W^T + H_z^T F_j^- W^T,$$

then the observation error dynamic system (9)–(10) is ultimately bounded stable by taking $B = P_2^{-1} W$. Matrix $h_{zi}$ denotes the i-th row of $H_z$.

**Proof.** See Appendix.

The optimization of scalar $\Delta$ can be easily carried out by means of a bisection algorithm or similar. The conditions are linear matrix inequalities with design parameters $P_1, P_2, W, \tau$, so the problem can be easily solved using appropriate software.

## IV. UNKNOWN INPUT ESTIMATION

This section is devoted to the unknown input estimation properties of the variable structure observer designed previously. Specifically, the switching function $\nu(k)$ contains useful information about the mismatch between the model used to define the observer and the actual plant.

Let $\bar{w}(k)$ denote the estimate of the unknown input. From the dynamics of the observation error (9), the actual error can be estimated as

$$e_1(k) = A_{11}^k e_1(0) + \sum_{i=0}^{k-1} (A_{11}^{k-1-i} D_1 \bar{w}(i)).$$

As $A_{11}$ is stable, the first term vanishes in some steps, so:

$$e_1(k) \approx \sum_{i=0}^{k-1} (A_{11}^{k-1-i} D_1 \bar{w}(i)). \quad (19)$$

Note that actual $e_1(k)$ depends on past values of $\bar{w}(k)$. On the other hand, using the output error dynamics (10), and assuming that slow disturbances will imply slow $e_z$, then:

$$\begin{aligned} e_z(k) &\approx e_z(k+1), \\ &\approx A_{21} e_1(k) + A_{22}^s e_z(k) - D_2 \bar{w}(k) + \nu(k). \end{aligned}$$

As the observer is asymptotically stable when $|e_z(k)| > \Delta$, it can be assumed that it is evolving inside the ball for sufficiently large $k$. In that case,

$$\nu(k) = \frac{B}{\Delta} e_z(k) \quad \Rightarrow \quad e_z(k) = \Delta B^{-1} \nu(k),$$

assuming nonsingular $B$. Substituting in the previous equation:

$$D_2 \bar{w}(k) \approx A_{21} e_1(k) + [\Delta(A_{22}^s - I) B^{-1} + I] \nu(k).$$

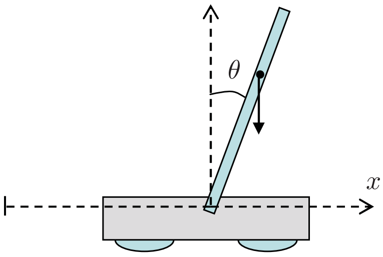


Fig. 1. Scheme of the inverted pendulum with cart

By least squares, the unknown input can be estimated using the $e_1(k)$ given in (19) and the actual value of $\nu(k)$:

$$\bar{w}(k) \approx M\left(A_{21}e_1(k) + [\Delta(A_{22}^s - I)B^{-1} + I]\nu(k)\right),$$

where $M = (D_2^T D_2)^{-1}D_2^T$. Here the matrix $D_2^T D_2$ is nonsingular because $D_2$ has full column rank.

## V. EXAMPLE: INVERTED PENDULUM

The problem of the inverted pendulum with a cart constitutes a benchmark study for the application of nonlinear design methods, [17]. The problem also lends itself to assessment of linear frameworks, as linearization errors are a motivation for control engineers to employ robust control techniques and observers. Consider the inverted pendulum with a cart shown in Figure 1. Using the same model given in [3], the equations of motion are

$$\begin{aligned} (M+m)\ddot{x} + F_x\dot{x} + ml(\ddot{\theta}\cos\theta - \dot{\theta}^2\sin\theta) &= u, (20) \\ J\ddot{\theta} + F_\theta\dot{\theta} - mlg\sin\theta + ml\ddot{x}\cos\theta &= 0, (21) \end{aligned}$$

where the values of the physical parameters used are given in Table I.

TABLE I
MODEL PARAMETERS FOR THE INVERTED PENDULUM WITH CART

| | | | | | |
|---|---|---|---|---|---|
| $M$ | $(kg)$ | 3.2 | $F_x$ | $(kg/s)$ | 6.2 |
| $m$ | $(kg)$ | 0.535 | $F_\theta$ | $(kg\ m^2)$ | 0.009 |
| $J$ | $(kg\ m^2)$ | 0.062 | $g$ | $(m/s^2)$ | 9.8 |
| $l$ | $(m)$ | 0.365 | | | |

To evaluate the performance of the proposed unknown input observer, it will be compared with the continuous sliding mode observer proposed in [3]. However, to make an appropriate comparison, additional sampling will be introduced between the plant and observer in this continuous version to reflect the practical situation whereby signals from the plant are sampled. In both cases, the system must be linearised around the equilibrium point at the origin. Using $x, \theta, \dot{x}$ and $\dot{\theta}$ as system states, and assuming that only $\theta, x$ and $\dot{x}$ are available as measured outputs, the discrete-time model with sampling time $T_m = 0.1s$ is given by the triple

$$A = \left[\begin{array}{c|ccc} 0.1051 & 0 & -7.4143 & 2.5489 \\ \hline -0.0003 & 1 & -0.0119 & 0.0915 \\ 0.1053 & 0 & 2.2255 & -0.3016 \\ -0.0084 & 0 & -0.2681 & 0.8447 \end{array}\right], B = \left[\begin{array}{c} 0.0435 \\ \hline 0.0015 \\ -0.0049 \\ 0.0293 \end{array}\right]$$

$$C = \left[\begin{array}{c|ccc} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right]$$

It is assumed that in both cases, the same control signal is used. Specifically, a sliding mode controller is implemented. Furthermore, it is assumed that the unknown inputs enter the system through the same channel as the control input, that is, $D = B$. Using Lemma 2 with $\beta = 0.01$ and $\xi = 0.15$, the optimization problem leads a minimum bound of $\Delta = 0.02$. The variable structure observer is defined by

$$\begin{aligned} A_{22}^s &= \begin{bmatrix} 0.3000 & 0 & 0 \\ 0 & 0.2000 & 0 \\ 0 & 0 & 0.4000 \end{bmatrix}, \\ B &= \begin{bmatrix} -0.0062 & 0.0000 & -0.0000 \\ 0.0000 & -0.0042 & 0.0000 \\ -0.0000 & 0.0000 & -0.0088 \end{bmatrix} \end{aligned}$$

Figures 2 and 3 compare the results of both observers. The simulations have been performed using the nonlinear model for the pendulum given in equations (20)-(21). The sum of two sinusoidal functions has been applied as the unknown input.

Figure 2 shows that the proposed observer exhibits an initial transient time after which a good estimate of the unknown input is obtained. This is because some of the assumptions made in Section IV are correct only if the observation error is close to zero. Moreover, linearization errors are more apparent at the transient, when the pendulum is far from the equilibrium. However, after this transient time, Figure 3 reveals that the discrete-time observer produces a better estimate than its continuous counterpart. Comparing the error in the steady state, the continuous observer has a maximum absolute error of $0.0295$ units whereas the discrete observer achieves a maximum error of $0.0183$. There is thus a circa $60\%$ difference in the error bound, with the proposed observer providing the greater accuracy of reconstruction.

## VI. CONCLUSIONS

Following the same framework as in [15], an unknown input observer has been presented for sampled systems. The proposed method has several advantages. On the one hand, the design exploits all the degrees of freedom available in the observer framework proposed in [15] in a positive way. On the other hand, the quality of the estimate of the unknown inputs is higher when compared with a classical continuous sliding mode observer. In the nonlinear bench mark inverted pendulum with cart example, a circa $60\%$ improvement in the error bound is achieved with the proposed synthesis. It is clear that the current methodology relies upon knowledge of the sample rate used for implementation and development of methods which incorporate sampling of uncertain or variable rate must now be considered.

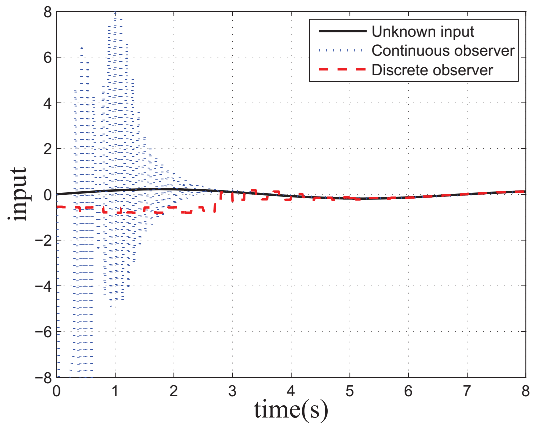


Fig. 2. Unknown input estimation with the continuous SMO (dotted line) and the discrete VSO (dashed line)

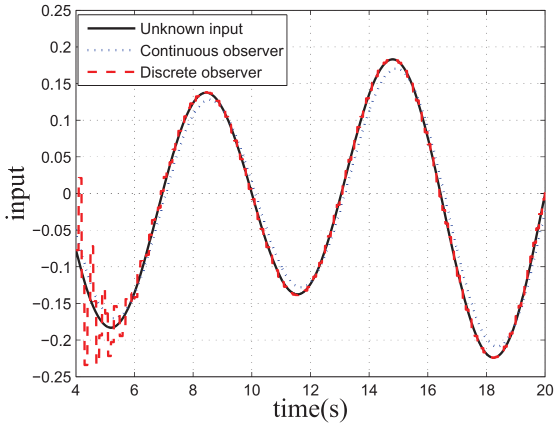


Fig. 3. Unknown input estimation with the continuous SMO (dotted line) and the discrete VSO (dashed line)

## Appendix

It will be shown that the conditions given in Lemma 2 and in Theorem 1 are equivalent. The first set of inequalities (13), $\Xi_j < 0$ can be rewritten as:

$$\begin{bmatrix} -P_1 & 0 & 0 \\ * & -P_2+\tau\frac{\xi^2}{\Delta^2}I & 0 \\ * & * & -\tau I \end{bmatrix} + \\ + \begin{bmatrix} A_{21}^T \\ M_{2j}^T \\ -D_2^T \end{bmatrix} P_2 \begin{bmatrix} A_{21} & M_{2j} & -D_2 \end{bmatrix} + \\ + \begin{bmatrix} A_{11}^T \\ 0 \\ -D_1^T \end{bmatrix} P_1 \begin{bmatrix} A_{11} & 0 & -D_1 \end{bmatrix} < 0,$$

for $j = 1, ..., 2^p$. Using the Schur complement, the previous inequalities are equivalent to

$$\begin{bmatrix} -P_1 & 0 & 0 & A_{11}^T & A_{21}^T \\ * & -P_2+\tau\frac{\xi^2}{\Delta^2}I & 0 & 0 & M_{2j}^T \\ * & * & -\tau I & -D_1^T & -D_2^T \\ * & * & * & -P_1^{-1} & 0 \\ * & * & * & * & -P_2^{-1} \end{bmatrix} < 0,$$

for $j = 1, ..., 2^p$. Pre- and post-multiplying the previous inequalities by $diag\{I, I, I, P_1, P_2\}$ and its transpose, conditions (17) are obtained. Then, the proof is finished by direct application of Theorem 1. □

## References

[1] U. Itkis, *Control systems of variable structure*. New York: Wiley, 1976.

[2] V. I. Utkin, *Sliding modes in Control Optimisation*. Berlin: Springer-Verlag, 1992.

[3] C. Edwards and S. K. Spurgeon, *Sliding Mode Control: Theory and applications*. Taylor and Francis, 1998.

[4] S. K. Spurgeon, “Sliding mode observers - a survey,” *International Journal of Systems Science*, vol. 39, no. 8, pp. 751–764, 2008.

[5] C. Edwards, S. K. Spurgeon, and R. J. Patton, “Sliding mode observers for fault detection and isolation,” *Automatica*, vol. 36, no. 4, pp. 541–553, 2000.

[6] T. Floquet and J. P. Barbot, “Super twisting algorithm based step-by-step sliding mode observers for nonlinear systems with unknown inputs,” *International Journal of Systems Science*, vol. 38, no. 10, pp. 803–815, 2007.

[7] T. Floquet, C. Edwards, and S. K. Spurgeon, “On sliding mode observers for systems with unknown inputs,” *International Journal of Adaptive Control and Signal Processing*, vol. 21, no. 8-9, pp. 803–815, 2007.

[8] C. Edwards and C. P. Tan, “Sensor fault tolerant control using sliding mode observers,” *Control Engineering Practice*, vol. 14, no. 8, pp. 897–908, 2006.

[9] ——, “Fault tolerant control using sliding mode observers,” in *43rd IEEE Conference on Decision and Control*, Nassau, Bahamas, December 2004, pp. 5254–5259.

[10] X. Han, E. Fridman, and S. K. Spurgeon, “A sliding mode observer for fault reconstruction under output sampling: A time-delay approach,” in *50th IEEE Conference on Decision and Control and European Control Conference*, Orlando, FL, USA, December 2011, pp. 77–82.

[11] A. J. Koshkouei and A. S. I. Zinober, “Sliding mode state observers for siso linear discrete-time systems,” in *UKACC International Conference on Control*, Swansea , UK, September 1998, pp. 837–842.

[12] ——, “Sliding mode state observers for discrete-time linear systems,” *International Journal of Systems Science*, vol. 33, no. 9, pp. 751–758, 2002.

[13] S. M. Lee and B. H. Lee, “A discrete-time sliding mode controller and observer with computation time delay,” *Control Engineering Practice*, vol. 7, no. 8, pp. 943–955, 2005.

[14] K. C. Veluvolu, S. Pavuluri, Y. C. Soh, W. Cao, and Z. Y. Liu, “Observers with multiple sliding modes for uncertain linear MIMO systems,” in *IEEE Conference on Industrial Electronics and Applications*, Singapore, May 2006, pp. 1–6.

[15] L. Orihuela, X. Yan, S. K. Spurgeon, and F. R. Rubio, “Variable structure observer for discrete-time multi-output systems,” in *12th International Workshop on Variable Structure Systems*, Mumbai, India, January 2012.

[16] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in system and control theory*. Philadelphia: Society for Industrial and Applied Mathematics, 1994.

[17] J. Aracil and F. Gordillo, “El péndulo invertido: un desafío para el control no lineal,” *Revista Iberoamericana de Automática e Informática Industrial*, vol. 2, no. 2, pp. 8–19, 2005.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Free Flight Concept Formulation Exploiting Neighbouring Optimal Control Concepts

S.N.Hasan
Department of Automatic Control and
Systems Engineering
University of Sheffield
Mappin Street, Sheffield S1 3JD
Email: saam.najat@sheffield.ac.uk

J.A.Rossiter
Department of Automatic Control and
Systems Engineering
University of Sheffield
Mappin Street, Sheffield S1 3JD
Email: j.a.rossiter@sheffield.ac.uk

***Abstract*—The goal of this paper is to develop a new approach for the Free Flight concept based on neighbouring Optimal Control (NOC) that can deliver an aircraft to its designated position safely and reliably when faced with conflicts, and to have some form of recovery measure when the pilot is faced with uncertainty that are due to wind and other uncertainties in the FF environment. The benefits of the algorithm are that it reduces fuel consumption, and it achieves optimal safe inter-aircraft separation while maintaining near-optimality under disturbances.**

***Index Terms*—NOC, Air Traffic Control, Fuel Consumption, Conflict Detection, Separation Constraints.**

## I. Introduction

Recent advancements in navigation aids, communication technologies, and computing power makes it possible for new concepts of Air Traffic Control (ATC) to be implemented, namely the Free Flight (FF), which is an environment where pilots have the liberty to select their routes in real time, and also define their own cruising altitudes and speeds for better fuel efficiency [2]. The workload of ATC would be greatly reduced, since pilots would be responsible for immediate threat avoidance when encountered. ATC then would be responsible for higher level tasks, such as monitoring traffic flow, airspace control, sectorisation, and scheduling runways. In the current centralised environment, ATC has to deal with potentially thousands of aircraft spread over a vast geographic area in order to ensure resolution of all possible conflicts [2]. Furthermore, as the number of aircraft increases, the complexity of conflict resolution grows and quickly becomes computationally intractable. Similarly, the possibility that a single failure could disable a significant portion of the centralized system creates a highly undesirable risk.

One of the most important tasks for ATC is conflict detection and resolution (CDR). Much of the current research for automating ATC focuses on this problem. Solutions for Conflict Resolution (CR) that are based on fixed rule sets that dictate actions based on situational geometry has been widely studied in both centralized and distributed algorithms. Very few of these techniques incorporate free flight rules in the algorithm, and that FF research has mainly been focused on the concept rather than defining automation tools for this concept. The goal of this paper is to exploit an approach based on neighbouring Optimal Control (NOC) to achieve a near-optimal FF air traffic control when faced with conflicts, and to have some form of recovery measure when the pilot is faced with uncertainty that are due to wind and other uncertainties in the FF environment. In addition, reduction in fuel consumption is desirable, and to achieve near-optimal safe inter-aircraft separation while maintaining near-optimality under disturbances.

The rest of the paper is organized as follows. Section II presents the Free Flight concept and its associated rules and conditions and some previous work that was carried out will be described. Section III outlines a multi aircraft model used, its associated constraints and a subsection also outlines fuel efficient resolution maneuvers. Subsequent sections will outline a derivation for Trajectory redirection using NOC and it also shows that feasible trajectories exist when faced with uncertainties that the pilot may experience. Finally a simulation example with a worst case scenario will be shown. A summary (Pseudo Code) of the algorithm is also shown on Section V.

## II. The Free Flight Concept

The following definition was defined in the Report of the Radio Technical Commission for Aeronautics (RTCA) Board of Directors' Select Committee on Free Flight [2]. ”Free Flight” was defined as :

*”A safe and efficient flight operating capability under instrument flight rules (IFR) in which the operators have the freedom to select their path and speed in real time. Air traffic restrictions are only imposed to ensure separation, to preclude exceeding airport capacity, to prevent unauthorized flight through special use airspace, and to ensure safety of flight. Restrictions are limited in extent and duration to correct the identified problem. Any activity which removes restrictions represents a move towards free flight.”*

Sometimes the words ”Free Flight” are used for concepts, which include direct routing but no airborne separation. ATC will be overseeing traffic flow management (TFM) and intervening only when there is a loss of separation.

### A. Conditions and Rules of Free Flight

The FF definition used by RTCA roughly defines the concept. To define the context, a more detailed definition and rules are required. The following questions need to be clarified after applying the RTCA definition:

1) What is the role & responsibility of ATC?
2) How is an alert zone and protected zone defined?

Based on the questions above, the following choices and assumptions were made:

1) NO ATC: An inflexible proposition of FF is chosen with no ATC on the ground. The idea behind this concept is to explore the limits of the FF concept.

2) ALL AIRCRAFT FULLY EQUIPPED: All aircraft are assumed to be set up with Automatic Dependent Surveillance-Broadcast (ADS-B) transmitter & receiver and CDR advisory modules. The transmitter sends the aircraft's position and intent information needed by the CD module to all other aircraft. The ADS-B receiver collects all the information of the traffic within a certain range in the FF sector. The scenarios of mixed equipage is not considered here even though it is required for the transition to FF.

3) UPPER AIRSPACE ONLY: This work is limited to the uppermost airspace only, in order to concentrate on general conceptual problems. In other flight phases, a transition to controlled flight is foreseen and required. Where and how that transition should be implemented is not addressed in this work, since more benefit is attained when contemplating direct routing of aircraft(s) through FF in the uppermost airspace, so the feasibility for this airspace is a desirable result in itself.

4) PROTECTED ZONE: A conflict is defined in this work as an intrusion of a protected zone. The protected zone is defined using current ATC standards to be able to relate to existing traffic densities, even though studies have shown that the protected zone can be smaller. The protected zone must not be entered by any other aircraft. The task of the CD module is to predict an intrusion of the protected zone. This protected zone was chosen to reflect current ATC separation standards, which is 5 nautical mile radius and a height of 2000 feet ($-1000ft \leqslant altitude \leqslant +1000ft$). This means the ratio diameter to height is about 30 to 1.

The following section outlines the aircraft model & objective function formulation with its associated constraints.

## III. Multi Aircraft Model & Constraints Formulation

Consider $I$ aircraft flying within an area of interest, their continuous-time dynamics for level flight cruise is outlined in [9]. An aircraft in level flight can be modelled as a four state Point Mass Model (PMM). The states considered in this model are the following

- horizontal position $(X)$ and $(Y)$
- the true airspeed $(V)$
- flight path angle $(\psi)$

The control inputs considered for the model are the engine thrust $(T)$, and the bank angle $(\phi)$. Assuming a PMM, and that the applied forces acting on the aircraft will lead to the following equations of motion [9]:

$$\begin{bmatrix} \dot{X}_i \\ \dot{Y}_i \\ \dot{V}_i \\ \dot{\psi}_i \end{bmatrix} = \begin{bmatrix} V_i \cos(\psi_i) \\ V_i \sin(\psi_i) \\ \frac{-C_{D_i} S_i \rho}{2} \frac{V_i^2}{m_i} + \frac{1}{m_i} T_i \\ \frac{C_{L_i} S_i \rho V_i}{2m_i} \sin(\phi_i) \end{bmatrix} = f(t,x,u) \qquad (1)$$

for $i \in \Upsilon \triangleq \{1, ....., I\}$ and $S$ is the surface area of the wings, $\rho$ is the air density and $C_D$, $C_L$, are aerodynamic lift and drag coefficients. Since we are considering commercial aircraft, it is assume that the aircraft operates on a near trimmed flight conditions [9]. Eurocontrol's Base of Aircraft Database (BADA) [9] manual defines this as Total Energy Model (TEM) [9]. The TEM combines the rate of work done by all the forces that is affecting the aircraft to the rate of increase in its total energy (Potential and Kinetic).The aircraft motion has four states $(X, Y, V, \psi)$ and two inputs $(T_i, \phi_i)$. For a specific aircraft, the values of such bounds can be derived from the BADA database [9]. There are many models available in the literature when it involves predicting the current and future positions of an aircraft. For example in [8], an interpolating algorithm was developed to fit a vector field to the observed velocities at aircraft positions and at given sample times. Worst case techniques were used in [2] to estimate aircraft's position. Probabilistic modeling was covered and outlined in [1,3]. In this paper we are using a four state point mass model to estimate the aircraft's position.

### A. Velocity and acceleration constraints

Aerodynamics impose physical constraints on the minimum and maximum speeds an aircraft can fly at each altitude. In addition, passenger comfort and other factors impose constraints on the acceleration and the turning rate. These constraints can be expressed in the following form:

$$V_{max} \geqslant V_i(t) \geqslant V_{min} \qquad (2)$$

$$|\dot{V}_i(t)| \leqslant \delta V \qquad (3)$$

$$|\dot{\psi}_i(t)| \leqslant \delta\psi \qquad (4)$$

### B. Maneuvering constraints

Considering enroute traffic, the speed range of an aircraft is narrow. Passenger comfort is also taken into consideration which requires smooth trajectories, so in this case additional constraints are required for the maneuvers. The approach was proposed in [2], and was also used in [8] where constraints are imposed on speed changes that are bounded on a set around the current aircraft speed. The possible changes is the convex set of possible speed commands and can be described by a combination of quadratic and linear constraints as follows:

$$\|V_{0i} + \dot{V}_i\| \leq V_{max} \tag{5}$$

$$(V_{0i} + \dot{V}_i)^T \frac{V_{0i}}{\|V_{0i}\|} \geq V_{min} \tag{6}$$

Values of $V_{max}$ and $V_{min}$ can be obtained from the BADA database [15] but generally for most commercial aircraft cruising at high altitudes it can be approximated by:

$$\frac{V_{max} - V_{min}}{V_{max}} \leq 0.1 \tag{7}$$

### C. Conflict Avoidance Constraints & Overall Cost Function

Conflict avoidance constraints can be mathematically formalised in numerous ways. Many authors express collision avoidance constraints in terms of a given minimum miss distance, and it appears to be the most attractive option from a geometrical point of view, In this work, a time-based separation criterion is considered. All aircraft should remain separated at all times, by at least a minimum distance, which is set to $d_s = 5nm$ for cruising altitudes. This constraint can be described as follows:

$$\left\| [X_i, Y_i]^T - [X_j, Y_j]^T \right\|_2 \geqslant d_s \tag{8}$$

for all times $t > 0$ and for all aircraft pairs $i \neq j$ , where $(i, j) \in \Upsilon$. The constraint described by (8) is not convex and, in order to be able to handle it computationally, it requires tightening. One approach would be using the norm inequality $\|\bullet\|_2 \geq \|\bullet\|_\infty$ where $\|\bullet\|$ is the euclidean norm, then the so-called "*big-M*" technique (Refer to the formulation in [1]) can be used. This will ensure that at least one of the inequality constraints is active, and consequently that the two aircraft are separated by the required distance along at least one of the axes. The problem with this approach is that a model mismatch may be prevalent and there is a possibility that aircraft will encounter a situation where a solution satisfying both the dynamics and the separation constraints may no longer exist.

So, in this case an alternative separation constraint which was adopted in [8] is considered. For notational convenience, let's define $P_i = [X_i, Y_i]^T$, $P_j = [X_j, Y_j]^T$ which denotes the aircraft position and $P_{ij} = P_i - P_j$ is the relative position. $V_{ij} = V_i - V_j$ is the relative speed and $\dot{V}_{ij} = \dot{V}_i - \dot{V}_j$ is the relative velocity $P_{i,0}$ and $V_{i,0}$ are the initial position and initial speed of aircraft $i$ respectively. the constraint can be described as:

$$P_{0ij}^T (V_{0ij} + \dot{V}_{ij}) + \|V_{0ij} + \dot{V}_{ij}\| \sqrt{\|P_{0ij}\|^2 - d_s^2} \geq 0 \tag{9}$$

Non-convexity is also prevalent in the above formulation, so a new slack variable is introduced $\varphi_{ij}$ where $\|V_{0ij}\|^2 \geq \varphi_{ij}$ and $\varphi_{ij} \geq 0$ which leads to

$$P_{0ij}^T (V_{0ij} + \dot{V}_{ij}) + \varphi_{ij} \sqrt{\|P_{0ij}\|^2 - d_s^2} \geq 0 \tag{10}$$

It can be seen that the above constraints and cost function form a non-convex quadratic program of the form

$$\begin{array}{ll} \min & h^T P_0 h + 2\xi_0^T h + r_0 \\ s.t & h^T P_i h + 2\xi_i^T h + r_i \leq 0 \end{array} \tag{11}$$

Although the above quadratic program can be very difficult to solve in general, it has received a fair share of research attention. Approximate solutions can be obtained based on convex optimization. We also have $H = [\dot{V}_1, \dot{V}_2, ....., \dot{V}_n, \varphi_{12}, ......, \varphi_{(n-1),n}]^T$. If the optimal solution $H$ to the semi-definite relaxation has unit rank, then $\overline{h}$ is the solution. If not, then a form of randomization procedure must be applied. If we consider a gaussian distribution with mean $\overline{h}$ and $cov(H - \overline{h}\,\overline{h}^T)$ then samples can be chosen in accordance to the distribution. As outlined in [8] a linearisation procedure is required to pick the crossing pattern for each aircraft pair by computing the following

$$C_p = sgn(P_{0ij} \times (V_{0ij} + \dot{V}_{ij})) = \frac{P_{0ij} \times (V_{0ij} + \dot{V}_{ij})}{|P_{0ij} \times (V_{0ij} + \dot{V}_{ij})|} \tag{12}$$

If $C_p = 1$ the crossing pattern is then chosen to be counter-clockwise and clockwise if $C_p = -1$. We will assume that the crossing pattern is clockwise in the very unlikely case when $C = 0$. Subsequently, the corresponding convex optimization problem can be solved using barrier methods.

### D. Fuel Efficient Resolution Maneuvers

In this section, a formulation for fuel resolution maneuvers outlined in [10] is considered. The strategic encounter is concerned with minimizing the economics of a CR. Economics defined in this context by considering direct operating cost (DOC), which comprises of fuel and flight time costs involved in maneuvering to avoid a conflict. Only trajectory solutions that geometrically pass around the Protected Airspace Zone of an intruder aircraft is considered. In the FF context, it is assumed that the heading maneuver will consist of a series of standard vectoring heading, and speed changes. in such a way that straight line motion materialises in between heading, and speed. The DOC penalty function used incorporates both fuel and time elements. Included in the DOC are

1) The additional fuel required due to the increased drag and flight path distance traveled during a maneuver [10].
2) The additional operating costs due to the additional time required to execute the maneuver and return back to course [10].

$$DOC = C_{Fuel}\Delta W_{Fuel} + C_{Time}\Delta T \quad (13)$$

where $C_{Fuel}$ is the cost of fuel, $\Delta W_{Fuel}$ is the additional fuel used in the maneuver, $C_{Time}$ is the time dependent aircraft operating cost and $\Delta T$ is the additional time used in the maneuver. This DOC is investigated by [10], where $C_{Fuel} = £0.10/lb$ and $C_{Time} = £15.22/min$. Another issue that needs to be considered is the additional DOC incurred by neighbouring traffic that may be affected by CR maneuvers; however, it's not considered here. The fuel burn equation is:

$$\frac{dW}{dT} = T_R C_j \quad (14)$$

where $W$ is the weight, $C_j$ is the specific fuel consumption, and $T_R$ is the required thrust. The required thrust can be described as:

$$T_R = D + sin(\psi_{ss}) \quad (15)$$

where $D$ is the drag and $\psi_{ss}$ is the steady state flight path angle. The fuel consumption model varies linearly with airspeed. The lift coefficient $C_L$, is described in terms of the aircraft weight $W$ and speed $V$ as follows:

$$C_L = \frac{2Wcos(\psi_{ss})}{S_w V^2 \rho} \quad (16)$$

where $\rho$ is air density, and $S_w$ is the reference wing area. The drag is described as:

$$D = \frac{C_D V^2 \rho W S_w}{S_w V^2 \rho} \quad (17)$$

$C_D$ is the drag coefficient which is approximated by:

$$C_D = C_{D0} + K^2_{CL} \quad (18)$$

$C_{D0}$ is the zero lift drag coefficient, $K = \frac{1}{ARe\pi}$ is the induced drag factor, $e$ is Oswald's efficiency factor, and $AR$ is the wing aspect ratio. Furthermore, the thrust specific fuel consumption of a turbojet or turbofan engine increases nearly linearly with Mach number [10]. For this reason, the fuel consumption can be modeled as:

$$C_j = \sigma V \quad (19)$$

where $\sigma$ is the coefficient. Next, these endurance equations can be combined to obtain the fuel burn:

$$\frac{dW}{dT} = -\left[0.5\rho V^2 S_w (C_{D0} + K\left(\frac{2Wcos(\psi_{ss})}{\rho V^2 S_w}\right)^2 + \quad (20) \right.$$
$$\left. Wsin(\psi ss)\right]\sigma V$$

Equation (20) is integrated to calculate the fuel burn, given the geometry the previously defined heading maneuvers or speed control maneuvers. For the acceleration case, the equation is integrated numerically [10].

## IV. Trajectory redirection using neighboring optimal control (NOC)

Solving a nonlinear optimal control problem with a long horizon is generally computationally expensive, especially when FF concepts are considered, since there is limited computation power onboard an aircraft. If a nominal optimal solution is known, it is advantageous to approximate the optimal control solution when the parameters of the optimal control problem are slightly perturbed. The Neighbouring optimal control (NOC) method provides a first order approximation to the optimal solution corresponding to an initial state perturbed from the nominal value. When a disturbance significantly affect the forces on an aircraft, and modifies the flight path, redirecting the trajectory may recover the aircraft. Disturbances considered here may correspond to the wind or cross track errors, so in this case the nominal states and controls can be used to redirect the trajectory [11]. An off-line optimal trajectory is always assumed to be found from a feasible trajectory database. When a failure occurs, NOC is used to determine its neighbouring feasible trajectory on-line. If such a neighboring feasible trajectory exists, the aircraft may then be recovered from that failure in real time [11]. The formulation outlined in [11,4] is outlined in this section, and they are applicable to the formulation of FF concepts.

### A. Off-line trajectory optimization for the Shortest Path Evaluation

A shortest path problem can be used for the off-line optimal trajectory search. To find the shortest path, several algorithms can be used. In the free flight concept, Dijkstra's algorithm [5] is a reasonable choice since it guarantees to find an optimal path by repeatedly selecting a vertex with the minimum shortest-path estimate. The weight is defined by the mean value of the angle of attack of the aircraft. The reader is reffered to [5] for further information.

### B. On-line trajectory optimization using NOC

Based on the off-line optimal trajectory, a small perturbation is imposed on each state variable and new control can be obtained by NOC. The aircraft can be recovered from the disturbances. The NOC approach was outlined by Bryson and Ho (1975) [4], Consider the following nonlinear system

$$\dot{x}(t) = f(t, x, u), x(t_0) = x_0 \quad (21)$$

where $x(t)$ are the state variables and $u(t)$ control inputs and the cost function is defined as

$$\min \qquad J(u) = \phi(t_f, x_f, u_f) \quad (22)$$

s.t. Dynamical, Separation constraints, Velocity and acceleration constraints.

"$f$" denotes the final time. An assumption is made that the optimal solution $(x^*(t), u^*(t))$ can be determined, in which the optimal control is $u^*(t) = g(t, x^*(t), \lambda(t))$ under the following constraints

$$x_{min} \leq x^*(t) \leq x_{max} \ , \ u_{min} \leq u^*(t) \leq u_{max} \quad (23)$$

when small perturbations $\delta x(t_0)$ and small variations $\delta\lambda$ imposed on $x_0$ and $\lambda$, then the neighbouring feasible trajectory $z(t)$ and new optimal control $\nu(t)$ can be found by $\delta x(t_0, \delta\lambda)$ and $\dot{x}(t)$. In order to ensure that a neighbouring trajectory exists the following assumptions are made:

- **Assumption 1.** For the nonlinear system (22), $f(t, x, u)$ is continuous.
- **Assumption 2.** The open loop optimal solution $x^*(t)$ and optimal control $u^*(t)$ is well defined for system (22).

For a neighbouring feasible initial condition $z(t_0)$ which will be equal to $x_0$, there will be a neighbouring feasible trajectory $z(t)$ and neighbouring optimal control $\nu(t)$ and satisfies

$$\sup\|z(t) - x^*(t)\| \leq K \tag{24}$$

Where $K$ is an upperbound. And when system parameters varies from$\lambda_1$ to $\lambda_2$ and $\lambda_2 = \lambda_1 + \delta\lambda$ then the neighbouring feasible state trajectory $z(t)$ and neighbouring optimal control $\nu(t)$ is:

$$\nu(t) = u^*(t) + G(t)(Z(t) - X{*}(t)) \qquad or \tag{25}$$

$$\nu(t) = u^*(t) + G(t)\Delta X(t), \tag{26}$$

Where $Z(t) = [z(t), \lambda_2]^T$ , $X^*(t) = [x^*(t), \lambda_1]$, and $\Delta X(t)$ is a deviation vector of the new state vector $Z(t)$ from the open loop nominal state trajectory $X^*(t)$, $G(t)$ is the feedback gain matrix which is given by

$$G(t) = \delta U(t)\delta\overline{X}^{-1}(t) \tag{27}$$

Where $\delta U(t)$ are the control errors caused by the perturbation on the state vector $X(t)$ and $\delta\overline{X}(t)$ is a perturbation matrix in the initial states $x_{t_0}$ and initial system parameter $\lambda_1$. Since the control law can be put in the form of $u = g(t, x, \lambda)$ then system (22) becomes an unforced system $\dot{x} = \overline{f}(t, x, \lambda), x(t_0) = x_0$, under the above conditions the controller can satisfy the constraints in. In order to ensure that that initial value problem does exist then theorem 3.5 outlined in [6] proves that the conditions are satisfied then in this case perturbations can be imposed.

In the neighbourhood of $x^*$, the state vector $z$ can be defined as $z = x^* + \delta x$. It was mentioned above that the parameters changed from $\lambda_1$ to $\lambda_2$ and $\lambda_2 = \lambda_1 + \delta\lambda$. Carrying out a Taylor series expansion for $\nu^*$ is the following

$$\nu^* = g(t, z, \lambda_2) = g(t, x^* + \delta x, \lambda_1 + \delta\lambda) \tag{28}$$

$$= g(t, x^*, \lambda_1) + \frac{\partial g}{\partial x}\partial x + \frac{\partial g}{\partial \lambda}\partial\lambda + H.O.T \tag{29}$$

$$= u^* + \frac{\partial g}{\partial x}(z - x^*) + \frac{\partial g}{\partial \lambda}(\lambda_2 - \lambda_1) + H.O.T \tag{30}$$

$$= u^* + \begin{bmatrix} \frac{\partial g}{\partial x} & \frac{\partial g}{\partial \lambda} \end{bmatrix} \left( \begin{bmatrix} z \\ \lambda_2 \end{bmatrix} - \begin{bmatrix} x^* \\ \lambda_1 \end{bmatrix} \right) + H.O.T \tag{31}$$

H.O.T denotes higher order terms. If the perturbation $\delta x$ is small enough or $z$ is in the neighbourhood of $x^*$ then the higher order terms of $\delta x$ and $\delta y$ must be small as well. Therefore the higher order terms can be neglected at this point. The new controller now is $\nu = u^* + G(Z - X^*)$, where $Z \in \mathbb{R}^{m\times(n+p)}$ and $X^* \in \mathbb{R}^{n+p}$. If closed loop system was to be considered, the perturbations $\delta x$ and $\delta\lambda$ will be the measured deviations, in this case the gain matrix is defined by:

$$G = \begin{bmatrix} \frac{\partial g}{\partial x} & \frac{\partial g}{\partial \lambda} \end{bmatrix} = \frac{\partial g}{\partial X} \in \mathbb{R}^{m\times(n+p)} \tag{32}$$

The gain matrix $G$ can be calculated numerically at every time $t \in [t_0, t_f]$. The reason for this being is that no partial derivatives will be evaluated. Now from the $u = g(t, x, \lambda)$ and $X = [x^T, \lambda^T]^T$, the differential of $u$ is $du = \frac{\partial g}{\partial X}dX + \frac{\partial g}{\partial t}dt$. For a known perturbation, the perturbation vectors are $\delta X = X_{perturbed} - X^*$ and $\delta u = u_{perturbed} - u^*$. If a linear approximation is used and also noting that $\delta t = 0$ for all $t \in [t_0, t_f]$, $\delta u$ can be obtained by

$$\delta u = \frac{\partial g}{\partial X}\delta X + \frac{\partial g}{\partial t}\delta t = G\delta X \tag{33}$$

Next an $(n + p) \times (n + p)$ square matrix is constructed for $\delta\overline{X}$, and another for $\delta U$ which is an $m \times (n + p)$:

$$\delta\overline{X} = [\delta X^1 .........\delta X^{n+p}] \in \mathbb{R}^{(n+p)\times(n+p)} \tag{34}$$

$$\delta U = [\delta u^1 .........\delta u^{n+p}] \in \mathbb{R}^{m\times(n+p)} \tag{35}$$

Where $\delta X^i$, $i = 1, 2, ....., n + p$ and $\delta u^i$, $i = 1, 2, ....., n + p$ are the $ith$ perturbation of $\delta X$ and $\delta u$ then equation () can be expressed by the following:

$$\delta U = G\delta\overline{X} \tag{36}$$

If the perturbations are chosen to be linearly independent and at the same time small enough, then $\delta\overline{X}$ can be guaranteed to be invertible at every instance of time.

## V. NOC Application to the Aircraft trajectory retargeting under conflict

It was shown above that a neighbouring feasible trajectory exists when applying NOC to an aircraft [11,4]. Since we have four states $(X, Y, V, \psi)$ and two control variables $(T, \phi)$. If $C_L$ and $C_D$ are the parameters that vary then a neighbouring solution exists. As mentioned above that a nominal optimal solution exists using Dijikstra's algorithm [5] then a new controller is calculated $\Psi^*$ then the perturbed controller is calculated as $\Psi_{perturbed} = \Psi^* + G(Z - X^*)$. Since it is required to enforce the aircraft to meet a final time, then the state $X$ is chosen as an independent variable. So five perturbations are used which are $V, \psi, Y, C_L, C_D$. Define $\delta\overline{X} = [V^1 - V^* - \psi^1 - \psi^* x^1 - x^* C_{L1}^1 - C_{L1}^* C_{D1}^1 - C_{D1}^*]^T$............$[V^5 - V^*, \psi^5 - \psi^* x^5 - x^* C_{L1}^5 - C_{L1}^* C_{D1}^5 - C_{D1}^*]^T$ and $\delta\alpha = [\Psi^1 - \Psi^*, ...., \Psi^5 - \Psi^*]^T$ and $n = 1, ..., 5$ which is the perturbation instances for each state and parameter. Then the gain matrix can be evaluated as $G = \delta\Psi\delta\overline{X}^{-1}$. The algorithm can be summarised below:

**Algorithm** Free Flight Concept Exploiting NOC

```
Require: $P_{i0}(t)$ and $P_i(t)$ for all $i \in 1, ..., I$
Require: $T_i$ and $\phi_i$ for the nominal shortest path
While $\exists i \; P_{0ij}^T(V_{0ij} + \dot{V}_{ij}) + \|V_{0ij} + \dot{V}_{ij}\| \sqrt{\|P_{0ij}\|^2 - d_s^2} \geq 0$ do
    Solve the Optimisation problem (11)
    Compute the matrix Gain $\delta\overline{X}$ and $\delta U$ from (34),(35)
    Compute the matrix Gain $G(t)$ from (36)
    Measure new aircraft position $p_i(t)$
End While
```

## VI. Simulation

A conflict situation with five aircraft are initially located on a circle of 400 km radius and are heading to mid air collision. Without any corrective action, all aircraft would collide at the center. Although unrealistic, this scenario allowed us to test the effectiveness of our method. All conflict constraints are enforced, and fuel consumption maneuvers are computed.

In order to differentiate between the nominal solution and the perturbed on,the results are shown as a box-and-whisker diagram in Figure 1 & 2. Qualitatively, both solutions are very similar, except for the fuel consumption of the aircraft which is outlined in Fig. 2, as the average fuel consumption in the perturbed case is lower than the nominal case which is an excellent demonstration of the ability of NOC to achieve near-optimal results without resorting the need to recompute the optimal solution. On the other hand, in both the nominal and perturbed case one conflict occurred which was not detected but resolved by the algorithm. However, there is a clear disadvantage that introducing any additional constraint in this formulation may impose restrictions when implementation issues are considered since computational times becomes higher, and intractability is also apparent sometimes in the simplest of optimisation problems. While the fuel savings is not very significant, the important point is that a near-optimal conflict resolution that maintains separation has been performed without increasing fuel use.

## VII. Conclusion

A Free Flight formulation was presented to deal with conflict detection and resolution. A point mass model for aircraft dynamics was used that allowed a quadratic program formulation for the problem. A new approach based on neighbouring Optimal Control (NOC) that can deliver an aircraft to its designated position safely when faced with conflicts and uncertainties was outlined. A simulation example was provided that illustrated the effectiveness of our approach. The benefits of the algorithm are that it reduces fuel consumption, and it achieves optimal safe inter-aircraft separation while maintaining near-optimality under disturbances. Ongoing research focuses on combining this algorithm with complexity metrics that defines the level of disorder are being explored.

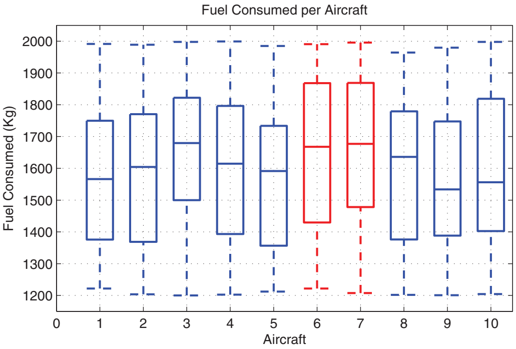


Fig. 1. Fuel Consumption. Blue: No conflict and Red: Conflicting Aircraft

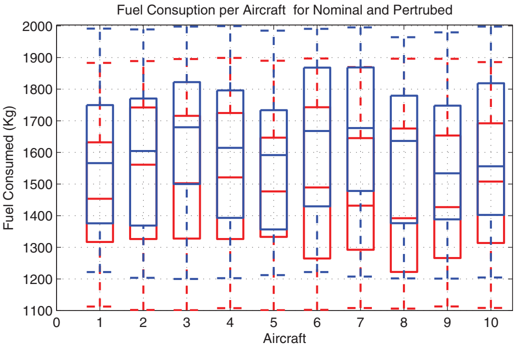


Fig. 2. Fuel Consumption. Blue: Nominal (Dijkstra Algorithm) and Red: Perturbed (NOC)

## References

[1] Prandini, M., Hu, J., Lygeros, J., Sastry, S., 2000. A probabilistic approach to aircraft conflict detection. IEEE Transactions on Intelligent Transportation Systems

[2] Eby, M., Kelly, W., 1999. Free Flight separation assurance using distributed algorithms. Proceedings of the IEEE Aerospace Conference, vol. 5. Snowmass, CO, pp. 429441.

[3] Erzberger, H., Paielli, R., 1997. Conflict probability estimation for Free Flight. AIAA Journal of Guidance, Control, and Dynamics 20 (3).

[4] Bryson, A. E., Jr., & Ho, Y. C. (1975). Applied optimal control: Optimization, estimation, and control. New York: Hemisphere.

[5] Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. Numerische Mathematik, 1, 269-271.

[6] Khalil, H. K. (2002). Nonlinear systems (3rd ed.) (p. 97). Upper Saddle River, NJ: Prentice Hall

[7] W. Glover and J. Lygeros, A stochastic hybrid model for air traffic control simulationin Hybrid Systems: Computation and Control.Springer Verlag, 2004, no. 2993, pp. 372386.

[8] E. Frazzoli, Z. Mao, J. Oh, and E. Feron, Aircraft conflict resolution via semi-definite programming, AIAA Journal of Guidance, Control, and Dynamics, vol. 24, no. 1, pp.79 86, 2001.

[9] Eurocontrol Experimental Centre, User Manual for the Base of Aircraft Data (BADA), 2004. [Online]. Available: http://www.eurocontrol.fr/projects/bada/

[10] Krozel, J. and Peters, M. (1997). Strategic Conflict Detection and Resolution for Free Flight, Proceedings of the Conference on Decision and Control, San Diego, USA.

[11] Z. Jiang and R. Ordnez, On-Line Robust Trajectory Generation on Approach And Landing for Reusable Launch Vehicles, Automatica, vol. 45, no. 7, Jul. 2009

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Path following for small UAVs in the presence of wind disturbance

Cunjia Liu* Owen McAree† and Wen-Hua Chen‡
Department of Aeronautical and Automotive Engineering
Loughborough University, Lecistershire, UK. LE11 3TU
Email: *c.liu5@lboro.ac.uk †o.mcaree@lboro.ac.uk ‡w.chen@lboro.ac.uk

*Abstract*—**This paper presents an alternative approach of designing a guidance controller for a small Unmanned Aerial Vehicle (UAV) to achieve path following in the presence of wind disturbances. The wind effects acting on the UAV are estimated by a nonlinear disturbance observer. Then the wind information is incorporated into the nominal path following controller to formulate a composite controller so as to compensate wind influences. The globally asymptotic stability of the composite controller is illustrated through theoretical analysis and its performance is evaluated by various simulations including the software-in-the-loop. Initial flight tests using a small aircraft are carried out to demonstrate its actual performance.**

## I. Introduction

The application of Unmanned Aerial Vehicles (UAVs) has been found in various areas not limited to military operations, but also in civil areas such as aerial photography and precision agriculture. Most of the UAV operations essentially are composed of commanding UAVs to fly through a series of spatial locations or paths either with or without a temporal requirement. This requirement categories flight patterns into two types, namely the trajectory tracking and path following [1]. The former suggests that the UAV needs to be in a particular position at a prespecified time, whereas the latter requires the UAV to converge to a geometric path with any feasible speed profile. In this paper, the path following problem is considered because it is less likely to push UAVs to their performance limits [2]. Moreover, with the influences of wind disturbance the trajectory tracking ability of an UAV can be easily compromised, which leaves the path following as an effective way to execute a task.

Path following as one of the motion control problems has been extensively studied especially for wheeled robots. In the field of UAV application, although vehicle dynamics are more complicated, the most recent micro aircraft are equipped with autopilots that provide the inner-loop stabilisation. For example, the UAV used in this study is installed with an Ardupilot autopilot which implements three PID controllers to achieve altitude-hold and airspeed-hold using elevator and throttle and the coordinated turn using aileron and rudder, which endows the UAV the ability to track the heading rate demand from the guidance controller. This means that the out-loop behaviour of the UAV in level flight can be abstracted at a kinematic level by using a unicycle model. To this end, this paper adopts a path following method stemmed from wheeled robots [3], which has since been extended to 3-D case for UAV applications [4]. But the main purpose of this study is to improve path following accuracy under wind conditions.

In terms of the micro fixed-wing aircraft considered in this paper, their light structure and limited power allows wind disturbances to have a strong effect on them. A common strategy to eliminate the influence of wind on path following is to overlook the airspeed of an UAV and focus on the its ground track [5], [6]. In this case, the ground velocity and flight course are required in the feedback signals. This information can be calculated from GPS position by differentiation or more elaborately can be provided by an onboard inertial navigation system (INS). However, for a micro UAV equipped with a low-cost sensor suit, these flight data may be not of good quality, whereas the fast dynamics of small vehicles are highly susceptible to the low rate and delay of the GPS feedback signal. It is therefore more convenient to use the smooth airspeed measurement, magnetic heading and original GPS position data to realise the guidance function.

Another approach to solve wind effects on an UAV is to explicitly take them into account in path planning or control algorithms [7]–[9]. The knowledge of wind conditions is therefore required in such applications. Following on this direction, this paper adopts an alternative approach that exploits the use of a nonlinear disturbance observer [10]. The disturbance observer is designed to provide the estimates of external wind, which are then incorporated into the nominal path following controller. This results in a composite controller for UAV path following. Note that the wind disturbances are assumed to be near constant in the analysis because only these components cause a steady state error. However, the ability of estimating the varying wind are also demonstrated in the simulation section. In addition, the disturbance observer can provide estimates of disturbances other than wind gust such as aircraft trimming errors, which cause the system to behave differently from the nominal model [11].

The composite controller comprising the disturbance observer and the nonlinear guidance law is shown to be globally asymptotically stable by using the control theory on cascaded systems [12]. This suggests that the proposed control system can guarantee the path following accuracy in spite of wind disturbances. The performance of the proposed controller is demonstrated in the simulations as well as in real flight experiments using our newly developed flight test platform.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

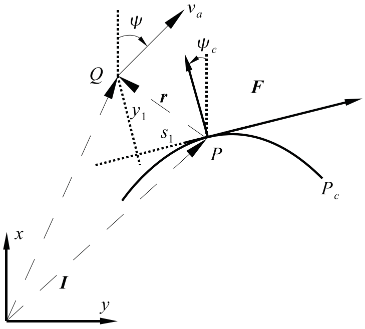


Fig. 1. Frame configuration

## II. Problem formulation

The problem considered in this paper is the accurate path following for UAVs in the presence of wind. The objective of path following is to guide the UAV to converge to a desired geometric path described by some parameters, e.g. the path length. When wind disturbances are introduced to the system, necessary actions need to be taken to prevent their effects on the path following performance.

The kinematics of a fixed-wing UAV can be described using the following unicycle model under the assumptions that the UAV is in level flight with a near constant air speed:

$$\begin{aligned} \dot{x} &= v_a\cos(\psi) + w_x \\ \dot{y} &= v_a\sin(\psi) + w_y \\ \dot{\psi} &= \omega \end{aligned} \tag{1}$$

where $(x, y)$ is the position of the UAV in the inertial frame $\mathcal{I}$, $\psi$ is the UAV heading angle, $v_a$ is the airspeed, $(w_x, w_y)$ is the velocity of the wind disturbance in $x$ and $y$ directions, respectively, and $\omega$ is the heading rate. By constructing a control input $\omega$, the position output should be able to follow a prescribed path $P_c(s) = [\ x_c(s) \quad y_c(s)\ ]^T$, which is a spatial curve and parameterised by its length $s$.

The path following function can be achieved by controlling the UAV to follow a virtual target running along the target path [3], [4]. To this end, it is natural to present the generalised error between the UAV and the virtual target in a moving frame attached to this virtual target. In the inertial frame $\mathcal{I}$, let $Q$ be the UAV position and $P$ denote the point on the path $P_c(s)$ to be followed, i.e. the virtual target. A Serret-Frenet frame $\mathcal{F}$ can be established on the point $P$ as shown in Fig.1, where the signed curvilinear abscissa is the path length $s$. Then, the UAV position $Q$ can be expressed in the frame $\mathcal{I}$ as a vector $\boldsymbol{q} = [\ x \quad y\ ]^T$ and in the frame $\mathcal{F}$ as a vector $\boldsymbol{r} = [\ s_1 \quad y_1\ ]^T$. Note that $\boldsymbol{r} = \boldsymbol{q} - \boldsymbol{p}$ is also the error vector, where $\boldsymbol{p}$ denotes the position vector of $P$ in $\mathcal{I}$.

To minimise the error vector $\boldsymbol{r}$, its dynamics in frame $\mathcal{F}$ need to be investigated. First, by defining the heading rate of the desired path $\dot{\psi}_c = \omega_c$, we have the following relations:

$$\begin{aligned} \dot{\psi}_c &= \omega_c = c_c(s)\dot{s} \\ \dot{c}_c(s) &= g_c(s)\dot{s} \end{aligned} \tag{2}$$

where $c_c(s)$ and $g_c(s) = \frac{dc_c(s)}{ds}$ are the path curvature and its derivative, respectively. Next, the velocity of $P$ in $\mathcal{I}$ can be expressed in frame $\mathcal{F}$, such that:

$$\left(\frac{d\boldsymbol{p}}{dt}\right)_F = \begin{bmatrix}\dot{s} & 0\end{bmatrix}^T \tag{3}$$

For the position $\boldsymbol{q}$ of the UAV, its inertial velocity can be expressed in a moving frame such that:

$$\left(\frac{d\boldsymbol{q}}{dt}\right)_I = \left(\frac{d\boldsymbol{p}}{dt}\right)_I + R_F^I\left(\frac{d\boldsymbol{r}}{dt}\right)_F + R_F^I(\omega_c \times \boldsymbol{r}) \tag{4}$$

where × denotes the vector cross-product and $R_F^I$ is the rotation matrix from frame $\mathcal{F}$ to $\mathcal{I}$. Left multiplying its inverse $R_I^F$ on both side of (4) yields

$$R_I^F\left(\frac{d\boldsymbol{q}}{dt}\right)_I = \left(\frac{d\boldsymbol{p}}{dt}\right)_F + \left(\frac{d\boldsymbol{r}}{dt}\right)_F + \omega_c \times \boldsymbol{r} \tag{5}$$

Using the relations

$$\begin{aligned} \left(\frac{d\boldsymbol{q}}{dt}\right)_I &= \begin{bmatrix}\dot{x} & \dot{y}\end{bmatrix}^T \\ \left(\frac{d\boldsymbol{r}}{dt}\right)_F &= \begin{bmatrix}\dot{s}_1 & \dot{y}_1\end{bmatrix}^T \end{aligned} \tag{6}$$

and the expansion of $\omega_c \times \boldsymbol{r}$, (5) can be rewritten as

$$R_I^F\begin{bmatrix}\dot{x} \\ \dot{y}\end{bmatrix} = \begin{bmatrix}\dot{s}(1 - c_c(s)y_1) + \dot{s}_1 \\ \dot{y}_1 + c_c(s)\dot{s}s_1\end{bmatrix} \tag{7}$$

Solving (7) for $\dot{s}_1$ and $\dot{y}_1$ and combining (1) yields the dynamics of the path following error in the frame $\mathcal{F}$ at the kinematics level:

$$\begin{aligned} \dot{s}_1 &= -\dot{s}(1 - c_c y_1) + v_a\cos(\psi_e) + w_{f_x} \\ \dot{y}_1 &= -c_c\dot{s}s_1 + v_a\sin(\psi_e) + w_{f_y} \\ \dot{\psi}_e &= \omega - c_c\dot{s} \end{aligned} \tag{8}$$

where $\psi_e = \psi - \psi_c$ is the heading error,

$$\begin{aligned} w_{f_x} &= \cos\psi_c w_x + \sin\psi_c w_y \\ w_{f_y} &= -\sin\psi_c w_x + \cos\psi_c w_y \end{aligned} \tag{9}$$

are wind disturbances expressed in the frame $\mathcal{F}$, respectively. The designed path following controller needs to regulate this system to eliminate the state error under the wind disturbances.

It is intuitive to work out that if the wind elements are known, the aircraft can fly into wind with certain trim angle so that the projection of forward speed normal to the track can be used to cancel the wind effect. This basic idea provides the guideline for the designing our path following controller.

## III. Disturbance observer based control

To obtain the estimates of wind disturbances, a nonlinear disturbance observer is first designed. This information is then incorporated into controller design. This design methodology is known as the disturbance observer based control (DOBC) [10].

### A. Disturbance observer design

The UAV kinematic model (1) can be cast into a compact mode:

$$\dot{\boldsymbol{x}} = f(\boldsymbol{x}) + g_1(\boldsymbol{x})\boldsymbol{u} + g_2(\boldsymbol{x})\boldsymbol{d} \tag{10}$$

where state $\boldsymbol{x} = [\ x \quad y \quad \psi\ ]^T$, control input $\boldsymbol{u} = \omega$ and disturbance $\boldsymbol{d} = [\ w_x \quad w_y\ ]^T$. The system functions $f(\boldsymbol{x})$, $g_1(\boldsymbol{x})$ and $g_2(\boldsymbol{x})$ are derived from (1), such that:

$$f(\boldsymbol{x}) = \begin{bmatrix} v_a \sin\psi \\ v_a \cos\psi \\ 0 \end{bmatrix}, \quad g_1(\boldsymbol{x}) = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad g_2(\boldsymbol{x}) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \tag{11}$$

A disturbance observer is adopted to estimate $w_x$ and $w_y$ under the assumption that they are near constant, i.e. $\dot{w}_x \approx 0$ and $\dot{w}_y \approx 0$. The disturbance observer follows the standard design [10]:

$$\begin{aligned} \hat{\boldsymbol{d}} &= \boldsymbol{z} + p(\boldsymbol{x}) \\ \dot{\boldsymbol{z}} &= -l(\boldsymbol{x})g_2(\boldsymbol{x})\boldsymbol{z} - l(\boldsymbol{x})(g_2(\boldsymbol{x})p(\boldsymbol{x}) + f(\boldsymbol{x}) + g_1(\boldsymbol{x})\boldsymbol{u}) \end{aligned} \tag{12}$$

where $\hat{\boldsymbol{d}} = [\ \hat{w}_x \quad \hat{w}_y\ ]^T$ is the estimate of wind disturbance, $\boldsymbol{z}$ is the internal state of the nonlinear observer, $p(\boldsymbol{x})$ is a nonlinear function to be designed, and $l(\boldsymbol{x})$ is the nonlinear observer gain given by

$$l(\boldsymbol{x}) = \frac{\partial p(\boldsymbol{x})}{\partial \boldsymbol{x}} \tag{13}$$

The estimation error in the above observer is defined as $\boldsymbol{e}_d = \boldsymbol{d} - \hat{\boldsymbol{d}} = [\ e_x \quad e_y\ ]^T$. Under the assumption that the disturbance is slowly varying compared to the observer dynamics and by combining (12)-(13) and the system function (10), it can be shown that the estimation error has the following property:

$$\dot{\boldsymbol{e}}_d = \dot{\boldsymbol{d}} - \dot{\hat{\boldsymbol{d}}} = -\dot{\boldsymbol{z}} - \frac{\partial p(\boldsymbol{x})}{\partial \boldsymbol{x}}\dot{\boldsymbol{x}} = -l(\boldsymbol{x})g_2(\boldsymbol{x})\boldsymbol{e}_d \tag{14}$$

Therefore, the observer design problem is converted to chose an appropriate observer gain $l(\boldsymbol{x})$ such that (14) is globally exponentially stable regardless of state $\boldsymbol{x}$.

In this paper, since the function $g_2$ is a constant matrix, the observer gain can be simply chosen as

$$l(\boldsymbol{x}) = \begin{bmatrix} l_x & 0 & 0 \\ 0 & l_y & 0 \end{bmatrix} \tag{15}$$

where $l_x$ and $l_y$ are positive constants to be tuned. Correspondingly, the nonlinear function $p(\boldsymbol{x})$ can be calculated by integrating $l(\boldsymbol{x})$ with respect to $\boldsymbol{x}$ based on (13).

### B. Control synthesis

After the wind estimates are obtained, the next step is to design a path following controller that regulates the error system (8) to its origin. According to the design guideline, the UAV can fly into wind with a trimming angle $\psi_0$ so that the side component of the forward velocity neutralises the wind component that drives the UAV away from the desired path.

Given that the wind estimates are available, the estimated trimming angle can be written as $\psi_{\hat{0}} = -\sin^{-1}\frac{\hat{w}_{f_y}}{v_a}$. Furthermore, the error dynamics (8) can be reformulated to facilitate the control design:

$$\begin{aligned} \dot{s}_1 &= -\dot{s}(1 - c_c y_1) + v_a\cos(\psi_{\hat{w}} + \psi_{\hat{0}}) + \hat{w}_{f_x} + e_{f_x} \\ \dot{y}_1 &= -c_c\dot{s}s_1 + v_a\sin(\psi_{\hat{w}} + \psi_{\hat{0}}) + \hat{w}_{f_y} + e_{f_y} \\ \dot{\psi}_{\hat{w}} &= \omega - c_c\dot{s} - \dot{\psi}_{\hat{0}} \end{aligned} \tag{16}$$

where $\psi_{\hat{w}} = \psi_e - \psi_{\hat{0}}$, $\hat{w}_{f_x}$ and $\hat{w}_{f_y}$ are calculated from (9) after $\hat{w}_x$ and $\hat{w}_y$ are estimated in (12), such that $\hat{w}_{f_x} + e_{f_x} = w_{f_x}$, $\hat{w}_{f_y} + e_{f_y} = w_{f_y}$. Carrying out the differentiation of $\psi_{\hat{0}}$ with respect to time gives:

$$\begin{aligned} \dot{\psi}_{\hat{0}} =& \frac{1}{\sqrt{v_a^2 - \hat{w}_{f_y}^2}} \underbrace{(\cos\psi_c\hat{w}_x + \sin\psi_c\hat{w}_y)}_{\hat{w}_{f_x}} c_c\dot{s} + \\ & \frac{1}{\sqrt{v_a^2 - \hat{w}_{f_y}^2}}(\sin\psi_c l_x e_x - \cos\psi_c l_y e_y) \end{aligned} \tag{17}$$

where the relations $\dot{\hat{w}}_* = \dot{w}_* - \dot{e}_* = -\dot{e}_*$ and $\dot{e}_* = -l_* e_*$ are used while '$*$' denoting $x$ and $y$. Moreover, one can define $\dot{\psi}_{\hat{0}} = \dot{\psi}_{\tilde{0}} + e_\psi$, where

$$\dot{\psi}_{\tilde{0}} = \frac{\hat{w}_{f_x}}{\sqrt{v_a^2 - \hat{w}_{f_y}^2}} c_c\dot{s} \tag{18}$$

and

$$e_\psi = \dot{\psi}_{\hat{0}} - \dot{\psi}_{\tilde{0}} = \frac{\sin\psi_c l_x e_x - \cos\psi_c l_y e_y}{\sqrt{v_a^2 - \hat{w}_{f_y}^2}} \tag{19}$$

At this stage, the problem of path following in the presence of wind can be solved by driving the states of system (16) to zero. This objective can be embodied in the Lyapunov function candidate:

$$V = \frac{1}{2}k_1 s_1^2 + \frac{1}{2}k_1 y_1^2 + \frac{1}{2}\psi_{\hat{w}}^2 \tag{20}$$

Its derivative can be calculated by invoking (16)-(19) as:

$$\begin{aligned} \dot{V} =& -k_1 s_1(\dot{s} - v_a cos\psi_e - \hat{w}_{f_x} - e_{f_x}) \\ & + k_1 y_1(v_a\sin\psi_e + \hat{w}_{f_y} + e_{f_y}) \\ & + \psi_{\hat{w}}(\omega - c_c\dot{s} - \dot{\psi}_{\tilde{0}} - e_\psi) \end{aligned} \tag{21}$$

where $k_1$ is a positive constant. In the observation of (21), a nonlinear control law is proposed for path following in conjunction with the disturbance observer (12):

$$\begin{aligned} \dot{s} &= k_2 s_1 + v_a\cos\psi_e + \hat{w}_{f_x} \\ \omega &= -k_3\psi_{\hat{w}} - k_1 y_1\frac{v_a\sin\psi_e + \hat{w}_{f_y}}{\psi_{\hat{w}}} + c_c(s)\dot{s} + \dot{\psi}_{\tilde{0}} \end{aligned} \tag{22}$$

where $k_2$ and $k_3$, together with $k_1$ are positive parameters to be tuned in the controller. It also can be shown that:

$$\lim_{\psi_{\hat{w}} \to 0} \frac{v_a\sin\psi_e + \hat{w}_{f_y}}{\psi_{\hat{w}}} = v_a\cos\left(\frac{\psi_e + \psi_{\hat{0}}}{2}\right) \tag{23}$$

The closed-loop stability under the composite controller needs to be investigated. In the following, we will first show the stability with no estimation error, and the stability with the disturbance observer.

*Proposition 1:* Given the estimation error $\boldsymbol{e}_d = 0$, i.e. the exact wind knowledge is available, the time-varying system (16) under the control of (22) is globally asymptotically stable.

*Proof:* Consider the Lyapunov function candidate (20). Substituting (22) into its time derivative (21) and using the assumption that $\boldsymbol{e}_d = 0$ gives:

$$\dot{V} = -k_1 k_2 s_1^2 - k_3 \psi_{\hat{w}}^2 \le 0 \tag{24}$$

Thus, $V$ is non-increasing, and this implies the states $s_1$, $y_1$ and $\psi_{\hat{w}}$ are bounded and $V$ converges to some limited value. According to Barbalat's Lemma, $\dot{V}$ converges to zero since it is uniformly continuous. From the observation of (24), $s_1$ and $\psi_{\hat{w}}$ asymptotically converge to zero.

Furthermore, by inserting the control law (22) into the system (16), it can be found that

$$\dot{\psi}_{\hat{w}} = -k_3 \psi_{\hat{w}} - k_1 y_1 \frac{v_a \sin \psi_e + \hat{w}_{f_y}}{\psi_{\hat{w}}} \tag{25}$$

As $\psi_{\hat{w}}$ approaches zero, $\dot{\psi}_{\hat{w}}$ also tends to zeros according to Barbalat's lemma. Because $\frac{v_a \sin \psi_e + \hat{w}_{f_y}}{\psi_{\hat{w}}} \neq 0$, $y_1$ is shown to converge to zero. ■

However, in practice the exact wind information is unknown, and its estimates are provided by the disturbance observer. Although the estimation error converges to zero regardless of the system states, the transit period still needs to be investigated. Thus, the stability of the overall system needs to take into account the observer dynamics. Such a system can be considered as a cascaded time-varying system:

$$\dot{\boldsymbol{x}}_1 = f_1(t, \boldsymbol{x}_1) + h(t, \boldsymbol{x}_1, \boldsymbol{x}_2)\boldsymbol{x}_2 \tag{26a}$$

$$\dot{\boldsymbol{x}}_2 = f_2(t, \boldsymbol{x}_2) \tag{26b}$$

where $\boldsymbol{x}_1 = [\begin{array}{ccc} s_1 & y_1 & \psi_{\hat{w}} \end{array}]^T$ and $\boldsymbol{x}_2 = [\begin{array}{cc} e_x & e_y \end{array}]^T$. The upper system corresponds to the system (16) and the lower system is (14), where function $f_1$ and $f_2$ represent the corresponding terms in these equations, respectively. The function $h(t, \boldsymbol{x}_1, \boldsymbol{x}_2)$ can be explicitly written as

$$h(t, \boldsymbol{x}_1, \boldsymbol{x}_2) = \begin{bmatrix} \cos \psi_c & \sin \psi_c \\ -\sin \psi_c & \cos \psi_c \\ \frac{-\sin \psi_c l_x}{\sqrt{v_a^2 - \hat{w}_{f_y}^2}} & \frac{\cos \psi_c l_y}{\sqrt{v_a^2 - \hat{w}_{f_y}^2}} \end{bmatrix} \tag{27}$$

This cascaded system can be examined by using the Theorems in [12], from which the following lemma can be drawn:

*Lemma 2:* [12] Consider the cascaded system (26). Assume that the upper system is globally uniformly asymptotically stable with a Lyapunov function of the form $V(t, \boldsymbol{x}) = k \|\boldsymbol{x}\|^p$, for all $k > 0$ and $p > 1$, whose derivative only needs to be negative-semi-definite and the lower system is global exponentially stable. If the function $h(t, \boldsymbol{x}_1, \boldsymbol{x}_2)$ satisfies

$$\|h(t, \boldsymbol{x}_1, \boldsymbol{x}_2)\| \le \theta_1(\|\boldsymbol{x}_2\|) + \theta_2(\|\boldsymbol{x}_2\|)\|\boldsymbol{x}_1\| \tag{28}$$

where $\theta_1$ and $\theta_2$ are continuous, then the cascaded system is globally uniformly asymptotically stable.

*Theorem 3:* Assume the wind disturbances are bounded and smaller than the airspeed $v_a > 0$, the error dynamics (16) of the path following problem is globally uniformly asymptotically stable under the control of the composite control law (22) and (12).

*Proof:* The proof follows the Lemma 2 to verify all the assumptions in it. First, from Proposition 1, the upper system is globally uniformly asymptotically stable and the Lyapunov function (20) satisfies the related assumptions. Then, the lower system is exponentially stable by choosing the observer gain according to (15). At last, the function $\|h(t, \boldsymbol{x}_1, \boldsymbol{x}_2)\| \le (2 + \frac{l_x^2 + l_y^2}{v_a})^{\frac{1}{2}}$ from its definition. ■

## IV. Simulations and Experiments

### A. Simulation results

The proposed path following controller based on the disturbance observer was first verified in simulation. A comparison test is presented here to compare it with the nominal controller without wind correction and two PID-like controllers based on the cross-track error [13] and the cross-track angle [14], respectively. In this test, the desired path is composed of line segments connecting four waypoints. The UAV was flying at a low airspeed of 5m/s with the wind condition $w_x = 1.5$m/s and $w_y = 2$m/s, and the yaw rate is saturated at 0.5rad/s.

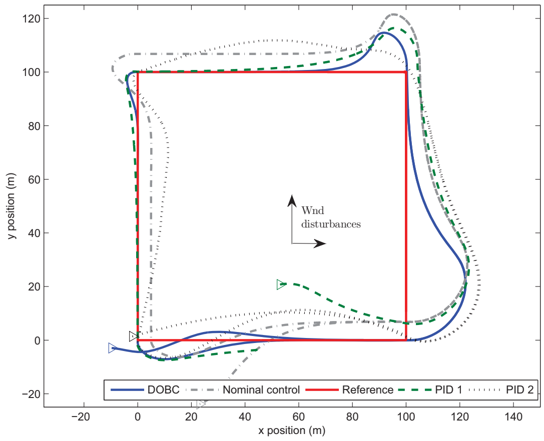


Fig. 2. Simulation Results

The simulation results are given in Fig.2. It can be seen that the proposed controller outperforms the others in terms of the path following accuracy, although the wind speed is 50% of the airspeed. The nominal controller without wind correction is able to follow the path but with steady state error due to the wind. The first PID controller based on the cross-track error can provide a competitive result but it suffers practical issues because noisy feedback signals [13]. The second PID control is based on the cross-track angle so that it aims at the next waypoint and converges to the path slowly. Note that all the controllers exhibit large converging errors at the lower right corner. This is because that the UAV heading deviated towards negative $y$ direction before the corner and need to fly into wind towards negative $x$ direction after the corner, so that a large heading change is experienced during the turning.

### B. Software-in-the-loop test

Before applying this new algorithm on the real UAV, more realistic tests need to be carried out to further evaluate its performance and minimise the risk in flight experiment. The software-in-the-loop (SIL) test is therefore performed to bridge up the gap between the numerical simulation and the practical experiment.

The structure of the SIL used in this study is shown in Fig 3. The test environment comprises three main components, namely the proposed algorithm to be tested, the Ardupilot code for the inner-loop stabilisation and the UAV dynamic model with a flight environment. The path following controller is implemented in the Simulink environment so that it can be easily debugged and tuned during the test. The Ardupilot is a commercial-of-the-shelf autopilot for inner-loop control of the UAV dynamics. It is also an open source project with the onboard code available in C/C++. Hence, its function can be simulated by recompiling its source code on a virtual machine based on a standard PC. On the other hand, although it does not belong to a flight control function to be designed, the UAV dynamic model plays an important role in the SIL test, since it needs to replicate the behaviour of the real aircraft in a software environment. To this end, the X-Plane software is adopted due to its ability to simulate realistic aircraft models and flight environment like wind conditions. The three components are connected and synchronised through the TCP/IP network connection and the flight data are transferred using a dedicated protocol.

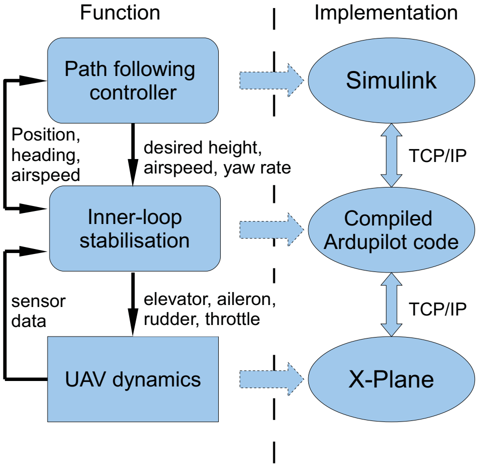

Function
Implementation
Path following controller
Simulink
Position, heading, airspeed
desired height, airspeed, yaw rate
TCP/IP
Inner-loop stabilisation
Compiled Ardupilot code
sensor data
elevator, aileron, rudder, throttle
TCP/IP
UAV dynamics
X-Plane

Fig. 3. Software-in-the-loop configuration

In the SIL simulation, an aircraft model with a similar dimension and power of the test UAV, which has a high fidelity of dynamics, was adopted in the X-Plane environment as the plant. The control gains in Ardupilot were then tuned to provide stabilisation for this aircraft. To mimic wind disturbances in reality, the weather condition in X-Plane was set up such that the wind was 6m/s towards north with the variances of ±2m/s on the speed and ±20deg on the direction.

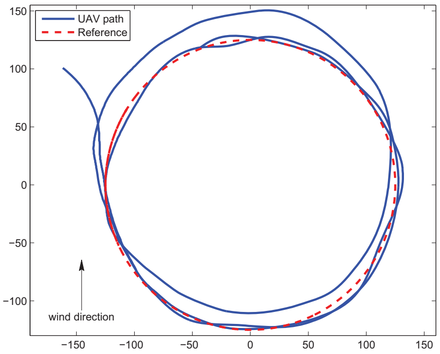


Fig. 4. SIL simulation result

One of the simulation results is presented in Fig.4. The UAV was required to follow a circle path at the airspeed of 15m/s, where the first circle was flown by the nominal control, then DOBC was switched on. It can be seen that the nominal control exhibits steady state error, whereas the proposed DOBC is able to provide accurate path following in spite of the varying wind disturbances. In addition to evaluate the performance of the proposed algorithm, the SIL simulation also verifies the integration of the high-level algorithm and the inner-loop autopilot control. This process helps to find out the potential software faults and mitigates the risk of applying the new algorithm on real UAV platform.

### C. Flight experiment

After the proposed algorithm has been tested thoroughly in simulations, it can be applied on the test UAV equipped with the Ardupilot hardware. In the flight experiment the proposed algorithm is located on a ground station and is implemented in Simulink with a sampling rate of 30Hz. The ground station is equipped with the ZigBee communication module connecting to the Ardupilot, so that the real-time fight data can be transferred back to ground station and control commands can be sent to Ardupilot. The flight experiment configuration is shown in Fig.5

Initial fight tests have been conducted with some promising results. Flight test results of using the nominal control and the DOBC are given in Fig.6 and 7, respectively, which are collected from the same test flight. The wind conditions during testing were southerly at approximately 5m/s, whose influence can be observed from Fig.6. In the flight test the DOBC was turned on after 100s, and its performance can be seen from Fig.7 where the path following accuracy was massively improved. However, there also shows oscillations on the path when the UAV flew into wind. This is because the light airframe and limited power of the test UAV and it can be alleviated by further tuning the inner-loop controller in the future flight test.

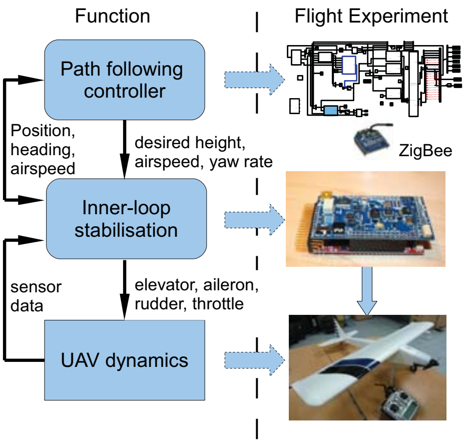

Function
Flight Experiment
Path following controller
Position, heading, airspeed
desired height, airspeed, yaw rate
ZigBee
Inner-loop stabilisation
sensor data
elevator, aileron, rudder, throttle
UAV dynamics

Fig. 5. Flight experiment configuration

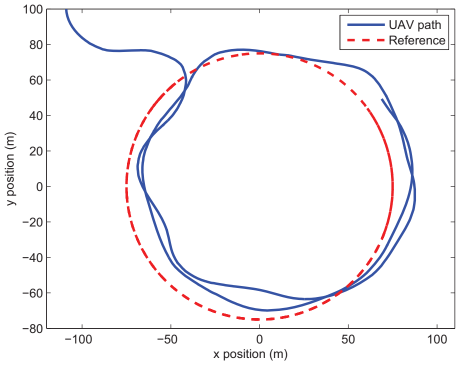


Fig. 6. Flight experiment without DOBC

## V. Summary

This paper describes a disturbance observer based design of a path following controller for small UAVs in the presence of wind. The proposed controller incorporates the wind estimates into the nominal path following controller in an intuitive way such that the UAV flies into wind with a trimming angle to cancel the wind component perpendicular to the path. The formulated composite controller including the disturbance observer is proven to be globally asymptotically stable in the theoretical analysis. Its performance is evaluated in the simulation against some other control strategies and is shown to be effective. The SIL simulation is then carried out to verify its function in a more realistic environment. The initial flight experiment is also performed and some promising results are obtained. Future work following the proposed approach include the extension to 3-D case and the incorporation of UAV's lateral dynamics.

## Acknowledgment

This work is supported by UK Engineering and Physical Science Research Council (EPSRC) under the Grant EP/J011525/1.

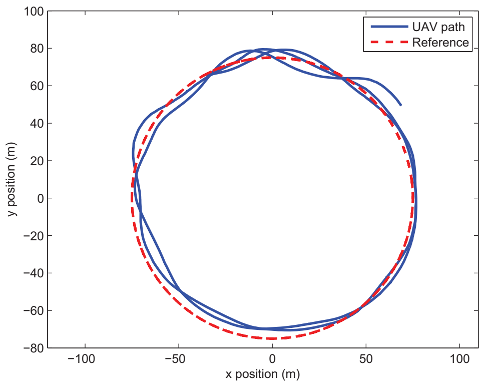


Fig. 7. Flight experiment using DOBC

## References

[1] P. Encarnacao and A. Pascoal, "Combined trajectory tracking and path following: an application to the coordinated control of autonomous marine craft," in *Decision and Control, 2001. Proceedings of the 40th IEEE Conference on*, vol. 1, 2001, pp. 964 –969 vol.1.

[2] A. Aguiar, J. Hespanha, and P. Kokotovic, "Path-following for non-minimum phase systems removes performance limitations," *Automatic Control, IEEE Transactions on*, vol. 50, no. 2, pp. 234 – 239, feb. 2005.

[3] D. Soetanto, L. Lapierre, and A. Pascoal, "Adaptive, non-singular path-following control of dynamic wheeled robots," in *Decision and Control, 2003. Proceedings. 42nd IEEE Conference on*, vol. 2, Dec. 2003, pp. 1765 – 1770.

[4] I. Kaminer, A. Pascoal, E. Xargay, N. Hovakimyan, and V. Dobrokhodov, "Path following for unmanned aerial vehicles using l1 adaptive augmentation of commercial autopilots," *Journal of guidance, control, and dynamics*, vol. 33, no. 2, pp. 550–564, 2010.

[5] I. Kaminer, O. Yakimenko, A. Pascoal, and R. Ghabcheloo, "Path generation, path following and coordinated control for time critical missions of multiple uavs," in *American Control Conference, 2006*, June 2006, pp. 4906 –4913.

[6] D. Nelson, D. Barber, T. McLain, and R. Beard, "Vector field path following for miniature air vehicles," *Robotics, IEEE Transactions on*, vol. 23, no. 3, pp. 519 –529, june 2007.

[7] J. Osborne and R. Rysdyk, "Waypoint guidance for small uavs in wind," in *Infotech@Aerospace*, Sep. 2005.

[8] T. McGee and J. Hedrick, "Path planning and control for multiple point surveillance by an unmanned aircraft in wind," in *American Control Conference, 2006*, june 2006, p. 6 pp.

[9] R. Rysdyk, "Unmanned aerial vehicle path following for target observation in wind," *Journal of guidance, control, and dynamics*, vol. 29, no. 5, pp. 1092–1100, 2006.

[10] W.-H. Chen, "Disturbance observer based control for nonlinear systems," *Mechatronics, IEEE/ASME Transactions on*, vol. 9, no. 4, pp. 706 –710, dec. 2004.

[11] C. Liu, W.-H. Chen, and J. Andrews, "Tracking control of small-scale helicopters using explicit nonlinear mpc augmented with disturbance observers," *Control Engineering Practice*, vol. 20, no. 3, pp. 258 – 268, 2012.

[12] E. Panteley and A. Loria, "On global uniform asymptotic stability of nonlinear time-varying systems in cascade," *Systems & Control Letters*, vol. 33, no. 2, pp. 131 – 138, 1998.

[13] E. Frew, T. McGee, Z. Kim, X. Xiao, S. Jackson, M. Morimoto, S. Rathinam, J. Padial, and R. Sengupta, "Vision-based road-following using a small autonomous aircraft," in *Aerospace Conference, 2004. Proceedings. 2004 IEEE*, vol. 5, march 2004, pp. 3006 – 3015 Vol.5.

[14] S. Bayraktar, G. Fainekos, and G. Pappas, "Experimental cooperative control of fixed-wing unmanned aerial vehicles," in *Decision and Control, 2004. CDC. 43rd IEEE Conference on*, vol. 4, dec. 2004, pp. 4292 – 4298 Vol.4.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Design of an Optimized Charge-Blended Energy Management Strategy for a Plugin Hybrid Vehicle

Ravi Shankar
Automotive Mechatronics Centre
Dep. of Automotive Engineering
School of Engineering
Cranfield University, UK
r.shankar@cranfield.ac.uk

Dr James Marco
Automotive Mechatronics Centre
Dep. of Automotive Engineering
School of Engineering
Cranfield University, UK
j.marco@cranfield.ac.uk

Prof Francis Assadian
Automotive Mechatronics Centre
Dep. of Automotive Engineering
School of Engineering
Cranfield University, UK
f.assadian@cranfield.ac.uk

*Abstract* - **This paper introduces the design of a charge blended energy management control system for use within a plugin hybrid electric vehicle. The approach taken extends the local cost function optimization routine associated with an Equivalent Fuel Consumption Method (EFCM) in which the charge-sustaining penalty factor is calculated online from an integrated PI controller rather than being derived from a pre-calibrated lookup table. The performance of the controller for a hybrid vehicle exercised over a number of different drive-cycles is presented. The powertrain model used to design and evaluate the system is derived from data logged onboard a number of different electric vehicles under real-world driving conditions.**

**Plugin Hybrid Electric Vehicle (PHEV), Energy Management, Equivalent Fuel Consumption Medthod (EFCM) , Controller Optimization**

## I. INTRODUCTION

Within the automotive and road transport sector, one of the main drivers for technological development and innovation is the need to reduce the vehicle's fuel consumption and the emission of Carbon Dioxide ($CO_2$) [1]. Legislative requirements are motivating manufacturers and subsystem suppliers to develop new and innovative electric vehicles (EV) and hybrid electric vehicles (HEV) concepts. In recent years, plug-in hybrids (PHEV) have also attracted considerable interest from both academia and industry.

A PHEV will typical have two primary modes of operation, namely a charge depleting mode (CD) and a charge sustaining mode (CS). Within a CD mode the vehicle operates as a zero emissions vehicle and the battery is depleted until it reaches a lower threshold. Conversely within a CS mode, an internal combustion engine (ICE), or equivalent is used to maintain the battery state of charge (SoC) within the required range. For a given journey that exceeds the zero emissions range of the vehicle, a number of publications describe a PHEV operating initially in its CD mode until the battery has depleted and then transitioning to the CS mode until the vehicle has reached its destination. Research published in [2] advocates a third mode of operation, the charge blended (CB) mode in which both the ICE and the electrical subsystems are optimally used throughout the entire journey. Because the ICE is able to operate in its most efficient region for comparatively longer, simulation results presented in [2] demonstrate an overall reduction in $CO_2$ for the journey.

The aim of this paper is to design and evaluate a CB energy management strategy for a PHEV. The instantaneous power split between the electrical subsystems and the ICE is calculated using the established Equivalent Fuel Consumption Method (EFCM) of local cost function optimization. However, the ECFM is extended by means of integrating it with a Proportional plus Integral (PI) controller in order to obtain the required SoC trajectory for the high voltage (HV) battery throughout the trip. In order to benchmark the performance of this new approach, the resulting powertrain efficiency is compared against that achieved from both a conventional EFCM controller and a thermostat strategy in which the ICE is only activate during the CS mode of vehicle operation.

This paper is structured as follows; Section II describes the derivation of the PHEV plant model employed as the foundation of this study. Sections III and IV present the design of the different energy management control systems for the PHEV. Section V describes the method of comparison used to simulate and evaluate the different control systems. Section VI presents the Results and Discussion. Conclusions and Further work are discussed in Section VII.

## II. DEVLOPMENT OF THE PHEV POWERTRAIN MODEL

### A. *Model Aims and Structure*

In order to support the evaluation of the different energy management techniques, a powertrain model for the PHEV is required of appropriate fidelity to facilitate vehicle simulations over the different drive-cycles. With simulation times in excess of 2000 seconds, a pseudo steady-state model is required in which the primary elements of the powertrain are represented by their non-linear efficiency characteristics. The powertrain model is designed to be an *enhanced* backward-facing model. Power calculations *flow* from the wheels to the main system components. However, forward-facing constraints are imposed to ensure that none of the component power ratings are invalidated. The simulation terminates if the required velocity trace cannot be met.

## B. Data Source

Parameterization of the electrical subsystems is based on the real-world vehicle usage data obtained from the Smart Move 2 Electric Vehicle Trail. As part of this program, 7 Smart Electric Drives (Smart EDs) were employed as the test vehicles. A full description of the vehicle trail is provided in [3] and will therefore not be repeated here. Over the course of 4268 km of driving, values of battery current ($i_b$), battery terminal voltage ($v_t$), vehicle speed ($v$) and inverter current ($i_m$) were recorded from the vehicle's Controller Area Network or CAN bus. As presented below, this data forms the basis for the derivation of the efficiency maps for both the electrical machine and HV battery. The remainder of this section describes the three main elements of the PHEV powertrain model, namely the vehicle mass, the electrical architecture and the ICE.

## C. Vehicle Mass

A vehicle coast down curve was experimentally obtained for the Smart ED. The coast down curve defines the resistive force ($F_r$) within the powertrain as a function of vehicle speed. The equation below defines the 3rd order polynomial best-fit approximation to the test data. The final term of 146.8 N represents the tire rolling resistance of the vehicle. The slope of the terrain (α) during the test was calculated using the measured height data obtained from the onboard GPS.

$$F_r = \left(8\times10^{-5}\right)v^3 + \left(0.0241\right)v^2 + \left(0.1456\right)v + 146.8\cos\left(\alpha\right) \quad (1)$$

The associated wheel power ($P_w$) for the vehicle is given below:

$$P_w = M_v\frac{dv}{dt} + F_r + \sin\left(\alpha\right) \quad (2)$$

The mass of the vehicle ($M_v$) was measured as 1036 kg.

## D. Electrical Architecure

### 1) Electrical Machine Model and Efficiency

The Smart ED employs a 50 kW brushless DC machine. For the PHEV model, the electric machine and the associated inverter have been considered as a single integrated system. Equation (3) was employed to calculate the efficiency of the electrical drive system $(\eta_e)$:

$$\eta_e = \frac{P_w}{v_t i_b} \quad (3)$$

The values of $v_t$ *and* $i_b$ were recorded during the EV evaluation trail under a number of different driving conditions [3]. *Fig.1* shows the mean efficiency for the electrical machine as a function of both shaft torque $\left(\tau_m\right)$ and rotor velocity $\left(\omega_m\right)$.

### 2) Battery Model and Efficiency

The Smart ED employs a 16.5kWh Lithium Ion battery, with a peak power rating of 30kW during discharge and 10kW for charge. A steady-state, equivalent circuit model is employed to represent the efficiency of the HV battery. This method is widely reported within the literature [4]. *Fig.2* presents the circuit, which comprises of a controlled open circuit voltage ($v_{oc}$) in series with a non-linear variable resistance ($R$).

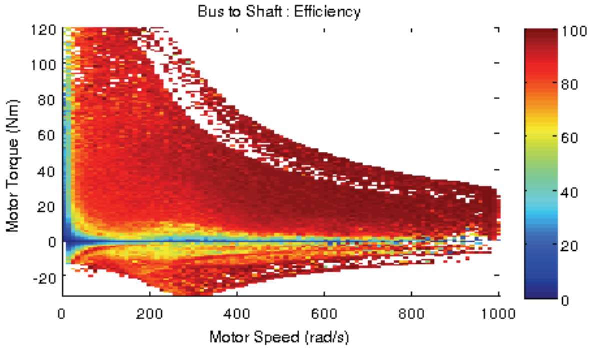


Fig. 1: Measured efficiency of the electrical drive system

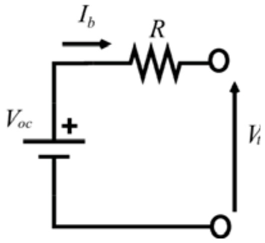


Fig. 2: Battery model

The value of $v_{oc}$ was estimated using the data recorded from [3]. The data was analyzed and the points in which $i_b$ is zero were noted. Under these conditions $v_{oc} = v_t$. *Fig.3* shows the results of this exercise and presents the estimated $v_{oc}$ as a function of measured pack temperature and recorded SoC. From Figs. 2 and 3, it is possible to derive an expression for the efficiency of the HV battery $(\eta_b)$:

$$\eta_b = \frac{v_{ocv} i_b}{v_t i_b} \quad (4)$$

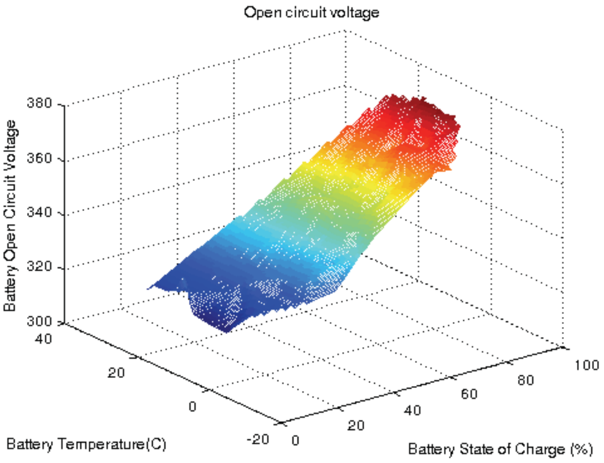


Fig. 3: Estimated battery open circuit voltage

As reported in [5] above 20% SoC, the internal resistance of lithium cells is largely insensitive to variations in battery SoC. However, battery efficiency does varies considerably due to changes in ambient temperature. *Fig. 4* presents the efficiency of the battery system as a function battery current and SoC. For the purpose of this study, battery pack temperature was assumed constant at 20 $^0$C.

## *E. The ICE Model and Efficiency*

When considering the design of an ICE model there are various levels of possible fidelity. As reported in [6], for a backward facing model in which the emphasis is on fuel economy estimation rather than transient load prediction, the ICE is often represented as a steady-state look-up table. This map defines the instantaneous mass flow rate of fuel and Best Specific Fuel Consumption (BSFC) line for the ICE. *Fig. 5* presents the engine BSFC map employed within this study. The map defines the steady-state performance for a 0.7 liter naturally aspirated 4 cylinder spark ignition engine with a maximum power of 46kW. The characteristics of this ICE are typical of the low power ICE variants often employed as a “range-extender” or Auxiliary Power Unit (APU) within a PHEV.

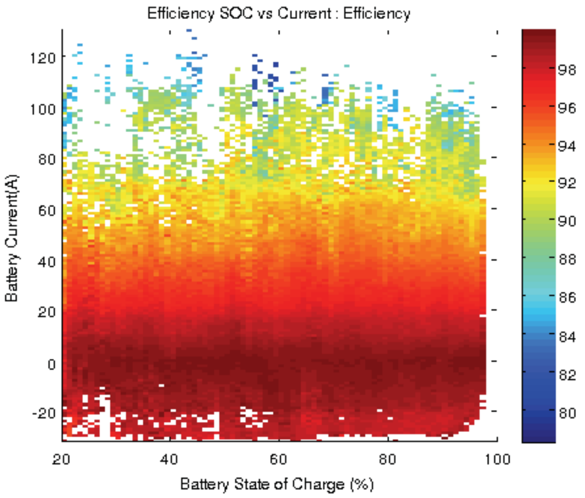


Fig. 4: Estimated battery efficiency as a function of SoC

## *F. Summary of Model Equations*

Given the non-linear parameterization data and efficiency characteristics given above, for the battery discharge conditions, equations (5)-(16) summaries the PHEV plant model employed within the control study discussed in the proceeding sections.

Based on the input drive-cycle, the wheel power ($P_w$) can be calculated directly from (1) and (2). Taking into account the efficiency of the electrical drive system including the efficiency of the final drive, the power on the HV bus can be calculated as follows:

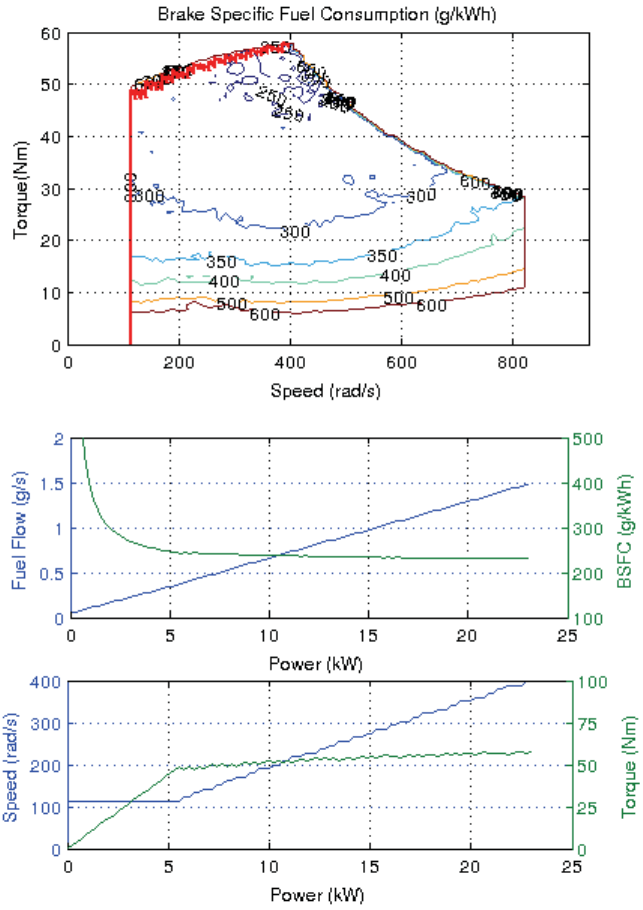


Fig. 5: Engine and Generator.

(a) Brake Specific Fuel Consumption Map (b) Best Fuel Consumption Operation Line (c) Target for engine controller

$$P_B = \frac{P_w}{\eta_e} \qquad (5)$$

$$\eta_e = f^n(\tau_m, \omega_m) \qquad (6)$$

$$\tau_m = F_r r_w \qquad (7)$$

$$\omega_m = \frac{v}{r_w} \qquad (8)$$

The term $r_w$ defines the rolling radius of the wheel. All other parameters are as previously defined. Given the power on the HV bus, the total battery power can be calculated as follows:

$$P_{Bat} = P_B + P_{aux} + P_{ICE} \qquad (9)$$

The term $P_{aux}$ defines an average auxiliary power demand for the vehicle and $P_{ICE}$ the power contribution from the ICE. Given $P_{Bat,}$ it is possible to calculate the required values of battery current ($i_b$) and battery SoC:

$$i_b(t-1) = \frac{P_{Bat}}{v_t(t-1)} \tag{10}$$

$$v_{oc} = f^n\left(SoC, T\right) \tag{11}$$

$$SoC = SoC_{ic} - \frac{1}{3600Q}\int_{t=0}^{t=STOP} i_b \, dt \tag{12}$$

$$i_b = \frac{v_{oc} - v_t}{R_{Bat}} \tag{13}$$

$$v_t = v_{oc} - i_b R_{Bat} \tag{14}$$

where Q defines the manufacturers rated battery capacity, $R_{Bat}$ defines the equivalent value of battery internal resistance derived from *Fig. 4*, t represents simulation time and finally, $SoC_{ic}$ the initial conditions for the battery SoC at the start of the simulation. All other parameters are as previously defined.

The required power from the ICE is a function of the energy management strategy and will be discussed in the following Section. With respect to *Fig. 5,* given $P_{ICE}$, the instantaneous mass flow rate of fuel ($g$) and the total fuel consumed over the cycle ($g_{total}$) can be calculated as shown below:

$$g = f^n\left(P_{ICE}\right) \tag{15}$$

$$g_{total} = \int_{t=0}^{t=stop} g \, dt \tag{16}$$

Numerical simulations and validation of the above model is provided in [5] and will therefore not be repeated again here.

## III. Design of a Thermostate energy management strategy

When designing a thermostat energy management strategy, there is relatively little analytical work that can be undertaken. The lower and upper thresholds for the strategy are defined that determine when the ICE / APU is switched on to replenish battery SoC and turned off respectively. Additional rules are often added to the heuristic control system to prevent the APU from rapidly switching, thereby causing unwanted noise, vibration and harshness (NVH) or driveability concerns within the vehicle. For the purpose of this study a simple thermostat strategy is employed in which the upper and lower SoC thresholds are defined in addition to a hysteresis function to prevent excessive cycling of the APU.

Within industry, much of the effort when integrating a thermostat approach relates to the calibration of the upper threshold. The significance of this parameter for achieving a low value of $CO_2$ is discussed further in Section VI. *Fig. 6* presents the sensitivity for the final $CO_2$ output of the PHEV over the New European Drive-Cycle (NEDC), for a lower threshold of 25% SoC to a range of different upper thresholds. Since there is an inherent coupling between the physical size of the energy storage medium, the useable SoC range and the characteristics of the journey profile (distance or kWh demand) the final $CO_2$ is highly variable and non-deterministic.

For the purpose of this study, two calibrations for the thermostat controller have been employed; the first relates to a strategy that has been optimized for use over the NEDC. The second, relates to a modified calibration that is optimized for each route under investigation. *Fig. 6,* shows the controller calibration options for a controller tuned specifically to achieve the best Tank to Wheel (TTW) emissions from the vehicle over the legislative, NEDC, drive-cycle.

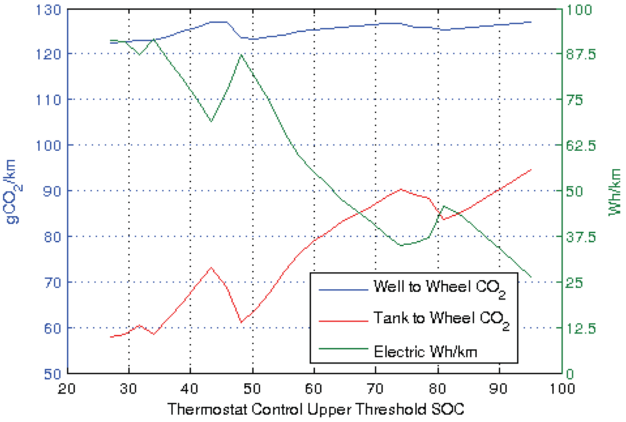


Fig. 6: Calibration of the thermostat controller (lower SoC threshold set to 25%)

## IV. Design of a EFCM energy management strategy

The EFCM is a real-time, local optimization technique that was first proposed in [8] and has been subsequently refined and extended in a number of further publications, for example [7]. The technique has also been applied to a wide range of different HEV powertrain architectures, each employing different degrees of hybridization and subsystem technologies.

### *A. Derivation of the EFCM Strategy*

The aim of this section is to present the structure, cost function and constraints associated with the proposed EFCM.

The primary function of the energy management controller is to ascertain the optimized value of $P_{ICE}$. For a given value of $P_B$, at each time-step, the strategy calculates the power demands to be sent to both the ICE and the electrical machine:

$$\begin{aligned} P_{ICE} &= (\beta) P_B \\ P_m &= (1-\beta) P_B \end{aligned} \tag{17}$$

where β represents the power split ratio. A controller sample rate of 100 ms was selected. Given this constraint, 100 unique values of β are computed for each iteration of the controller. The normalized values of β are calculated such that the following constraints are not invalidated:

$$\begin{aligned} &P_{ICE\_MIN} \le P_{ICE} \le P_{ICE\_MAX} \\ &P_m \le P_{m\_MAX} \\ &P_{B\_MIN} \le P_B \le P_{B\_MAX} \end{aligned} \tag{18}$$

The locally optimized value of β that provides the most efficient power split between the ICE and electrical machine is calculated at each time-step, by minimizing the following cost function ($J$):

$$J = MIN\left(g + g_{equiv}\zeta\right) \tag{19}$$

where $g$ is calculated from (15), $g_{equiv}$ represents the equivalent fuel consumption of the energy sunk and sourced from the HV battery taking into account the battery efficiency presented in *Fig. 4.* Finally ζ defines the charge-sustaining penalty function:

$$0 \leq \zeta \leq \zeta_{\max} \tag{20}$$

In the majority of published research, ζ represents a static look-up table that is either a linear or sigmoidal function of SoC. For the purpose of this study, two approach's to the definition of ζ have been investigated. The first is to define ζ as a linear function of SoC that weights the use of the electrical subsystems for progressively lower values of SoC (similar approaches have been discussed in [7,9]). The second is to employ a time varying value of ζ(t), in which the value is calculated from an outer PI control loop. *Fig. 7* presents the structure of the integrated PI-EFCM control system.

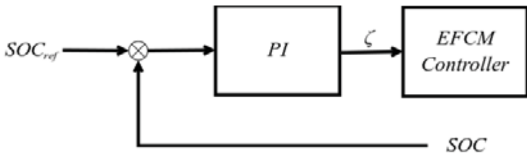


Fig. 7: Integrated PI–EFCM energy management approach

In order for this technique to be successfully applied to a PHEV as part of a CB energy management strategy, the desired set-point trajectory of SoC across the trip ($SoC_{ref}$) must be known in advance.

## V. Method of evaluation for control system performance

The aim of this Section is to introduce the method of control system evaluation employed within this study. Three drive-cycles have been used; the NEDC, the Artemis cycle and a real-world mixed urban-highway cycle that has ben logged as part of the evaluation of the Smart Move 2 trail. The speed profile of this latter cycle is 29.83 km long, it has a top speed of 28 $ms^{-1}$ and a number of start-stop events. Further information is presented in [3]. For each drive-cycle, four sets of simulations have been undertaken and the $CO_2$ output of the PHEV recorded for each study:

**Study 1:** A thermostat controller with an SoC range optimized for each specific cycle under investigation.

**Study 2:** A thermostat controller with an SoC range of 25 - 34% (i.e.: tuned for the NEDC) used on each cycle.

**Study 3:** A controller based on the EFCM in which the target SoC is fixed at 25%

**Study 4:** A controller based on the EFCM and the target SoC is managed across the cycle to facilitate a CB strategy.

In order to draw the final comparisons between a CB strategy and the different forms of a CD-CS strategy, the final energy content of the HV battery is equalized back to the initial SoC value. Using a value of $CO_2$ equal to 594 $gCO_2$ / kWh for the UK electrical grid mix, the total $CO_2$ output of the powertrain can be calculated:

$$CO_2 = CO_2\left(ICE\right) + CO_2\left(grid - kwh\right) \tag{21}$$

## VI. Results and discussuion

Table 1 presents the $CO_2$ output for each of the four studies introduced above. For each drive-cycle and control method, the Table shows the overall well-to-wheel (WTW) $CO_2$ value. In addition, Table 1 also presents how this value is broken down into the corresponding TTW $CO_2$ and the $CO_2$ contribution from the electrical supply network. The final SoC of the HV battery is also shown for completeness.

| | Study 1 | Study 2 | Study 3 | Study 4 |
|---|---|---|---|---|
| | Thermostat | Thermostat<br>(tuned controller) | EFCM<br>(CS-Mode) | EFCM<br>(CB-Mode) |
| NEDC | | | | |
| WTW $CO_2km^{-1}$ | 143.47 | 143.47 | 143.95 | 141.87 |
| TTW $CO_2km^{-1}$ | 59.87 | 59.87 | 58.10 | 57.30 |
| $Whkm^{-1}$ | 123.10 | 123.10 | 127.4 | 125.48 |
| Final SoC (%) | 27.15 | 27.15 | 24.81 | 25.85 |
| ARTEMIS | | | | |
| WTW $CO_2km^{-1}$ | 169.08 | 168.04 | 177.13 | 168.44 |
| TTW $CO_2km^{-1}$ | 52.97 | 48.27 | 54.76 | 47.32 |
| $Whkm^{-1}$ | 179.84 | 188.41 | 189.87 | 189.96 |
| Final SoC (%) | 28.75 | 25.65 | 25.12 | 25.09 |
| REAL-WORLD ROUTE | | | | |
| WTW $CO_2km^{-1}$ | 123.37 | 122.36 | 122.25 | 121.71 |
| TTW $CO_2km^{-1}$ | 59.87 | 58.00 | 56.99 | 56.53 |
| $Whkm^{-1}$ | 89.24 | 91.25 | 93.07 | 93.06 |
| Final SoC (%) | 27.86 | 26.39 | 25.04 | 25.04 |

Table 1: PHEV efficiency results for different energy management techniques

For the real-world drive-cycle, *Fig. 8* presents the simulation results from studies 2-4. The figure shows the vehicle speed profile, the instances in which the PHEV is operating as an EV or when additional power is supplied from the ICE. From Table 1, given the variations in the TTW $CO_2$ values obtained, one of the main drawbacks of the thermostat approach are highlighted. For a PHEV, final $CO_2$ achieved for the trip or drive-cycle, is highly dependent on the nature of the journey (distance, level of driver demand), the physical size of the battery (kWh) and the upper turn-off threshold for the controller. The ideal calibration for the control system is one that ensures the battery at its lowest allowable SoC point at the end of the journey.

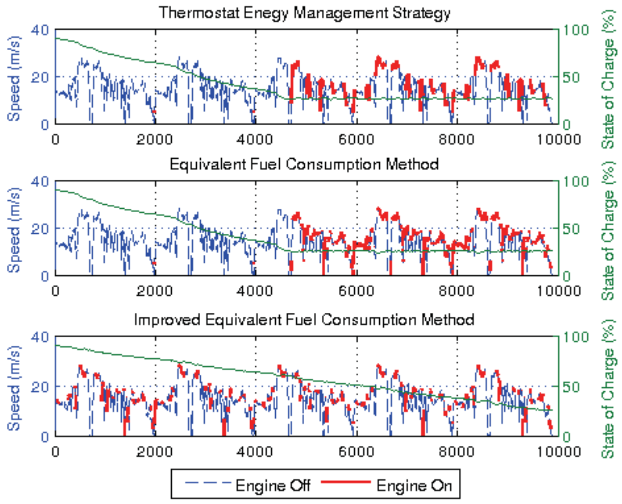


Fig. 8: Drive-cycle results for the PHEV

From Table 1, the primary conclusion that can be drawn is that the CB strategy, designed using an EFCM, consistently results in a PHEV powertrain system with the lowest WTW $CO_2km^{-1}$ values. For each cycle, this is true even when compared against a thermostat strategy that has been calibrated to optimize the performance of the vehicle over that cycle. Even though the differences presented here are marginal (in the order of 1-2%) two points should be noted. Firstly, this improvement is achieved simply by adopting a new control methodology. No change to the physical design of the vehicle has been assumed. The second point to note, is that if consideration is restricted to the TTW analysis, then the differential between the different control methods is considerably greater. The $CO_2$ reduction possibilities are in the order of 4-10%. While the TTW comparisons do not take into account the residual energy content of the HV battery, this method of evaluation is inline with current EU fuel economy measurement processes for PHEVs [10]. Within Regulation 101, two fuel economy metrics are cited, the $CO_2km^{-1}$ from the vehicle over the cycle and separately a $Whkm^{-1}$ is also published representing the energy required from the grid to replenish the battery. Currently within the UK only the former is employed when setting vehicle taxation levels and to alleviate congestion charges.

The improvement in efficiency when comparing either a CS or CB control system using EFCM can be seen from *Fig. 8.* When a closed-loop CB control system is employed, the ICE is consistently used more to support the high power demands from the driver and hence more power is fed directly to the road wheels. With a CS strategy, in which the ICE is used only when the battery SoC has depleted to its lower threshold, the ICE is forced to operate for longer periods within it's more inefficient low power regions.

## VII. CONCLUSIONS AND FURTHERWORK

A PHEV powertrain model has been developed to support the design and optimization of a CB energy management control system. The electrical subsystems within the model have been calculated from data logged from a Smart ED vehicle under real-world usage conditions.

The energy management technique extends a conventional EFCM controller by integrating it with a closed loop PI control system. The output of the PI controller is a time-varying charge sustaining penalty function that allows for a more robust control of the battery SoC, irrespective of the nature of the drive-cycle or the physical size of the battery.

In order to implement this CB control system, it has been assumed that the control system knows the distance of the journey that the vehicle is traversing. This assumption is valid for a number of applications such as public transportation or vehicle fleet operation. In order to support the wider application of this control method, research is currently ongoing to investigate the accuracy of on-line estimation techniques in which the required SoC trajectory for a CB strategy across a trip can be calculated as a function of the different road types, traffic congestion levels and the differing energy demands from with the driver.

## VIII. ACKNOWLEDGEMENTS

The research presented within this paper is supported by the UK Technology Strategy Board (TSB) through the Knowledge Transfer Partnership (KTP) scheme in collaboration with the Morgan Motor Company .

## IX. REFERENCES

[1] C. C. Chan (2002), The state of the art of electric and hybrid vehicles, Proceedings of the IEEE, vol. 90, issue 2, ppP704

[2] T. Markel, (2005), Energy Storage Systems Considerations for Grid-Charged Hybrid Electric Vehicles, IEEE Conference on Vehicle Power and Propulsion, pp. 62, USA.

[3] R. Shankar and J. Marco, (2011), Performance of an EV During Real-World Usage, Cenex Hybrid Electric Vehicle Conference, UK.

[4] O.,Tremblay O., et al., (2007), A Generic Battery Model for the Dynamic Simulation of Hybrid Electric Vehicles, IEEE Conference on Vehicle Power and Propulsion, pp. 284, USA.

[5] R. Shankar, J. Marco and F. Assadian, (2012), A Methodology to Determine Drivetrain Efficiency Based on the External Environment, IEEE Electric Vehicle Conference (IEVC), USA.

[6] Rizzoni G., L. Guzzella and B. M. Baumann, (1999), Unified Modelling of Hybrid Electric Vehicle Drivetrains, IEEE Transactions on Mechatronics, vol. 4, issue 3, pp. 246-257.

[7] C. Musardo, G. Rizzoni and B. Staccia, (2005), A-ECMS: An Adaptive Algorithm for Hybrid Electric Vehicle Energy Management, IEEE Conference on Decision and Control, and the European Control Conference, pp. 1816-1823, Spain.

[8] G. Paganelli, S. Delprat, T. M. Guerra, J. Rimaux and J. J. Santin, (2002), Equivalent Consumption Minimization Strategy for Parallel Hybrid Powertrains, 55th IEEE *Vehicular Technology Conference (VTC),* vol. 4, pp. 2076-2081, USA.

[9] G. Paganelli, et al., (2001), General Supervisory Control Policy for the Energy Optimization of Charge-Sustaining Hybrid Electric Vehicles, JSAE, vol. 22, issue 4, pp. 511.

[10] unece.org, Regulation 101 - Battery Electric Vehicles with Regard to Specific Requirements for Construction and Functional Safety, Last Accessed January, 2011.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Modelling Channel Flow over Riblets: Calculating the Energy Amplification

Aditya Kasliwal, Stephen Duncan and Antonis Papachristodoulou

*Abstract*— **Riblets have been considered as a passive method for drag reduction. Riblets are structures on a surface that run parallel to one another, which are aligned longitudinally to the flow. It has been shown experimentally that when the shape, spacing and height of the riblets are optimized, the drag coefficient over the surface can be reduced by up to 10%. These results have also been confirmed by direct numerical simulation studies. Although the benefits of riblets have been known since the early 1980's, the mechanism of drag reduction is still not fully understood. This paper examines the effect of riblet structures on the amplification of background noise within channel flow between two parallel plates (Poiseuille flow), where riblets are present on the surface of one of the plates. A linearized version of the Navier-Stokes equation about the steady flow is developed and through a coordinate transformation, the boundary conditions associated with the riblets are transferred into the partial differential equations. Previous work has used spectral methods to discretize these equations, leading to a large-scale state space model, and the energy amplification was calculated for the streamwise constant component of the flow from the controllability gramian. However, solving the associated Lyapunov equation can be computationally prohibitive, which limits the density of the discretized grid. This paper shows how the problem can be transformed to decouple the system, so that the gramian can be obtained by solving a set of smaller Lyapunov equations, which has the potential to allow the energy amplification to be calculated for systems with a dense discretization grid.**

*Index Terms*— **Fluid flow, drag reduction, Lyapunov equation, energy amplification, riblets**

## I. Introduction

One approach that has been considered in the past that results in a significant reduction in the drag caused by a fluid flowing past a body is the introduction of riblet structures on the surface of the body [1], [2], [3]. Riblets are structures that run parallel to one another, that are positioned longitudinally to the flow and usually have a triangular cross-section in the transverse direction. By contrast with approaches such as suction/blow mechanisms, riblets can be regarded as a passive drag reduction mechanism [4]. Previous research has shown that riblet structures can be optimized to produce a reduction in the drag coefficient of up to 10% [2]. It is believed that the mechanism responsible for this level of drag reduction is the interaction of the riblets with the structure of the boundary layer, which leads to a reduction in drag, despite the significant increase in the area of the modified surface that is in contact with the flow compared to a smooth surface. This drag-reduction mechanism is observed widely in nature, for example on shark skin and scallop shells [5], [6]. Static riblet structures (often in the form of films that cover the surface) have been used in a number of fields and in the 1990's tests on a scale model of an Airbus A320 cruising at Mach 0.7 have shown reductions in viscous drag of 4.85% [7], [8]. However, despite these benefits, there has not as yet been a satisfactory explanation for the mechanism of drag reduction using riblet structures. Experimental and Computational Fluid Dynamics (CFD) studies have suggested that riblets induce streamwise vortices that sit within the riblets [9], [10], although recent work has shown that riblets induce spanwise vortices close to the riblet surface and it is these that change the drag over the surface [11], [12].

The results presented here build on previous work [13], [14] that combines ideas from control theory and aerodynamics, by extending an existing direct numerical simulation of the flow between two parallel plates (Poiseulle flow), where riblets are present on the surface of one of the plates. The model uses concepts from linear control theory to determine the transient growth of background noise energy amplification in the linearized Navier-Stokes equations [15], [16]. We investigate how the transient growth in energy alters with the introduction of riblet structures by using the model to predict how the shape and positioning of the riblets affect this transient energy growth. The model calculates the amplification of background noise energy, together with the drag on the surface of the plates, for a given riblet geometry, by using a transformation of coordinates that will 'passes' the non-uniformity of the boundary associated with the riblet structure on one of the walls, into the equations of motion. Once this equation is obtained, a numerical model is developed using a Chebyshev discretization approach in the wall-normal direction and a Fourier representation in the streamwise and spanwise directions and estimates of the reduction in energy amplification are obtained as a function of different geometries. Previous work [13], [14] has modelled the flow at relatively low Reynolds numbers due to the computational complexity of the problem, because in order to find the energy amplification, it was necessary to solve a high dimension Lyapunov equation, which is a computationally intensive step. However, by exploiting the inherent structure of the problem that is associated with the Fourier representation in the spanwise direction, it is possible to decompose the problem into a set of decoupled Lyapunov equations of lower dimension. This has the potential to allow the flow to be modelled at higher Reynolds numbers and reduced riblet spacing that require denser discretization grids, which is where the benefits of the riblets structures are

A. Kasliwal, S. Duncan and A. Papachristodoulou are with the Department of Engineering Science, University of Oxford, Parks Road, Oxford, OX1 3PJ, United Kingdom. Emails aditya.kasliwal@mansfield.ox.ac.uk, stephen.duncan@eng.ox.ac.uk, antonis@eng.ox.ac.uk

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

observed. It will also allow comparison with experimental results and with CFD simulations that have been reported in the literature.

The paper is organized as follows. Section II develops the equations describing the channel flow and uses a coordinate transformation to convert the boundary conditions associated with the riblet structures into a uniform domain. In Section III, the equations describing the linearized flow about the steady flow are derived. Section IV shows how Fourier methods can be used to create a large-scale, finite-dimensional state space model and describes how the structure of the model can be exploited to calculate the amplification of the energy associated with background noise in a computationally efficient manner. Section V gives the results for the effect on the energy amplification of introducing riblets for a specific flow regime and Section VI concludes the paper.

## II. Transforming the Equations of Motion

We consider channel flow between two stationary plates with the geometry of the problem as shown in Fig. 1, where $\tilde{x}$, $\tilde{y}$ and $\tilde{z}$ are the coordinates of the streamwise, wall-normal and spanwise directions, respectively. The upper wall is a flat plate, while the lower wall is a plate with riblets aligned with the streamwise direction. The dimensions of the problem are normalized, so that the upper boundary of the flow occurs at the plate positioned at $\tilde{y} = 1$, while the lower boundary is at

$$\tilde{y} = -1 + f(\tilde{z}) \tag{1}$$

where $f(\tilde{z})$ describes the “shape” of the riblets. The analysis will be restricted to riblets that are aligned with the streamwise direction and are independent of $\tilde{x}$.

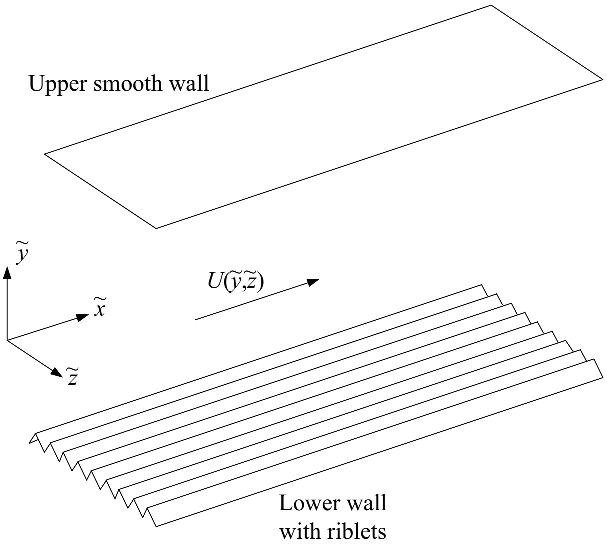


Fig. 1. Three-dimensional view of computational domain showing riblet structure on lower wall of channel. $U(\tilde{y}, \tilde{z})$ denotes the direction of the steady, streamwise flow.

The streamwise, wall-normal and span-wise components of the flow in the coordinate system $(\tilde{x}, \tilde{y}, \tilde{z})$ are denoted by $u(\tilde{x}, \tilde{y}, \tilde{z}, t)$, $v(\tilde{x}, \tilde{y}, \tilde{z}, t)$ and $w(\tilde{x}, \tilde{y}, \tilde{z}, t)$ respectively, and the non-dimensionalised Navier-Stokes equations in this coordinate system are given by

$$\left(\frac{\partial}{\partial t} + u\frac{\partial}{\partial \tilde{x}} + v\frac{\partial}{\partial \tilde{y}} + w\frac{\partial}{\partial \tilde{z}}\right) u = -\frac{\partial p}{\partial \tilde{x}} + \frac{1}{\mathrm{Re}}\tilde{\Delta} u \tag{2}$$

$$\left(\frac{\partial}{\partial t} + u\frac{\partial}{\partial \tilde{x}} + v\frac{\partial}{\partial \tilde{y}} + w\frac{\partial}{\partial \tilde{z}}\right) v = -\frac{\partial p}{\partial \tilde{y}} + \frac{1}{\mathrm{Re}}\tilde{\Delta} v \tag{3}$$

$$\left(\frac{\partial}{\partial t} + u\frac{\partial}{\partial \tilde{x}} + v\frac{\partial}{\partial \tilde{y}} + w\frac{\partial}{\partial \tilde{z}}\right) w = -\frac{\partial p}{\partial \tilde{z}} + \frac{1}{\mathrm{Re}}\tilde{\Delta} w \tag{4}$$

$$\frac{\partial u}{\partial \tilde{x}} + \frac{\partial v}{\partial \tilde{y}} + \frac{\partial w}{\partial \tilde{z}} = 0 \tag{5}$$

where $p$ is the pressure, $\mathrm{Re}$ is the Reynolds number and

$$\tilde{\Delta} = \frac{\partial^2}{\partial \tilde{x}^2} + \frac{\partial^2}{\partial \tilde{y}^2} + \frac{\partial^2}{\partial \tilde{z}^2} \tag{6}$$

For the smooth wall at $\tilde{y} = 1$, the boundary conditions are

$$u(\tilde{x}, 1, \tilde{z}) = v(\tilde{x}, 1, \tilde{z}) = w(\tilde{x}, 1, \tilde{z}) = 0 \tag{7}$$

together with

$$\left.\frac{\partial v}{\partial \tilde{y}}\right|_{\tilde{y}=1} = 0 \tag{8}$$

while for the riblet wall at $\tilde{y} = -1 + f(\tilde{z})$, the boundary conditions are

$$\begin{aligned} u(\tilde{x}, -1 + f(\tilde{z}), \tilde{z}) &= v(\tilde{x}, -1 + f(\tilde{z}), \tilde{z}) \\ &= w(\tilde{x}, -1 + f(\tilde{z}), \tilde{z}) = 0 \end{aligned} \tag{9}$$

with

$$\left.\frac{\partial v}{\partial n}\right|_{\tilde{y}=-1+f(\tilde{z})} = 0 \tag{10}$$

where $n$ is the normal to the surface of riblets.

We now apply a change of coordinates

$$\begin{aligned} x &= \tilde{x} \\ y &= F(\tilde{y}, \tilde{z}) \\ z &= \tilde{z} \end{aligned} \tag{11}$$

where

$$y = F(\tilde{y}, \tilde{z}) = \frac{2\tilde{y} - f(\tilde{z})}{2 - f(\tilde{z})} \tag{12}$$

which has the effect of mapping $\tilde{y} \in [-1 + f(\tilde{z}), 1]$ to $y \in [-1, 1]$.

The Navier-Stokes equations in the new coordinate system then become [14]

$$\left(\frac{\partial}{\partial t}+u\frac{\partial}{\partial x}+v\frac{\partial F}{\partial \tilde{y}}\frac{\partial}{\partial y}+w\frac{\partial F}{\partial \tilde{z}}\frac{\partial}{\partial y}+w\frac{\partial}{\partial z}\right)u = -\frac{\partial p}{\partial x}+\frac{1}{\mathrm{Re}}\tilde{\Delta}u \tag{13}$$

$$\left(\frac{\partial}{\partial t}+u\frac{\partial}{\partial x}+v\frac{\partial F}{\partial \tilde{y}}\frac{\partial}{\partial y}+w\frac{\partial F}{\partial \tilde{z}}\frac{\partial}{\partial y}+w\frac{\partial}{\partial z}\right)v = -\frac{\partial F}{\partial \tilde{y}}\frac{\partial p}{\partial y}+\frac{1}{\mathrm{Re}}\tilde{\Delta}v \tag{14}$$

$$\left(\frac{\partial}{\partial t}+u\frac{\partial}{\partial x}+v\frac{\partial F}{\partial \tilde{y}}\frac{\partial}{\partial y}+w\frac{\partial F}{\partial \tilde{z}}\frac{\partial}{\partial y}+w\frac{\partial}{\partial z}\right)w = -\frac{\partial p}{\partial z}-\frac{\partial F}{\partial \tilde{z}}\frac{\partial p}{\partial y}+\frac{1}{\mathrm{Re}}\tilde{\Delta}w \tag{15}$$

$$\frac{\partial u}{\partial x}+\frac{\partial F}{\partial \tilde{y}}\frac{\partial v}{\partial y}+\frac{\partial F}{\partial \tilde{z}}\frac{\partial w}{\partial y}+\frac{\partial w}{\partial z}=0 \tag{16}$$

where

$$\begin{aligned}\tilde{\Delta} &= \frac{\partial^2}{\partial x^2}+\left(\frac{\partial F}{\partial \tilde{y}}\right)^2\frac{\partial^2}{\partial y^2}+\left(\frac{\partial F}{\partial \tilde{z}}\right)^2\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\\ &\quad+\frac{\partial}{\partial \tilde{y}}\left(\frac{\partial F}{\partial \tilde{y}}\right)\frac{\partial}{\partial y}+\frac{\partial}{\partial \tilde{z}}\left(\frac{\partial F}{\partial \tilde{z}}\right)\frac{\partial}{\partial y}+2\frac{\partial F}{\partial \tilde{z}}\frac{\partial^2}{\partial y\partial z}\end{aligned} \tag{17}$$

## III. Linearized Equations

The steady state flow, $(U,0,0)$, satisfies

$$U\frac{\partial U}{\partial x} = -\frac{\partial P}{\partial x}+\frac{1}{\mathrm{Re}}\tilde{\Delta}U \tag{18}$$

$$0 = -\frac{\partial F}{\partial \tilde{y}}\frac{\partial P}{\partial y} \tag{19}$$

$$0 = -\frac{\partial P}{\partial z}-\frac{\partial F}{\partial \tilde{z}}\frac{\partial P}{\partial y} \tag{20}$$

$$\frac{\partial U}{\partial x} = 0 \tag{21}$$

Since we are restricting our attention to the case of straight riblets aligned with the streamwise direction, so that $f(\hat{z})$ does not depend upon $\hat{x}$, then $P=P(x)$ and $U=U(y,z)$, where $U(y,z)$ is the solution of

$$\tilde{\Delta}U=-2\mathrm{Re} \tag{22}$$

We now redefine the streamwise flow as $u(x,y,z,t)+U(y,z)$, where $u(x,y,z,t)$ denotes the flow *relative* to the solution of steady flow problem, $U(y,z)$, and the pressure as $p(x,y,z,t)+P(x)$, where $p(x,y,z,t)$ is the pressure *relative* to $P(x)$. Linearizing about this steady solution and applying the change of coordinates, the Navier-Stokes equations become

$$\left(\frac{\partial}{\partial t}+U\frac{\partial}{\partial \tilde{x}}\right)u+v\frac{\partial U}{\partial \tilde{y}}+w\frac{\partial U}{\partial \tilde{z}} = -\frac{\partial p}{\partial \tilde{x}}+\frac{1}{\mathrm{Re}}\tilde{\Delta}u \tag{23}$$

$$\left(\frac{\partial}{\partial t}+U\frac{\partial}{\partial \tilde{x}}\right)v = -\frac{\partial p}{\partial \tilde{y}}+\frac{1}{\mathrm{Re}}\tilde{\Delta}v \tag{24}$$

$$\left(\frac{\partial}{\partial t}+U\frac{\partial}{\partial \tilde{x}}\right)w = -\frac{\partial p}{\partial \tilde{z}}+\frac{1}{\mathrm{Re}}\tilde{\Delta}w \tag{25}$$

$$\frac{\partial u}{\partial \tilde{x}}+\frac{\partial v}{\partial \tilde{y}}+\frac{\partial w}{\partial \tilde{z}} = 0 \tag{26}$$

Introducing

$$\eta=\frac{\partial u}{\partial \tilde{z}}-\frac{\partial w}{\partial \tilde{x}} \tag{27}$$

and rearranging, we obtain

$$\begin{aligned}\left(\frac{\partial}{\partial t}+U\frac{\partial}{\partial \tilde{x}}\right)\eta &- \frac{\partial U}{\partial \tilde{z}}\frac{\partial v}{\partial \tilde{y}}+\frac{\partial U}{\partial \tilde{y}}\frac{\partial v}{\partial \tilde{z}}+v\frac{\partial^2 U}{\partial \tilde{y}\partial \tilde{z}}+w\frac{\partial^2 U}{\partial \tilde{z}^2}\\ &= \frac{1}{\mathrm{Re}}\tilde{\Delta}\eta\end{aligned} \tag{28}$$

The flow is now described by equations that depend upon $v$, $\eta$ and $w$. To eliminate $w$, define

$$\eta=\frac{\partial u}{\partial \tilde{z}}-\frac{\partial w}{\partial \tilde{x}} \tag{29}$$

which leads to

$$w=-\left(\frac{\partial^2}{\partial \tilde{z}^2}+\frac{\partial^2}{\partial \tilde{x}^2}\right)^{-1}\left(\frac{\partial \eta}{\partial \tilde{x}}+\frac{\partial^2 v}{\partial \tilde{y}\partial \tilde{z}}\right) \tag{30}$$

We now assume that the flow is streamwise constant, so that $v$, $\eta$ and $w$ are independent of $\tilde{x}$, and $\frac{\partial}{\partial \tilde{x}}=\frac{\partial}{\partial x}=0$, and following some lengthy manipulations (details are given in [13], [14]), the linearized flow can be described by

$$\frac{\partial}{\partial t}\tilde{\Delta}v = \frac{1}{\mathrm{Re}}\tilde{\Delta}\tilde{\Delta}v \tag{31}$$

$$\frac{\partial}{\partial t}\eta = \tilde{\mathcal{C}}v+\frac{1}{\mathrm{Re}}\tilde{\Delta}\eta \tag{32}$$

where

$$\tilde{\mathcal{C}}=\frac{\partial U}{\partial \tilde{z}}\frac{\partial}{\partial \tilde{y}}-\frac{\partial U}{\partial \tilde{y}}\frac{\partial}{\partial \tilde{z}}-\frac{\partial^2 U}{\partial \tilde{y}\partial \tilde{z}}+\frac{\partial^2 U}{\partial \tilde{z}^2}\left(\frac{\partial^2}{\partial \tilde{z}^2}\right)^{-1}\frac{\partial^2}{\partial \tilde{y}\partial \tilde{z}} \tag{33}$$

## IV. Representation in Spatial Fourier Domain

Using (17), the expression for $\tilde{\Delta}$ and $\tilde{\Delta}\tilde{\Delta}$ can be expressed in terms of the steady-state flow $U$ and the partial differentials of $F(\tilde{y},\tilde{z})$. In order to obtain expressions for these operators, we require

$$y=F(\tilde{y},\tilde{z})=\frac{2\tilde{y}-f(\tilde{z})}{2-f(\tilde{z})} \tag{34}$$

and because $\tilde{z}=z$, then

$$\tilde{y}=y+\frac{f(z)}{2}(1-y) \tag{35}$$

so that

$$\frac{\partial F}{\partial \tilde{y}} = \frac{2}{2 - f(z)} \tag{36}$$

$$\frac{\partial F}{\partial \tilde{z}} = \frac{(y-1)\, f'(z)}{2 - f(z)} \tag{37}$$

where $f'(z)$ denotes the differential of $f(z)$ with respect to $z$. The higher partial derivatives of $F(\tilde{y}, \tilde{z})$ can also be derived, but the resulting expressions for the transformed operators are rather cumbersome and are omitted from this paper. Full details are given in [14].

The operator $\tilde{\Delta}$ can be written as

$$\tilde{\Delta} = K_1(y,z)\frac{\partial}{\partial y} + K_2(y,z)\frac{\partial^2}{\partial y \partial z} + K_3(y,z)\frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{38}$$

with

$$K_1(y,z) = \frac{\partial}{\partial z}\left(\frac{\partial F}{\partial \tilde{z}}\right) + \frac{\partial F}{\partial \tilde{y}}\frac{\partial}{\partial y}\left(\frac{\partial F}{\partial \tilde{y}}\right) + \frac{\partial F}{\partial \tilde{z}}\frac{\partial}{\partial y}\left(\frac{\partial F}{\partial \tilde{z}}\right) \tag{39}$$

$$K_2(y,z) = 2\frac{\partial F}{\partial \tilde{z}} \tag{40}$$

$$K_3(y,z) = \left(\frac{\partial F}{\partial \tilde{y}}\right)^2 + \left(\frac{\partial F}{\partial \tilde{z}}\right)^2 \tag{41}$$

The operators in (31) and (32) depend upon the steady state profile $U(y,z)$. Because the flow is taken to be periodic in the $\tilde{z}$ direction with period $2\pi$, and since $\tilde{z} = z$ then

$$U(y,z) = U(y, z + 2\pi) \tag{42}$$

the spanwise component of the flow can be approximated by a linear combination of Galerkin trial functions, $\mathrm{e}^{\mathrm{i}nz}$ for $n \in \mathbb{Z}$, so that

$$U(y,z) = \sum_{n=-\infty}^{\infty} \hat{U}_n(y)\mathrm{e}^{\mathrm{i}nz} \tag{43}$$

The riblet structure is also taken to be periodic in the $\tilde{z}$ (and $z$) direction, so that $f(z)$ can be expressed as

$$f(z) = \sum_{n=-\infty}^{\infty} \hat{f}_n \mathrm{e}^{\mathrm{i}nz} \tag{44}$$

However, assuming that there are an integer number of riblets, $P \in \mathbb{Z}^+$, in the region $z \in [0, 2\pi]$, then $\hat{f}_n = 0$ for all $n \neq mP$, where $m$ is an integer. When the spacing between riblets is small, as is usually the case for flow with high Reynolds numbers, then $P$ is large, which means that only coefficients $\hat{f}_n$ for values of $n$ that are multiples of $P$ are non-zero, and this sparsity is exploited in the solution of the problem.

The expression in (36) and (37) can be simplified by defining

$$g(z) = \frac{1}{2 - f(z)} \tag{45}$$

Expanding $g(z)$ as a linear combination of powers of $f(z)$

$$g(z) = \frac{1}{2 - f(z)} = \frac{1}{2}\left(1 + \frac{1}{2}f(z) + \frac{1}{4}[f(z)]^2 + \ldots\right) \tag{46}$$

then this is periodic and can be expressed as

$$g(z) = \sum_{n=-\infty}^{\infty} \hat{g}_n \mathrm{e}^{\mathrm{i}nz} \tag{47}$$

where $\hat{g}_n = 0$ for all $n \neq mP$. Since the expressions in (36) and (37) and the higher derivatives are periodic, the terms $K_1(y,z)$, $K_2(y,z)$ and $K_3(y,z)$ can also be expressed in terms of Fourier series of the form

$$K_1(y,z) = \sum_{n=-\infty}^{\infty} \hat{k}_n^{(1)}(y)\,\mathrm{e}^{\mathrm{i}nz} \tag{48}$$

$$K_2(y,z) = \sum_{n=-\infty}^{\infty} \hat{k}_n^{(2)}(y)\,\mathrm{e}^{\mathrm{i}nz} \tag{49}$$

$$K_3(y,z) = \sum_{n=-\infty}^{\infty} \hat{k}_n^{(3)}(y)\,\mathrm{e}^{\mathrm{i}nz} \tag{50}$$

Because both $\hat{f}_n$ and $\hat{g}_n$ are non zero when $n$ is an exact multiple of $P$, then this will also be true for the coefficients, $\hat{k}_n^{(1)}$, in these expansions.

Using these expansions in the flow equation (22) gives

$$\sum_{n=-\infty}^{\infty} \hat{k}_n^{(1)}(y)\,\mathrm{e}^{\mathrm{i}nz}\left(\frac{\partial}{\partial y}\sum_{m=-\infty}^{\infty}\hat{U}_m(y)\mathrm{e}^{\mathrm{i}mz}\right) + \sum_{n=-\infty}^{\infty} \hat{k}_n^{(2)}(y)\,\mathrm{e}^{\mathrm{i}nz}\left(\frac{\partial^2}{\partial y \partial z}\sum_{m=-\infty}^{\infty}\hat{U}_m(y)\mathrm{e}^{\mathrm{i}mz}\right) + \sum_{n=-\infty}^{\infty} \hat{k}_n^{(3)}(y)\,\mathrm{e}^{\mathrm{i}nz}\left(\frac{\partial^2}{\partial y^2}\sum_{m=-\infty}^{\infty}\hat{U}_m(y)\mathrm{e}^{\mathrm{i}mz}\right) + \sum_{n=-\infty}^{\infty}\frac{\partial^2 \hat{U}_n(y)}{\partial z^2}\mathrm{e}^{\mathrm{i}nz} = -2\mathrm{Re} \tag{51}$$

Defining an inner product as

$$\langle p(y,z)q(y,z)\rangle = \int_0^{2\pi} \overline{p(y,z)}\, q(y,z)dz \tag{52}$$

then taking the inner product of the expression in (51) with a Galerkin trial function, $\mathrm{e}^{\mathrm{i}\ell z}$ and exploiting the orthogonality of the trial functions with respect to $z$

$$2\pi \sum_{m=-\infty}^{\infty} \hat{k}_{\ell-m}^{(3)}(y)\frac{\mathrm{d}^2\hat{U}_m}{\mathrm{d}y^2} + \left[\hat{k}_{\ell-m}^{(1)}(y) + \mathrm{i}m\,\hat{k}_{\ell-m}^{(2)}(y)\right]\frac{\mathrm{d}\hat{U}_m}{\mathrm{d}y} - m^2\hat{U}_m(y) = b_\ell \qquad \text{for } \ell \in \mathbb{Z} \tag{53}$$

where

$$b_\ell = \begin{cases} -4\pi\mathrm{Re} & \text{for} \quad \ell = 0 \\ 0 & \text{for} \quad \ell \neq 0 \end{cases} \tag{54}$$

If $U(y,z)$ is approximated by its projection onto a finite basis, then the summation is limited to $2M{+}1$ terms, and (53 reduces to $2M{+}1$ coupled second order, ordinary differential equations. The values of $\hat{U}_m(y)$ can be solved numerically at the Chebyshev points on $y \in [-1, 1]$ by stacking the values of $\hat{U}_m(y)$ at $K$ sample points for each $\ell$ into a vector $\hat{\mathbf{U}} \in \mathbb{R}^{(2M+1)K}$, so that

$$\hat{\mathbf{U}} = \left[\; \hat{\mathbf{U}}_{-M} \; \ldots \; \hat{\mathbf{U}}_{-1} \; \hat{\mathbf{U}}_0 \; \hat{\mathbf{U}}_1 \; \ldots \hat{\mathbf{U}}_M \;\right]^{\mathrm{T}} \tag{55}$$

where $\hat{\mathbf{U}}_m \in \mathbb{R}^K$ are the samples of $U_m(y)$ at the $K$ Chebyshev points [17], [18]. The discretised version of the coupled ODE's in (53) take the form

$$\mathbf{E}\hat{\mathbf{U}} = \mathbf{b} \tag{56}$$

where $\mathbf{b} \in \mathbb{C}^{(2M+1)K}$ contains the terms $b_\ell$. The $\mathbf{E} \in \mathbb{C}^{(2M+1)K\times(2M+1)K}$ consists of a series of blocks $[\mathbf{E}]_{\ell,m} \in \mathbb{C}^{K\times K}$, that satisfy

$$\begin{aligned}[\mathbf{E}]_{\ell,m} = \operatorname{diag}\left\{\hat{\mathbf{k}}^{(3)}_{\ell-m}\right\}\mathbf{D}_K^2 + \operatorname{diag}\left\{\hat{\mathbf{k}}^{(1)}_{\ell-m} + \mathrm{i}m\,\hat{\mathbf{k}}^{(2)}_{\ell-m}\right\}\mathbf{D}_K \\ - \operatorname{diag}\left\{m^2\right\}\end{aligned} \tag{57}$$

where $\mathbf{D}_K \in \mathbb{R}^{K\times K}$ and $\mathbf{D}_K^2 \in \mathbb{R}^{K\times K}$ are the first and second Chebyshev differentiation matrices respectively, and $\hat{\mathbf{k}}^{(1)}_{\ell-m} \in \mathbb{C}^K$, $\hat{\mathbf{k}}^{(2)}_{\ell-m} \in \mathbb{C}^K$, $\hat{\mathbf{k}}^{(3)}_{\ell-m} \in \mathbb{C}^K$ and $\hat{\mathbf{k}}^{(4)}_{\ell-m} \in \mathbb{C}^K$ are vectors obtained by sampling $\hat{k}^{(1)}_{\ell-m}(y)$, $\hat{k}^{(2)}_{\ell-m}(y)$ and $\hat{k}^{(3)}_{\ell-m}(y)$ at the $K$ Chebyshev points.

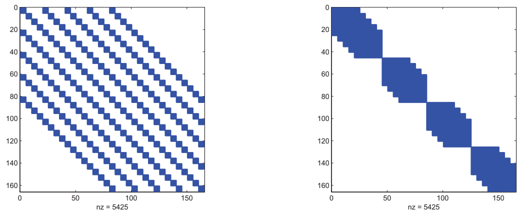


Fig. 2. Structure of $\mathbf{E}$ matrix (a) before reordering and (b) after reordering

The key point is that because $\hat{k}^{(1)}_{\ell-m}(y)$, $\hat{k}^{(2)}_{\ell-m}(y)$ and $\hat{k}^{(3)}_{\ell-m}(y)$ are only non zero when $\ell - m$ is an exact multiple of $P$, the $\mathbf{E}$ matrix has the structure shown in Figure 2(a), where each of the individual blocks has dimension $K$ by $K$. By rearranging the order of the terms in $\hat{\mathbf{U}}$, the structure of the $\mathbf{E}$ matrix can be arranged into the block diagonal form shown in Figure 2(b), which consists of $P$ blocks, each of dimension $[2(M/P)+1]K$ by $[2(M/P)+1]K$. This structure can be exploited to solve the overall problem as a set of $P$ individual sub-problems, which has two advantages. Firstly, solving $P$ smaller problems reduces the computational load by a factor of $P$, and secondly, the problem can be implemented as $P$ sub-problems, which can be solved separately.

*Remark 1.* For the case shown in the Figure 2, the first block has dimension $[2(M/P)+1]K$ by $[2(M/P)+1]K$, but the subsequent blocks have dimension $2(M/P)K$ by $2(M/P)K$. This is a consequence of using $2M+1$ terms in the finite Galerkin expansion in (43). To ensure that all blocks are of equal size, it is necessary to restrict the expansion to $2M$ terms so that $m \in \{-M+1, \ldots, -1, 0, 1, \ldots, M\}$, although this makes the indexing of the matrices more complicated.

*Remark 2.* To solve the steady state problem, it is only necessary to solve the sub-problem associated with the first sub-block, because the elements of the $\mathbf{b}$ vector associated with the other sub-problems are all zero. However, this will not be the case when solving the full linearised problem.

The linearised equations for the flow in (31) and (32) can be expressed in terms of Fourier series for

$$v(y,z,t) = \sum_{n=-\infty}^{\infty} \hat{v}_n(y,t)\mathrm{e}^{\mathrm{i}nz} \tag{58}$$

$$\eta(y,z,t) = \sum_{n=-\infty}^{\infty} \hat{\eta}_n(y,t)\mathrm{e}^{\mathrm{i}nz} \tag{59}$$

We discretize the problem using a Chebyshev grid in the $y$ direction, which reduces the differential operators in (31) and (32) to matrices [19]. Because of the orthogonality of the Galerkin approximation in the $z$ direction, the problem is reduced to a decoupled set of state space models of the form (the procedure is the exactly the same as in the derivation of (57) for the steady state solution, although the expressions are cumbersome, so details are given in [20])

$$\frac{d}{dt}\begin{bmatrix}\hat{\mathbf{v}}_\ell \\ \hat{\eta}_\ell\end{bmatrix} = \begin{bmatrix}\hat{\mathbf{L}}^\ell_{11} & \mathbf{0} \\ \hat{\mathbf{L}}^\ell_{21} & \hat{\mathbf{L}}^\ell_{22}\end{bmatrix}\begin{bmatrix}\hat{\mathbf{v}}_\ell \\ \hat{\eta}_\ell\end{bmatrix} \tag{60}$$

where $\hat{\mathbf{L}}^\ell_{11}$, $\hat{\mathbf{L}}^\ell_{21}$ and $\hat{\mathbf{L}}^\ell_{22}$ are the matrices obtained by discretizing the operators for each Fourier coefficient. The states are not functions of the streamwise position $x$ as it is assumed that the flow is streamwise constant.

The amplification of the energy at the output of the system resulting from the addition of Gaussian white noise at each sampling position in the Chebyshev grid can be found by solving the Lyapunov equation to find the controllability gramian, $\hat{\mathbf{X}}^\ell$ of the (discretized) system associated with each Fourier mode

$$\hat{\mathbf{A}}^\ell\hat{\mathbf{X}}^\ell + \hat{\mathbf{X}}^\ell\hat{\mathbf{A}}^{\ell\,\mathrm{T}} + \hat{\mathbf{Q}}^\ell = 0 \tag{61}$$

where $\mathbf{A}$ matrix is block diagonal

$$\hat{\mathbf{A}}^\ell = \begin{bmatrix}\hat{\mathbf{L}}^\ell_{11} & \mathbf{0} \\ \hat{\mathbf{L}}^\ell_{12} & \hat{\mathbf{L}}^\ell_{22}\end{bmatrix} \tag{62}$$

and $\hat{\mathbf{Q}}^\ell$ is the covariance matrix of the white noise, which has the structure

$$\hat{\mathbf{Q}}^\ell = \begin{bmatrix}\hat{\mathbf{Q}}^\ell_{11} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{Q}}^\ell_{22}\end{bmatrix} \tag{63}$$

where $\hat{\mathbf{Q}}^\ell_{11}$ and $\hat{\mathbf{Q}}^\ell_{22}$ are diagonal. The structure of the problem can be exploited to break it into three separate equations [15], each of which can be solved efficiently using Cholesky factorizations [14].

The total amplification of the noise energy is given by the trace of $\mathbf{C}\hat{\mathbf{X}}^\ell\mathbf{C}^\mathrm{T}$, but the maximum amplification is given by the mode associated with the largest eigenvalue of this matrix, where $\mathbf{C}$ is the discretization of the output operators that relate $\hat{u}_\ell(y,t)$ and $\hat{w}_\ell(y,t)$ to $\hat{v}_\ell(y,t)$ and $\hat{\eta}_\ell(y,t)$. In practice, it is more efficient to exploit the structure of the problem by finding the largest singular value of $\mathbf{CR}$, where $\mathbf{R}$ is the Cholesky decomposition of $\hat{\mathbf{X}}^\ell$ in (61). The mode associated with the largest amplification is then given by the corresponding singular vector [13].

## V. Results

As an initial study, the effects of including a sinusoidal riblet for the linearised flow at a Reynolds number Re = 4200 are investigated. The $u(\tilde{y}, \tilde{z})$ component of the mode associated with the maximum amplification when riblets with peak to peak separation of $2\pi/8$ and amplitude 0.15 are introduced onto the lower wall in the channel flow is shown in Fig. 3. It can be seen that in this component, the flow is raised up above the riblets and the largest component of this flow occurs in the upper half of the channel. This has the effect of reducing the flow in the lower half of the channel, which in turn, reduces the gradient of the shear stress at the riblet wall. It should be emphasised that this models the linearised flow; in practice, it is likely that the nonlinear effects will have a more important role at this Reynolds number.

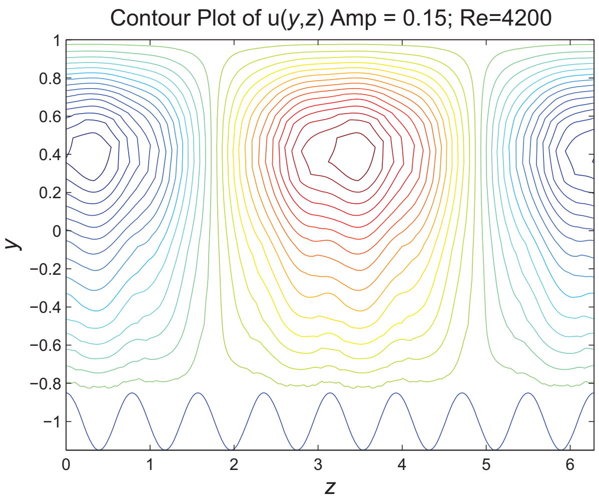


Fig. 3. Contour plot of streamwise velocity $u(\tilde{y}, \tilde{z})$ component of mode associated with maximum amplification for the case with riblets on lower wall.

## VI. Conclusion

This paper has developed a linearized model of the channel flow between two plates, where one of the plates has a riblet structure. The boundary conditions associated with the riblets are transferred into the partial differential equations describing the flow by applying a transformation of coordinates. The number of components in the flow equations are reduced by restricting attention to streamwise constant variations. By expressing the model in terms of spatial Fourier components in the spanwise direction, the structure of the resulting large-scale, finite-dimensional state space model can be exploited, allowing the amplification of noise energy to be obtained by calculating the controllability gramian via a set of decoupled Lyapunov equations. The model shows that for flow at low Reynolds numbers, the inclusion of riblets reduces the maximum amplification of the noise energy and it appears that this reduction in energy amplification may be associated with the presence of counter-rotating vortices at the peaks of the riblet structures. Current work is using the model to examine the effect of riblets on the energy amplification in flows at a range of Reynolds numbers and for a range of riblet amplitudes and separations.

## References

[1] M. Walsh, “Riblets as a viscous drag reduction technique,” *AIAA J.*, vol. 21, pp. 485–486, 1983.

[2] ——, “Riblets,” in *Viscous Drag Reduction in Boundary Layers*, D. Bushnell and J. Hefner, Eds. New York, NY: AIAA, 1990, pp. 203–261.

[3] D. Bechert, M. Bruse, W. Hage, J. van der Hoeven, and G. Hoppe, “Experiments on drag-reducing surfaces and their optimization with an adjustable geometry,” *J. Fluid Mechanics*, vol. 338, pp. 59–87, 1997.

[4] E. Coustols and A. Savill, “Turbulent skin-friction drag reduction by active and passive means,” in *Special Course on Skin Friction Drag Reduction*, J. Cousteix, Ed. AGARD Report 796, 1993, pp. 8–1 – 8–80.

[5] D. Bechert, M. Bruse, and W. Hage, “Experiments with three-dimensional riblets as an idealized model of shark skin,” *Experiments in Fluids*, vol. 28, pp. 403–412, 2000.

[6] B. Dean and B. Bhushan, “Shark-skin surfaces for fluid-drag reduction in turbulent flow: a review,” *Philosophical Transactions of the Royal Society A*, vol. 368, no. 1929, pp. 4775–4806, 2010.

[7] P. Viswanath, “Aircraft viscous drag reduction using riblets,” *Progress in Aerospace Sciences*, vol. 38, pp. 571–600, 2002.

[8] J. Reneaux, “Overview on drag reduction technologies for civil transport aircraft,” in *Proc. European Congress on Computational Methods in Applied Sciences and Engineering ECCOMAS*, Jyvaskyla, Finland, 2004.

[9] H. Choi, P. Moin, and J. Kim, “Direct numerical simulation of turbulent flow over riblets,” *J. Fluid Mechanics*, vol. 255, pp. 503–539, 1993.

[10] D. Chu and G. Karniadakis, “A direct numerical simulation of laminar and turbulent flow over riblet-mounted surfaces,” *J. Fluid Mechanics*, vol. 250, pp. 1–42, 1993.

[11] R. García-Mayoral and J. Jiménez, “Hydrodynamic stability and breakdown of the viscous regime over riblets,” *J. Fluid Mechanics*, vol. 678, pp. 317–347, 2011.

[12] ——, “Drag reduction by riblets,” *Philosphical Transactions of the Royal Society A*, vol. 369, no. 1940, pp. 1412–1427, 2011.

[13] S. Duncan and A. Papachristodoulou, “Energy amplification in channel flow over riblets,” in *Proc. IEEE Multi-Conf. on Systems and Control*, Denver, CO, 2011, pp. 456–461.

[14] ——, “Analyzing the energy amplification in channel flow over riblets,” Department of Engineering Science, University of Oxford, Oxford, UK, Tech. Rep. 2327/11, 2011.

[15] B. Bamieh and M. Dahleh, “Energy amplification in channel flows with stochastic excitation,” *Physics of Fluids*, vol. 13, no. 11, pp. 3258–3268, 2001.

[16] M. Jovanovic and B. Bamieh, “Componentwise energy amplification in channel flows,” *J. Fluid Mechanics*, vol. 534, pp. 145–183, 2005.

[17] J. Boyd, *Chebyshev and Fourier Spectral Methods*, 2nd ed. New York, NY: Dover, 2001.

[18] L. Trefethen, *Spectral Methods in Matlab*. Philapelphia, PA: SIAM, 2000.

[19] J. Weidman and S. Reddy, “A MATLAB differentiation matrix suite,” *ACM Transactions on Mathematical Software*, vol. 26, no. 4, pp. 465–519, 2000.

[20] A. Kasliwal, “Drag reduction over aerofoils fitted with riblets,” Department of Engineering Science, University of Oxford, Oxford, UK, Tech. Rep., 2012.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Assessment of Fault Tolerance for Actively Controlled Railway Wheelset

Mohammad Mirzapour
School of Computing, Science Engineering
The University of Salford
Great Manchester, M5 4WT, UK
m.mirzapour@edu.salford.ac.uk

T.X Mei
School of Computing Science Engineering
The University of Salford
Great Manchester, M5 4WT, UK
t.x.mei@salford.ac.uk

I Hussain
School of Electronic Engineering
Mehran University
Jamshoro, Pakistan
imtiaz.hussain@faculty.muet.edu.pk

*Abstract*— **this paper studies the key issue of fault tolerance for actively controlled railway wheelsets. It assesses failure modes in such systems, with a focus on actuator failures, and consequence of those hardware failures. It seeks to establish the necessary basis for control reconfiguration to ensure system stability and performance in the event of a faulty, without the need for hardware redundancies. A number of control schemes (with and without faults) are included in the study. Both analytical and simulation results are presented.**

***Keywords-Railway; Wheelset; Active control; Stability; Actuator; Fault tolerance.***

## I. INTRODUCTION

Conventional wheelset for the railway vehicle is composed of the two coned (or profiled) wheels rigidly fixed to a common axle to rotate at the same angular velocity. When an unconstraint wheelset rolling along the track it is displaced laterally due to track irregularities, the rolling radii therefore are different because of the profiles of these two wheelset. Consequently, different forward speeds obtained for each wheelset due to the difference rolling radii to provide a natural centering/curving action. However, an unconstrained wheelset also presents a problem of kinematic instability known as the "Kinematic Oscillation" or wheelset "hunting" [1,2].

Traditionally the wheelset is stabilized by using passive suspensions on conventional rail vehicle, but such additional stiffness affects the pure rolling action of the wheelset around the curve. It has been theoretically proven that to this design conflict between stability and curving performance can be solved by applying active control instead of conventional passive components within the primary suspension of railway vehicle [2].

Passive components in the primary suspensions can be designed in such a way not to fail in order to maintain the stability and steering performance of railway vehicle and they are generally accepted as "safe" in railway industry. However, any new technology must prove that it can cope with any failures to demonstrate that any component faults would not lead to the system failure such that passenger safety is not compromised under such conditions. From a practical point of view, any active control scheme must be also able to maintain an effective operation of a rail system in order to meet the necessary standard of reliability [3]. Hardware redundancy technique may be used in the system to guarantee safety operation of such a system. Whilst it may be acceptable to apply the above technique in sensors due to their relatively low cost, it is far more difficult to justify the use of multiple actuators in a cost effective manner for redundancy or accommodate those within the limited space of railway bogie [3]. There are two main approaches for fault tolerant control systems. The first philosophy relies on the existing system redundancies to achieve acceptable performance in the event of component failures. In this type of systems, once the controllers designed, it will remain stable. It should be noted that the redundancies in such a system are usually in hardware forms. The second methodology takes a completely different approach to achieve fault tolerance. It involves such procedures as real-time fault detection, isolation, and control system reconfiguration. The redundancy in such a system may be an analytical form [4] and help to minimize the use of the hardware redundancies in order to keep the overall cost down [5].

The object of this study is to develop the fault tolerance approaches without using redundant actuators to provide stability across a range of operation conditions with different failure modes. It investigates the possibilities/feasibilities of re-configuring the controller based on the use of remaining actuator(s) in the system. For this study, the paper will review first a number of different control methods for railway wheelsets in the normal condition to understand how control for stability and/or curving performance is achieved. A thorough assessment of failure modes and adverse effect of the faults on the system stability and performance is then carried out, followed by an investigation into control re-tuning/re-configuration for fault tolerance. The paper is organized as following. The mathematical dynamic model of railway vehicle is presented in Section п. Consideration of basic

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

control scheme is given in section ш. Section V demonstrates the different fail modes of railway vehicle with the actuator faults and re-configuration of the controller based on the remaining actuator Finally, conclusion and future work will be discussed.

## II. Mathematical model of the rail way vehicle

A railway vehicle mainly consists of a vehicle body and two bogie frames, and each bogie frame consists of a bogie frame and the two wheelsets. The wheelsets are connected to the bogie frame with springs and dampers in the longitudinal and lateral directions. For this study, only the plan-view dynamics of a half vehicle is used to analysis stability and steering performance of the vehicle, which is the accepted practice in railway industry [6]. Fig.1 gives a plan-view diagram of the half body vehicle model used for this study. The equations of motion for a railway vehicle when running along a track are mainly determined by the creep forces between wheel and rail contact patches. In this paper, a linear model has been considered, which is justified as the active control tends to reduce the effect of non-linearity in the wheelsets [7]. The linear model of the motion contains seven degrees of freedom, i.e. the lateral and yaw motions for each wheelset and for the bogie frame, and a lateral displacement for the vehicle body defined by (1) to (7). The model is therefore 14th order in total, and can be represented in the state space model by (8) [6].

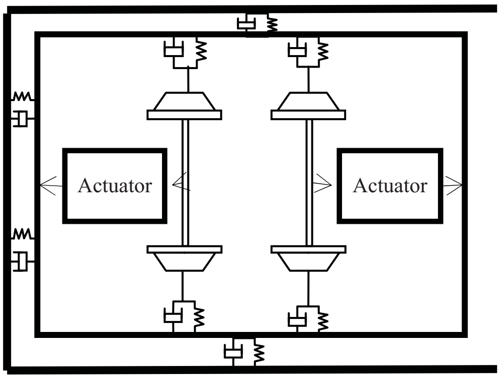


Figure 1: Plane-view of the vehicle

The input vector $u$ represents the control inputs to the wheelsets, and the vector $\mu$ is used to represent the inputs from the railway track, including the lateral displacement, cant, and curvature. The lateral track displacement is a random input, which represents track irregularities along the path track, whereas the track curvature and cant are the deterministic inputs [6]. More details of the vehicle parameters in the equations are provided in the Appendix A.

$$\ddot{y}_{w_1} = -\frac{1}{m_w}[(\frac{2f_{11}}{V_s}+C_s)\dot{y}_{w_1} - K_s y_{w_1} + 2f_{22}\Psi_{w_1} + C_s\dot{y}_g + K_s y_g$$

$$+C_s L_v \dot{\Psi}_g + K_s L_v \Psi_g + m_w(\frac{V_s^2}{R_1} - g\cdot\theta_{c_1})], \tag{1}$$

$$\ddot{\Psi}_{w_1} = -\frac{1}{I_w}[\frac{-2f_{11}L_g^2}{I_w V_s}\dot{\Psi}_{w_1} - \frac{-2f_{11}\lambda L_g}{I_w r_0}y_{w_1} + \frac{2f_{11}L_g^2}{I_w R_1}$$

$$-\frac{-2f_{11}\lambda L_g}{I_w r_0}y_{t_1} + \frac{\tau_{w_1}}{I_w}], \tag{2}$$

$$\ddot{y}_{w_2} = -\frac{1}{m_w}[(\frac{2f_{11}}{V_s}+C_s)\dot{y}_{w_2} - K_s y_{w_2} + 2f_{22}\Psi_{w_2} + C_s\dot{y}_g + K_s y_g$$

$$-C_s L_v \dot{\Psi}_g - K_s L_v \Psi_g + m_w(\frac{V_s^2}{R_2} - g\cdot\theta_{c_2})], \tag{3}$$

$$\ddot{\Psi}_{w_2} = -\frac{1}{I_w}[\frac{-2f_{11}L_g^2}{I_w V_s}\dot{\Psi}_{w_2} - \frac{-2f_{11}\lambda L_g}{I_w r_0}y_{w_2} + \frac{2f_{11}L_g^2}{I_w R_2} \tag{4}$$

$$-\frac{-2f_{11}\lambda L_g}{I_w r_0}y_{t_2} + \frac{\tau_{w_2}}{I_w}],$$

$$\ddot{y}_g = \frac{1}{m_g}[-(2\cdot C_s + C_{sc})\cdot\dot{y}_g - (2\cdot K_s + K_{sc})y_g + C_s\dot{y}_{w_1}$$

$$+K_s y_{w_1} + C_s\dot{y}_{w_2} + K_s y_{w_2} + C_{sc}\dot{y}_v + K_{sc}y_v \tag{5}$$

$$+m_g V_s^2(\frac{1}{R_1}+\frac{1}{R_2}) - m_g g(\frac{\theta_{c_1}}{2}+\frac{\theta_{c_2}}{2})],$$

$$\ddot{\Psi}_g = \frac{1}{I_g}[-2L_v^2 C_s\dot{\Psi}_g - 2L_v^2 K_s\Psi_g + L_v C_s\dot{y}_{w_1} - L_v C_s\dot{y}_{w_2} \tag{6}$$

$$+L_v K_s y_{w_1} - L_v K_s y_{w_2} - (\tau_{w_1} + \tau_{w_2})]$$

$$\ddot{y}_v = \frac{1}{m_v}[-C_{sc}\dot{y}_v - K_{sc}y_v + C_{sc}\dot{y}_g + K_{sc}y_g$$

$$+m_v V_s^2(\frac{1}{2R_1}+\frac{1}{2R_2}) - m_v g(\frac{\theta_{c_1}}{2}+\frac{\theta_{c_2}}{2})], \tag{7}$$

$$\dot{x} = A\cdot x + B\cdot u + \mu\cdot w, \tag{8}$$

$$x = [\dot{y}_{w_1} \quad y_{w_1} \quad \dot{\Psi}_{w_1} \quad \Psi_{w_1} \quad \dot{y}_{w_2} \quad y_{w_2} \quad \dot{\Psi}_{w_2} \quad \Psi_{w_2} \quad \dot{y}_g \quad y_g \quad \dot{\Psi}_g \quad \Psi_g \quad \dot{y}_v \quad y_v]$$

$$u = \lfloor \tau_{w_1} \quad \tau_{w_2} \rfloor,$$

$$\mu = \begin{bmatrix} y_{t_1} & \theta_{c_1} & \frac{1}{R_1} & y_{t_2} & \theta_{c_2} & \frac{1}{R_2} \end{bmatrix},$$

## III. BASIC WHEELSET CONTROL SCHEMES

The railway wheelset can be stabilized by using either passive suspension or through the use of active control. For active approaches, it is possible to achieve this by applying either a yaw torque or lateral force between the bogie and the wheelset, but the yaw control is preferred as it also tends to improve the ride quality experienced by passengers [8]. Therefore, this study only discusses approaches that apply control in the yaw direction to provide desired damping to stabilize the system. The review of the control strategies is to provide a background for the study of fault tolerant control issues and more detail of the controls can be found in the references provided [9,6,10]. The suspension/control schemes are considered in the study are:

- Passive Suspension that uses conventional passive yaw stiffness in the primary suspensions.
- Active Yaw Damping where the two wheelset of the bogie are controlled by applying a yaw torque proportional to the lateral velocity of the wheelset.
- Sky-hook Yaw Stiffness where the control output of each actuator is set to be proportional to the absolute yaw motion of each wheelset
- Optimal Control where the controllers for the two inputs (actuators) are designed with the use of full state feedback (from either direct measurements or through the use of an estimator).

The track input used in the simulation, to study the control performance on curves for both active controllers and passive suspension, represents a curved track with radius of 1250m connected to straight track via a transition of 2sec. The curved track is canted inward by 6 degrees to reduce the lateral acceleration experienced by the passengers (a normal features of railway track). The vehicle speed of 50m/s is used – parameters of the vehicle are provided is Appendix A. The simulation result in Fig.2 and Fig.3 clearly illustrates that active control can provide good curving performances to reduce the longitudinal and lateral creep forces, compared with passive suspension, when both leading and trailing actuator functioning normally.

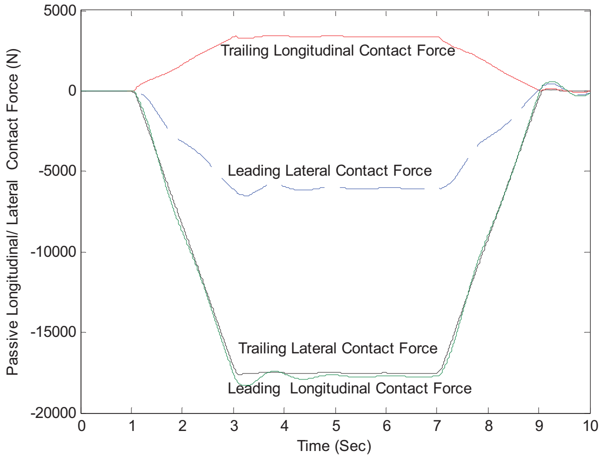


Figure 2: Passive Longitudinal/ Lateral creep forces

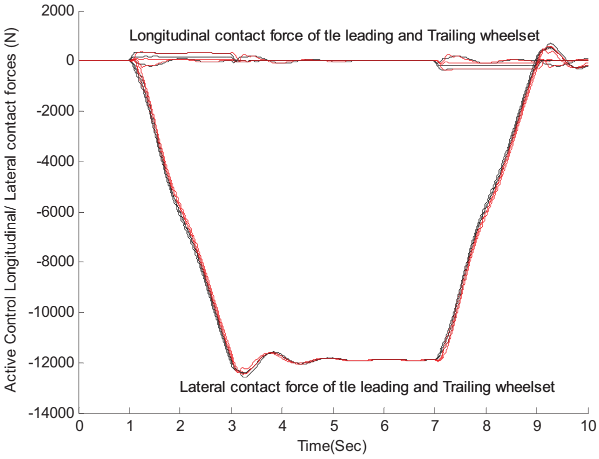


Figure 3: Active control Longitudinal/ Lateral contact forces

However, the active controllers are designed based on the assumption that both actuators function as expected and extra measures (possibly through the use of fault tolerance) will be needed in order to maintain the stability and if possible curving performance if one of the control channels fails to deliver. Therefore, it is essential to establish first a full understanding how the fault(s) affect the bogie dynamics in the system and then define what corrective actions can be taken [3].

## IV. CONTROL ANLYSIS IN FAULT CONDITIONS

In the normal condition with both actuators functioning, the bogie is designed to be stable. Fig.4 shows the minimum damping ratio of the wheelset modes with the different controllers where the stability is achieved across a wide range of speed with a critical speed of over 100m/s in the three of controllers except active yaw damper [3].

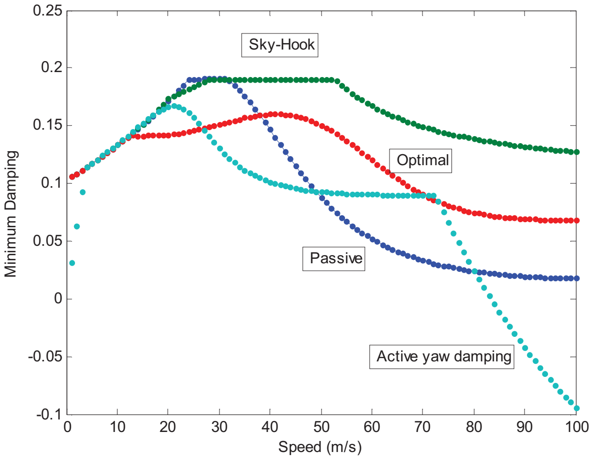


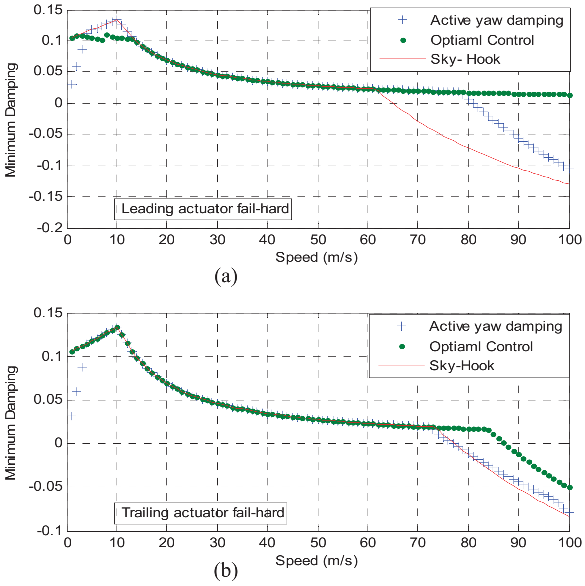
Figure 4: Comparison of different control scheme

However, this is expected to change dramatically, when one of the actuators fails. In this study, two failure modes are considered – one is fail-hard and the other fail-soft, representing an actuator jam and free-motion respectively.

The aim of fault tolerance for actively control system is to preserve stability conditions and maintain the current curving performance close to desired ones (or at least not worse than the passive system in the normal condition) in the presence of actuator faults. In this study, the full state feedback is considered as a start point and a control gain matrix is designed to control the remaining actuator in order to explore the fault tolerant control possibilities.

### *A. Fail-Hard*

In the fail hard mode, one of the actuators is assumed to be blocked and can be simulated as a spring with very high stiffness between the bogie frame and the axle. The bogie stability at different speed for the selected active control schemes are compared in Figs.5a and 5b, where the fail hard occurs in the leading and trailing wheelset respectively. In Figs.5a and 5b, the critical speed is reduced to 62m/s for the active sky-hook control scheme with the malfunction of the leading actuator and around 75m/s if the trailing actuator fails-hard respectively. A fault in the trailing actuator would result in an even lower critical speed for all three active control schemes. However the fail-hard condition poses a more problem for the curving performance of the bogie.



(a)

(b)

Figure 5: Stability of different control schemes with leading/trailing fault

The simulation result in Fig.6 and Fig.7 indicate clearly that when fail hard occurs in the leading and trailing control input respectively, the original controller will not be able to provide the 'right' control effort as the contact creep forces at the both leading and trailing wheelset increase significantly on the curved track, delivering a poor steering condition. The objective of the fault tolerance in the fail-hard case is therefore is to try and minimize the adverse impact of the actuator failure on the curving performance. It can be seen that the curving performance is more under risk when fail-hard occurs for the leading actuator. The simulation results in Fig. 6 and 7 suggest that the contact forces with the original controller are even worse than that with the passive suspension (Fig.2). However, retuning of the control gains for the remaining actuator does seem to provide a solution to improve curving performance of the bogie in the event of fail-hard in the leading or trailing actuator as evidenced by results in Fig.8 and Fig.9 respectively. The simulation results in Fig.8 and 9 in comparison with Fig.6 and 7 of the original controller indicate that the re-tuned optimal controller for the remaining actuator can reduce the contact forces and maintain curving performance close to that of the passive suspension. In this approach optimal controllers are tuned manually by choosing different values for weighting factor. Although the stability is guaranteed with the optimal control design, the re-design of the other active control schemes is less straight-forward and the research is ongoing to ensure such designs will meet both stability and performance requirements, e.g. by applying optimization technique to search for the best control structures and control gains.

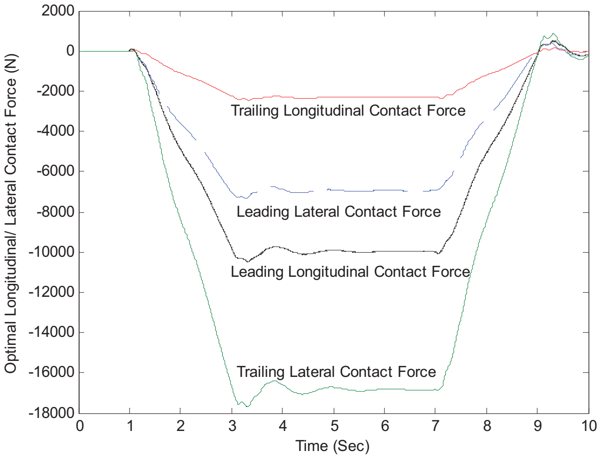


Figure 6: Original controller optimal controller /Leading Fail Hard

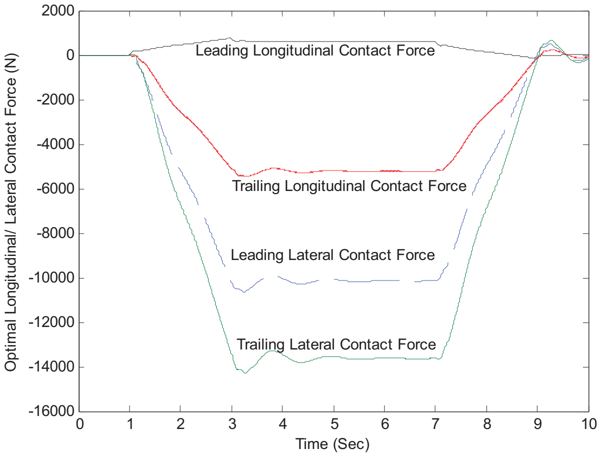


Figure 7: original optimal controller/Trailing Fail Hard

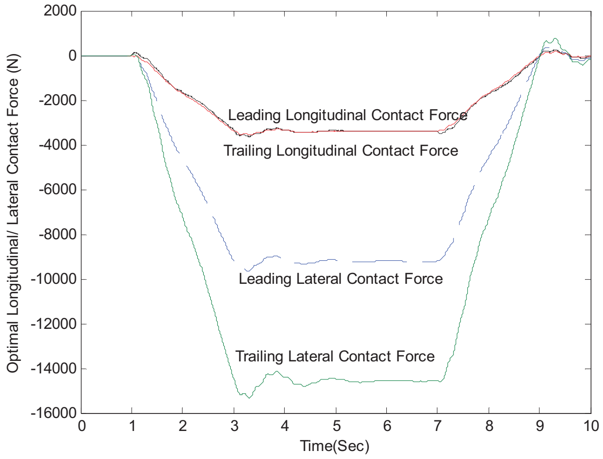


Figure 8: Retune manually optimal controller/ Leading fail Hard

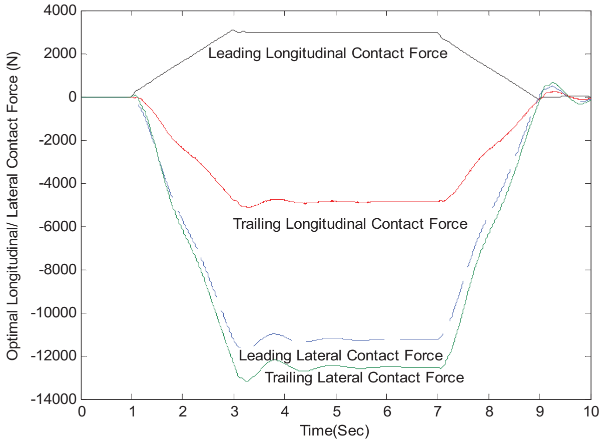


Figure 9: Retune manually Optimal Controller/ Trailing Fail Hard

## B. *Fail-Soft*

The second type of the actuator malfunction is known as fail-soft, which is when one of the actuator is unconstraint from its control input. Control torque for the failed actuator in this scenario is zero and therefore not able to stabilize the kinematic mode of the wheelset. Figs.10a and 10b compares with the stability of the bogie with different active control schemes when one of the actuators fail-soft [3]. In Figs.10a and 10b, the critical speed is reduced to approximately 10m/s for all three active control schemes in the event of an actuator fault. Clearly, in the event of fail-soft condition the active controller would not able to stabilize the system and the operation speed of the system will have to be reduced very quickly to a very low level to avoid potentially dangerous situation if no other corrective actions are taken [3].

Therefore, in this scenario, the priority of fault tolerance for the active control systems is to preserve stability control in the presence of actuator fail-soft, with the curving performance a secondary design issue.

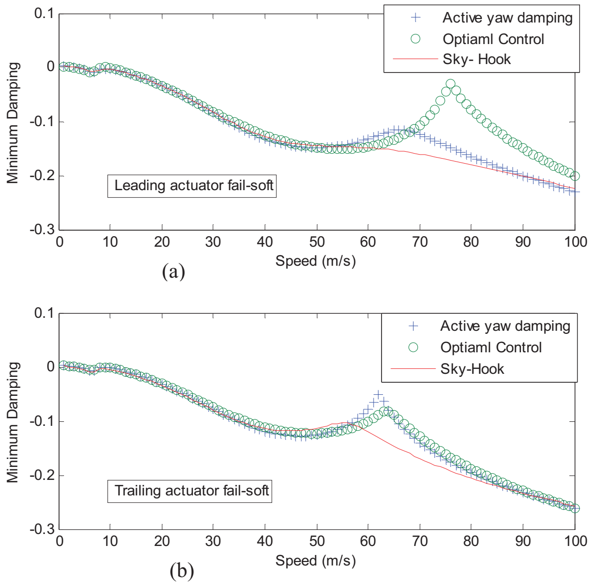


Figure 10: Stability of the different controller with leading/trailing actuator

In the event of fail-soft, the bogie stability is guaranteed if the number of control input is reduced from two actuators to one, and therefore more feedbacks are available for more sophisticated control design to provide desired control torque through the remaining actuator [3]. Fig.11 clearly reveals the bogie stability across a wide range of speed through re-designing of the optimal controller with one control input. It is also necessary to assess performance of redesign controller around the curve. Fig.12 gives the lateral and longitudinal contact forces at the wheel-rail contact points for the leading and trailing wheelset in the event of fail soft at the trailing actuator. The simulation result indicates clearly when the controller is re-designed with one control input (representing the remaining actuator), the perfect steering condition is achieved on the curved track, where the longitudinal contact forces at both leading and trailing wheelset are zero and the lateral contact forces of the two wheelsets are equal [13].

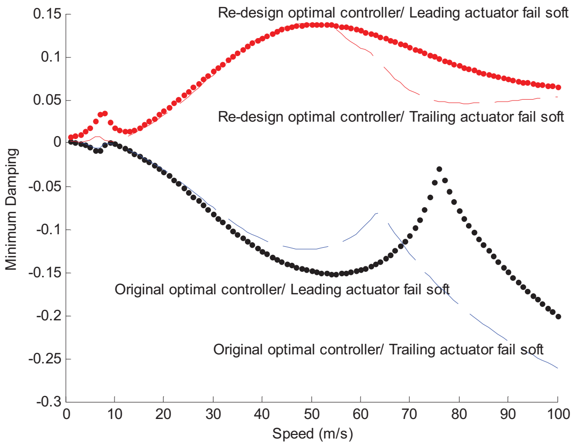


Figure 11: Stability of original / re-design optimal controller with leading/trailing fail soft in actuator

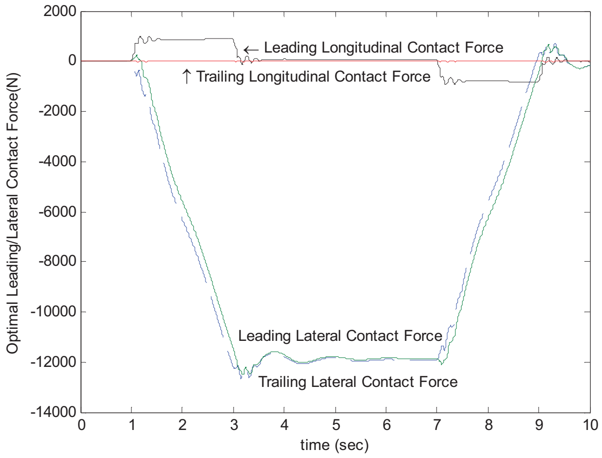


Figure 12: Curving performance of re-design optimal controller with trailing fail-soft

## V. CONCLUTION AND FUTURE WORK

This paper has studied the fundamental fault tolerant control issues for actively controlled railway vehicles through analytical redundancy to guarantee controllability of the system in the event of actuator fault. A reconfiguration based strategy for managing both soft and hard faults has been investigated, focusing on solving instability and curving performance issues respectively. The design reconfiguration controller has been evaluated by their performance capability as evidenced in simulation results.

Research is ongoing to develop optimization technique to search for the best control gain and control structure in the event of fail-hard in such a way that ensure both stability and curving performance. However, there is clearly scope for extending the work other failure modes with different actuator configurations.

## REFERENCES

[1] C. Esveld, Modern Railway Track, Delf University of Technology: BV Amesterdam, 2001.

[2] R. Goodall and H. Li, “Modelling and analysis of a railway wheelset for active control,” in UK Control, Swansea, 1998,September.

[3] T.X. Mei, “A Study of Fault Tolerance For Active Wheelset Control,” in 22 nd IAVSD, Manchester,UK, 2011.

[4] Q. Zhao and J. Jin, “Realiable Tracking Control System Design Against Actuator Failures,” in SICE 97, Tokushima,Japan, 1997.

[5] H. G. Guo, Automative Informaticand communicative system, London, UK: IGI Global, 2009.

[6] J. Pearson, R. Goodall, T.X. Mei and G. Himmelstein, “Active Stability Control Stategies For A High Speed Bogie,” Science Direct, vol. 12, no. Control Engineering Practice, pp. 1381-1391, 2004.

[7] T.X. Mei and H. Li, “Control Design for the Active Stabilization of Rail Wheelset,” Dynamic, System, Measurement and Control, vol. 130, no. 1, pp. 011002 - 011011, Jan 2008.

[8] T.X. Mei and R. Goodall, “Wheelset control strategies for a 2-axle railway vehicle,” Vehicle System Dynamics, vol. 33, pp. 653-664, 2000.

[9] S. Bruni, R. Goodall and T.X. Mei, “Control and monitoring for railway vehicle dynamics,” International Journal of Vehicle Mechanics and, vol. 45, p. 743–779, 2007.

[10] P. Aknin, J. Ayasse and A. Devallez, “Active Steering of Railway Wheelsets,” in 12th IAVSD Conference, Lyon, France., 1991.

[11] R. Iserman, “Model_Based Fault Detection and And Diagnosis Satus and Applications,” in 16th IFAC Symposium on Automatic Control in Aerospace, Osaka, Japan, 2004.

[12] R. Goodall and T.X. Mei, “LQG and GA solutions for Active Steering of Railway Vehicles,” IEE Proceedings-Control Theory and Applications, vol. 147, no. 1, pp. 111-116, 2000.

[13] S. Shen, T. Mei, R. Goodall, J. Pearson and G. Himmelstein, “A syudy of active steering strategies for a railway bogie,” in IAVSD, Kanagawa, Japan, 2003.

## APPENDIX A

***Vehicle symbol and parameter in the simulation***

| ***Symbols*** | ***Parameters*** |
|---|---|
| $y_{w_1}, y_{w_2}, y_g, y_v$ | Lateral diplacement of leading, trailing wheelset, bogie frame and vehicle body |
| $\psi_{w_1}, \psi_{w_2}, \psi_g$ | yaw diplacemet of leading, trailing and bogie fram |
| $V_s$ | Vehicle forward speed (50m/s) |
| $m_w, I_w$ | wheelset mass (1250 kg) and yaw inertia (700 kgm$^2$), respectively |
| $I_g, I_v$ | Half guage of wheelset(0.7 m), half spacing of axle(1.225 m) |
| $r_0, \lambda$ | wheel radius(0.45 m), and conicity (0.2) |
| $m_g, I_g$ | Bogie frame mass(6945 kg), and Yaw inertia (3153 kgm$^2$) respectively. |
| $K_{sc}, C_{sc}$ | Secondary Lateral and longitudinal stiffness (511 kNm$^{-1}$), and damping(37 kNsm$^{-1}$) respectively. |
| $K_s, C_s$ | primary Lateral stiffness (4750 kNm$^{-1}$), and damping (7705 N sm$^{-1}$ ) respectively. |
| $m_v$ | Half vehicle mass (15000kg) |
| $f_{11}, f_{22}$ | Longitudinal and lateral creepage coefficient (10MN). |
| $R_1, R_2$ | Radius of the curved track at the leading and trailing Wheeslet(1250 m). |
| $\theta_{c_1}, \theta_{c_2}$ | cant angle of the curved track at the leading and trailing wheelset (6$^\circ$) |
| $y_{t_1}, y_{t_2}$ | Track lateral diplacement for leading and trailing wheelsets, respectively |
| $\tau_{w_1}, \tau_{w_2}$ | Controlled torque for leading and trailing weelset respectively. |
| $g$ | Gravity (9.8 m/s$^2$) |

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Step change improvements in high-temperature thermocouple thermometry

Professor Graham Machin
Head, Temperature Standards
Engineering Measurement Division
National Physical Laboratory
Hampton Road, Teddington, Middlesex, TW11 0LW, UK
graham.machin@npl.co.uk

Dr Jonathan Pearce
Engineering Measurement Division
National Physical Laboratory
Hampton Road, Teddington, Middlesex, TW11 0LW, UK
jonathan.pearce@npl.co.uk

*Abstract*— **Thermocouples are the most widely used and most widely misunderstood temperature sensor in industry. This paper will begin with an introduction to thermocouples, how they work and the different types of thermocouple. The commonest and most misunderstood source of uncertainty, thermoelectric homogeneity, will be described. The paper will then discuss the concept of measurement traceability and new calibration methods that can give <1 °C uncertainty up to 1500 °C. The paper will conclude by describing some thermocouple innovations; a) industrial versions of pure thermoelement types (Pt/Pd) and b) self-validating "smart-sensors" which enable thermocouples to remain in calibration even when operating at extreme temperatures e.g. tungsten rhenium (W-Re) refractory metal types (e.g. type C) above 2200 °C.**

***Keywords- energy efficiency, industrial process, process control, smart sensors, high temperatures, thermocouples, calibration, traceability, uncertainty***

## I. Introduction

Thermocouples are used extensively in industry for process control. The reliable manufacture of, for example, iron and steel, glass, semiconductors, carbon and carbon composites, ceramics, refractory metals and nuclear fuel is dependent upon thermocouples. However thermocouples drift and degrade with use. If this drift can be ameliorated or reduced then the processes would run more optimally, so reducing energy consumption and improving product quality. New developments in thermocouples in particular, improved calibration, a new thermocouple type and approaches to self-validation at very high temperatures will lead to improved thermometry for more effective industrial production.

Most thermocouples in common use are identified by a letter designation. Thermocouples for industrial process control at high temperatures fall broadly into three categories, dependent upon the wires used for the thermoelements. Below 1300 °C base metal (nickel alloy) type K and increasingly, due to better stability, type N are used. Above those temperatures and up to about 1700 °C noble metal thermocouples (based on Pt and/or Pt/Rh alloys) are used; these are designated type S, R and B. Then finally, at still higher temperatures refractory metal thermocouples (based on alloys of tungsten (W) and Rhenium (Re)) are used up to approximately 2300 °C.

The foundation of reliable temperature measurement is traceability to an internationally recognized temperature scale. This is currently the International Temperature Scale of 1990 (ITS-90) [1]. All temperature measurement should be traceable to ITS-90 to ensure equivalence. This means that sensors should be calibrated to this temperature scale. In practice (and in particular) base metal thermocouples are manufactured to a specified tolerance and should, at least when new, give reliable temperature measurement. The wires or cables are batch tested before thermocouples are manufactured to ensure that they meet the required specification. Calibration is generally performed against reference thermometers in an isothermal calibration environment. The noble metal thermocouples, which are often the reference sensors, are calibrated against fixed points of known temperature. Recent innovations in the approach to thermocouple calibration, and in particular the advent of completely new fixed points, have meant that lower uncertainty and more reliable calibrations can be performed – at least to 1500 °C - for noble metal thermocouples.

In addition, it is known that type S, R and B thermocouples drift due to the alloy nature of one or both of the thermoelements. The recently developed pure thermoelement robust thermocouple based on platinum and palladium wires (the Pt/Pd thermocouple) is not subject to this source of drift and can give considerably more reliable results than the traditional noble metal thermocouples to 1500 °C.

At higher temperatures, and in particular above 1700 °C, refractory metal thermocouples (based on alloys of W and Re) are used for process control. However this type of thermocouple can experience significant drift in use leading to non-optimal control. Very recent work, particularly at NPL, has led to the development of self-validation where thermocouples such as these can be continually calibrated in process leading to a factor 10 or more improvement in their in-use performance.

Beginning with a short introduction to the temperature scale and the calibration of thermocouples this paper will describe how new high temperature fixed points (HTFPs) have improved thermocouple calibration to 1500 °C, how the

978-1-4673-1560-9/12/$31.00 ©2012 Crown

innovation of robust Pt/Pd thermocouples opens up the possibility of better temperature sensing and control to that temperature and how the implementation of self-validation at very high temperatures will make a step change improvement in the practice of high temperature contact thermometry.

## II. The International Temperature Scale of 1990 (ITS-90) and Calibration

### *A. The ITS-90*

The ITS-90 is the current temperature scale in use throughout the world, and all reliable temperature measurement is established through traceability to it. The ITS-90 ranges from 0.65 K above absolute zero, or -272.50 °C, to the highest temperatures (>3000 °C) for practical thermometry. The purpose of the ITS-90 is to define procedures by which certain specified practical thermometers can be calibrated in such a way that the values of temperature obtained from them are precise and reproducible, while at the same time approximating the corresponding thermodynamic (ideal) values as closely as possible. The "specified practical thermometers" are platinum resistance thermometers (PRT) from about 13.8 K (-259.4 °C) to the freezing point of pure silver (961.78 ºC) and radiation thermometers at higher temperatures. These thermometers are calibrated against a series of fixed points which are generally either triple points (e.g. argon, mercury, water), freezing points (e.g. silver, zinc, tin, indium) and in one case (gallium) a melting point. These have defined temperatures, formally with zero uncertainties. Below the silver point a calibration essentially determines the resistance ratio of the PRT at a specified set of fixed points relative to the water triple point. At and above the silver point the ITS-90 is established by non-contact thermometry using Planck's law in ratio form. A blackbody of either Ag, Au or Cu is used to provide the foundation for the ITS-90 in that range. The interested reader should consult references [1, 2] for further information about the ITS-90.

### *B. The calibration of thermocouples*

Calibration as applied to temperature measurement can be defined as *'the set of operations which establish, under specified conditions, the relationship between values indicated by a thermometer and the corresponding known values of temperature'* adapted from [3].

There are two principal types of calibration: comparison calibration, and fixed-point calibration. In a comparison calibration, the environment may be a variable-temperature bath or furnace which can be controlled at the desired uniform temperature. The thermometers to be calibrated are inserted into the environment along with standard thermometers which are calibrated in terms of the ITS-90, and measurements are made in an appropriate sequence. In this way the readings of the thermometers under calibration are related to the temperature as determined by the standards.

Fixed-point calibrations are generally more precise and demanding than comparison calibrations. Typically the fixed-point temperature is given by the freezing points (except gallium where the melting point is used) of pure materials, whose temperatures are defined in the ITS-90 (see Section II.*A* above). Generally the metallic fixed-points are in a state of near-equilibrium between the liquid and solid phases of a pure substance at a pressure of one standard atmosphere. The only exception to this is the mercury fixed-point where the triple point is realised. The freezing point technique requires heating the cell to about 5 to 10 °C above the melting temperature, insertion of a monitoring thermometer which indicates when the metal ingot is completely melted. The ingot is then slowly cooled while continuously logging the output of the monitoring thermometer. After an initial supercool the temperature rises to the freezing-point. The calibration measurements are made in the ensuing flat portion, or plateau, of the freezing curve, so giving the calibration of the thermometer at the freezing point. Typical fixed point materials used in the calibration of thermocouples are zinc (419.527 °C), aluminium (660.323 °C), silver (961.78 °C) and copper (1084.62 °C).

A discussion of what thermocouples are and how they work is deferred until Section IV.

## III. High Temperature Fixed Points, HTFPs

For the purpose of clarity HTFPs are those that operate above the freezing point of Cu (1084.62 °C) and are based on various types of binary alloys of metal (or metal-carbide) and carbon [4, and references therein]. These are new fixed points, under development in the last 10 years, and as such they have not yet had formal temperatures assigned to them. Table 1 lists the reported fixed-points to date, with their nominal transition temperatures.

The performance of HTFPs has been well characterised, particularly the repeatability and the reproducibility of their realised temperatures. A typical melt and freeze curve for a HTFP (Co-C) is given in Fig. 1.

Repeatability is a measure of how repeatable the temperature of an individual HTFP is when realized in quasi-identical conditions. That is, if a HTFP was melted in a furnace, then immediately frozen and melted again in the same furnace, repeatability is a measure of the agreement between the two melt temperatures. It is clear that individual HTFPs are highly repeatable, of the order of 0.02 °C even at the highest temperatures [4].

Reproducibility is a more exacting criterion than repeatability. It is a measure of the agreement in the melting temperature of two samples of HTFP material, especially if manufactured by two different suppliers. Good reproducibility is attainable when impurities in the metal of the ingot are carefully controlled. For example it has been shown that for Co-C, Pt-C and Re-C reproducibility of better than 0.1 °C can easily be achieved [5].

TABLE 1: METAL-CARBON, METAL-CARBIDE-CARBON AND METAL-CARBIDE-CARBON PERITECTICS WITH NOMINAL TEMPERATURES [6]

| Eutectic | Approximate temperature / K | Approximate temperature/ ºC |
|---|---|---|
| **Metal-carbon** | | |
| Fe-C | 1426 | 1153 |
| Co-C | 1597 | 1324 |
| Ni-C | 1602 | 1329 |
| Pd-C | 1765 | 1492 |
| Rh-C | 1930 | 1657 |
| Pt-C | 2011 | 1738 |
| Ru-C | 2227 | 1954 |
| Ir-C | 2565 | 2292 |
| Re-C | 2747 | 2474 |
| **Metal-carbide-carbon** | | |
| $B_4C$-C | 2659 | 2386 |
| δ(MoC)-C | 2856 | 2583 |
| TiC-C | 3032 | 2759 |
| ZrC-C | 3155 | 2882 |
| HfC-C | 3458 | 3185 |
| **Metal-carbide-carbon peritectic** | | |
| $Mn_7C_3$-C | 1604 | 1331 |
| $Cr_3C_2$-C | 2099 | 1826 |
| WC-C | 3022 | 2749 |

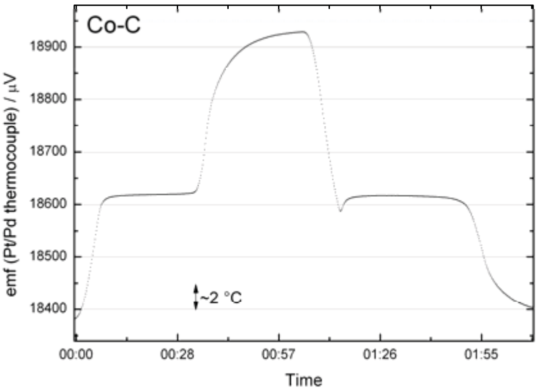


Figure 1: Melting and freezing curve of an NPL Co-C point as measured by a Pt/Pd thermocouple. The x-axis is time in hours:minutes..

The main use of HTFPs to date has been comparing high temperature scales for radiation thermometry [e.g. 5] and the calibration of thermocouples to 1500 °C. NPL now has fully implemented ISO 17025 calibration of noble metal thermocouples using HTFP ingots of Co-C and Pd-C. The accredited uncertainty of noble metal thermocouple calibrations is 0.5 °C (*k*=2) at 1324 °C (the Co-C point) and 0.7 °C (*k*=2) at 1492 °C (the Pd-C point). To obtain the lowest uncertainty calibration for industrial customers accredited laboratories are beginning to implement HTFPs within their calibration process as well [7].

## IV. THERMOCOUPLES

In this section we describe briefly what thermocouples are, how they work and the various types. The common and often largest source of uncertainty, that due to inhomogeneity is discussed. Then the pure thermoelement thermocouple based on Pt and Pd will be introduced and its superior performance described.

### *A. Thermocouples what they are and how they work*

Thermocouples are made of two dissimilar wires joined together at the measurement junction. As the wires pass through a region where the temperature changes (i.e. where there is a temperature gradient) they generate an emf, or thermovoltage. This is the Seebeck effect, named after its discoverer, and its magnitude, typically about 10-40 μV/°C, depends on the Seebeck coefficients of the thermocouple wires and the temperature difference along the total length. The emf can be measured by connecting the wires to a voltmeter. However, it depends on the temperature at the cold end, so for the most accurate measurements the wires are connected to copper leads and placed in melting ice to form the 'reference junction', and the copper wires are then connected to the voltmeter. Fig. 2 is a schematic circuit diagram of a thermocouple of wires *a* and *b* measuring temperature $t_1$, with reference junctions connected to copper wires at temperature $t_{ref}$, generating an emf , *E*.

It is important to remember that the emf is not generated at the measurement junction but in the region where the thermocouple experiences a temperature gradient. So for example when a thermocouple is being calibrated in a furnace most of the emf is generated in the region where the thermocouple emerges from the furnace, not in the isothermal part where the measuring junction is [8].

This is such an important point it is worth discussing in detail. The emf generated at any point along the wire can be written as the product of the Seebeck coefficient, *S(t)* and the temperature gradient at that point *dt*. The total emf is found by adding the emfs *S(t)dt* generated at all points along the length, starting from the reference junction and continuing to the measuring junction. Mathematically this is expressed as integrating *S(t)dt* from $t_{ref}$ to $t_1$. As what is measured is the difference between the emfs generated in the two wires of the thermocouple, with different Seebeck coefficients $S_a$ and $S_b$, the total emf, *E*, is the difference between two integrals (1):

$$E = \int_{t_{ref}}^{t_1} S_a dt - \int_{t_{ref}}^{t_1} S_b dt = \int_{t_{ref}}^{t_1} (S_a - S_b) dt \qquad (1)$$

In the simple case where the two Seebeck coefficients are independent of temperature, the equation reduces to (2)

$$E = (S_a - S_b)(t_1 - t_{ref}) \qquad (2)$$

so the emf is just the product of the difference in the two Seebeck coefficients and the temperature difference between the ends of the wires. This does not mean that the emfs are generated at the junctions but, as we have emphasized, the two junction temperatures are just the start and finish of the integration, or summation of emfs in the circuit.

Note that in the circuit of Fig. 2, the two copper wires, *c*, ideally produce equal and opposite emfs, so they do not contribute to the total result. Note also that, rather than use ice, these days an automatic electronic ice-point is often used. More commonly, however, the thermocouple wires are directly connected to instrumentation that has automatic ‘cold-junction compensation’.

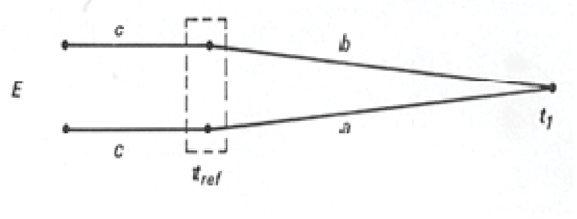


Figure 2: Schematic circuit diagram of a thermocouple being used to measure $t_1$, from [9]. The letters *a* and *b* represent two different types of thermoelement, letter *c* represents the copper wires and $t_{ref}$ the reference temperature – usually the ice point.

Provided that the junctions are isothermal, i.e. that there are no temperature gradients near them, the way in which they are made is not important; they can be twisted, clamped, soldered or welded together. The important point is that they must be isothermal and sufficiently robust to keep the wires connected together thus completing the circuit.

### B. *Types of thermocouples*

There are many types of thermocouple, and several are standardized and recognized by a letter designation. They also fall into three broad categories depending upon the metal the thermoelements are made from; these are base metal, noble (or precious) metal and refractory metal.

In base metal types for relatively low temperature applications, designated type T, J and E, the thermoelements are made of copper or copper-nickel alloys. For example the type T has a positive thermoelement of pure Cu and a negative thermoelement made of copper alloyed with about 45% nickel. This thermocouple is known as the copper-constantan thermocouple. Base metal types for higher temperature applications, up to 1300 °C, are designated types K and N. Here the thermoelements are based on nickel alloys. The type N positive thermoelement is Nicrosil from an alloy of Ni, Cr and Si, the negative thermoelement is Nisil, made from an alloy of Ni, Si and Mg.

Noble metal thermocouples (generally for high temperature applications) can be used continuously in oxidizing atmospheres up to about 1600 °C. They are designated types S, R and B and the thermoelements are based on pure platinum and varying proportions of Pt-Rh alloy. For example the type R thermocouple has a positive thermoelement of Pt alloyed with 13%Rh and a negative thermoelement made of pure Pt. There is also the pure thermoelement thermocouple Pt/Pd (see Section IV:*D* below) which has superior performance to other the noble metal alloy types, in particular circumstances, but its top temperature is limited to 1500 °C.

Refractory metal thermocouples are generally used above 1300 °C where noble metal thermocouples are too expensive for the application, and above 1600 °C to approximately 2300 °C where no other thermocouples are available for use. They are designated types C, G and D and the thermoelements are constructed from tungsten and/or various alloys of tungsten and rhenium. The most common refractory metal thermocouple, type C, is made from W5%Re (positive thermoelement) and W26%Re (negative thermoelement).

Thermocouples are manufactured to different tolerance classes depending on how closely they match the specified emf. The best is tolerance class I ranging to tolerance class III for some thermocouples. More details can be found elsewhere [10] but for illustrative purposes a tolerance class I type R thermocouple has a tolerance value of ±1 °C in the temperature range 0 °C to +1100 °C.

The purpose of this section was to give an introduction to the operation of thermocouples and the standard thermocouple types. More information can be found in classic texts such as Kinzie [8] or at [9, 10, 11]. The rest of this section will describe a common source of uncertainty, often the largest, and often not identified, the thermoelectric inhomogeneity and will conclude by describing the construction and performance of the non-letter designated Pt/Pd thermocouple.

### C. *Uncertainty arising from thermoelectric inhomogeneity*

As discussed in Section IV:*A* the thermovoltage from a thermocouple is generated where it experiences temperature gradients. In industrial applications this means that the thermovoltage is usually generated when the thermocouple passes through a furnace wall to the outside environment.

The manufacture of thermocouple wires is carefully controlled so that when new the wires are homogenous (uniform) in alloy and (particularly noble metal types) are generally annealed so that strain is removed. However in use several things happen to disturb the freshly constructed state. Examples are; non-uniform contamination of the wires from the external environment, movement of the alloying materials within the wires and crystal growth within the wires. These effects mean that after use the wires are no longer homogeneous in composition, and the effect of this is magnified when used at high temperatures.

The consequence of the wire being non-uniform in composition or crystal structure (or both) is that the Seebeck coefficient varies with position along the wire. This means that if the affected part of the thermocouple is where the temperature gradient is, the generated thermovoltage will be in error with a consequent error in the measured temperature. This effect can be quite severe, in base and refractory metal thermocouples, >10 °C errors are not uncommon.

Finally the problem of inhomogeneity means that it is not appropriate to remove base metal thermocouples for re-calibration. Unless the thermocouple experiences the same thermal environment as in use, a recalibration will fail to properly account for inhomogeneities. To effect a proper re-calibration the thermocouple should be calibrated *in-situ* with a freshly calibrated thermocouple brought alongside in the same thermal environment. However it is usually more cost effective to replace the used thermocouple with a new one of the same type and tolerance classification according to a set maintenance schedule.

### D. Pt/Pd thermocouples

A relative newcomer is a thermocouple based on thermoelements of pure platinum and palladium, the Pt/Pd thermocouple [12]. Although this has been around for some time it has not been widely used outside the laboratory environment because the measurement junction has up to now been made using a stress relieving coil of very fine platinum wire (typically 0.1-0.2 mm diameter) to accommodate the dissimilar thermal expansivities of the thermoelements. If this had not been allowed for, unacceptable strain would have been introduced into the thermocouple leading to a failure of the measurement junction. The introduction of the coil was gave access to the full performance of the thermocouple but its drawback was that it made the thermometer very fragile and unsuited to industrial applications.

Recent innovations in design, without the stress-relieving coil, have been trialed and demonstrate similar performance to the original versions [e.g. 13]. This development opens up very interesting application possibilities for these sensors that were not possible previously. It can be envisaged that in the near term these robust thermocouple versions will act as scale transfer devices from National Measurement Institutes to accredited laboratories – but with lower uncertainties than are possible with current type S, R or B thermocouples. It can also be envisaged that they will be used as the high temperature scale comparison device of choice between NMIs to 1500 °C. In the longer term it is envisaged that use of Pt/Pd thermocouples in industry may well become widespread as the lower sensor cost and improved process control overcome industries' inertia to change.

To take full advantage of the performance of this thermocouple a means had to be found to calibrate it at the highest temperature of operation (about 1500 °C). Until HTFPs became available this was not possible because there were no fixed points that could be used, but, with the advent of fixed points of Co-C (1324 °C) and Pd-C (1492 °C) Pt/Pd thermocouples can now be calibrated with low uncertainties over their whole operating range, as described in the preceding section, with remarkably low uncertainties.

A study of robust Pt/Pd thermocouples in an industrial laboratory environment showed that they exhibited no significant drift, compared to type R thermocouples, from their as-new calibrated state, even when experiencing over 300 hours of high temperature exposure (to 1300 °C) [14].

## V. SELF VALIDATION OF THERMOCOUPLES AT VERY HIGH TEMPERATURES

It is well known that thermocouples drift significantly at high temperatures. For example Fig. 3 shows the drift of a W-Re thermocouple (type C) when exposed to temperatures near 2000 °C over a number of hours [15]. There is a dramatic shift of about 90 °C in the first 5 hours but even after that large shift there is still an erratic output – e.g. the step upwards in output of about 20 °C at about 30 hours.

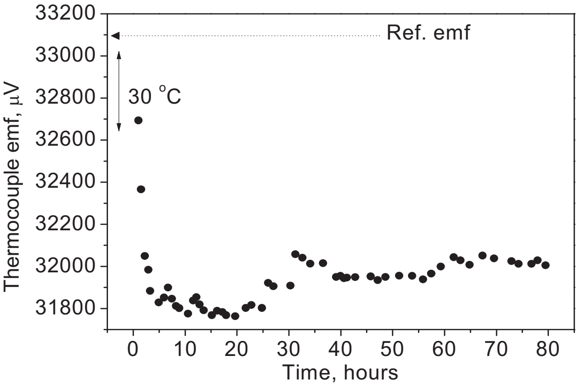


Figure 3. The emf of a new type C thermocouple as a function of time at temperature > 1950 °C. An emf shift is about 90 °C after about only 5 h is clearly observable.

This large and often undetected drift can be very detrimental to the control of an industrial process, particularly in terms of energy efficiency and product quality. The self-validation of temperature sensors has been implemented at lower temperatures for a number of years [16], but this was not really possible at higher temperatures until the advent of HTFPs. Here we introduce the principles of self-validation and then how this is implemented in practice with W-Re thermocouples.

### A. The principle of self-validation

Refractory metal thermocouples are the only practical contact temperature sensor above 1700 °C. However there are several problems with their implementation that currently limits their attainable uncertainty. When new the standard tolerance of such thermocouples is 1%, that is 20 °C at 2000 °C [17]. Removal for recalibration is not possible for two reasons. Firstly on use the thermoelements become very brittle: if the sensor was then removed it would almost certainly break. Secondly when used the W-Re thermocouple experiences strong inhomogeneity growth rendering, as described above, any calibration except *in-situ* calibration of little or no use.

It is clear that the W-Re thermocouple is beset with major performance limitations. One way to overcome all of these is to implement self-validation. That is to incorporate within the thermocouple a reference of known temperature. Then every time the sensor passes through the melting point of the temperature reference its thermovoltage is calibrated allowing the whole output of the thermocouple to be normalized. In this way the thermocouple performance is optimized even at the very highest temperatures.

### *B. Application of self-validation to high temperature thermocouples*

Self-validation of high temperature thermocouples requires the application of high temperature stable references. The HTFPs described in earlier sections are ideal for this application. Miniature fixed points of Co-C, Pt-C, Ru-C and Ir-C have been produced; these are shown in Fig. 4.

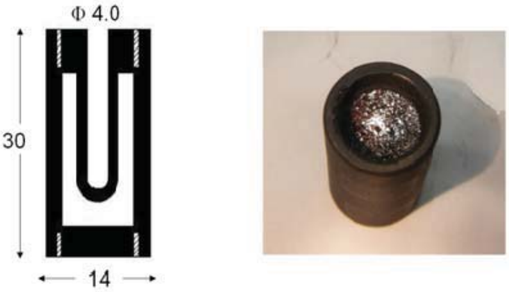


Figure 4: Cross-sectional diagram of a mini HTFP ingot used in self-validation studies of W-Re thermocouples [15].

These miniature HTFP ingots are then incorporated with the measurement junction of a metal-sheathed (usually tantalum) W-Re thermocouple (Fig. 5).

Figure 5: Incorporation of mini HTFP ingot with a metal sheathed tungsten rhenium (W-Re) thermocouple [15].

The test thermocouples are then placed in a furnace and raised in temperature to the melting point of the HTFP ingot. Each time the melting and freezing plateau was observed. For illustrative purposes see Fig. 6, but similar results have been obtained for all the HTFPs tested (up to the Ir-C point, 2292 °C) [18].

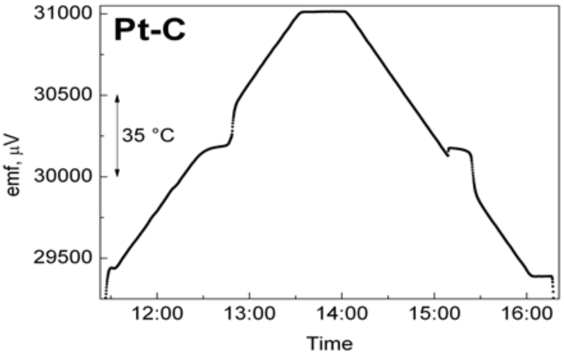


Figure 6: Typical Pt-C melt and freeze of the miniature HTFP (nominally 1738 °C), measured with a type C thermocouple [18].

It is clear that this transition can be used to normalize the thermocouple output and so maintain optimum control of a process. It has been shown [15, 18] that by implementing self-validation for type C thermocouples, instead of having to assign the manufacturing tolerance of 1% of temperature, the uncertainty could be about 2 °C – a potential factor of ten improvement in performance.

It should be noted that type C thermocouples generally operate in a vacuum or inert environment so there are no compatibility issues with the HTFP graphite crucibles, which are also compatible with the thermocouple tantalum sheathes.

In this section we have shown the possibility of improving high temperature thermometry in industry through proof of concept studies of self-validating W-Re thermocouples at high temperatures. However it is well known that there are a wide variety of unaddressed high temperature measurement problems in industry. Some of these are being tackled in a joint European Metrology Research Programme[1] (EMRP) project discussed briefly in the next section.

## VI. THE HITEMS PROJECT

HiTeMS, High Temperature Metrology for Industrial Applications, is a three-year, fifteen partner, research project part funded by the EMRP [19]. The objective of the research is to develop a suite of methods and techniques that will significantly improve the practice of industrial high temperature non-contact and contact thermometry, up to at least 2500 °C. Special emphasis is given to facilitating *in-situ* traceability, i.e. ensuring traceability to the International Temperature Scale of 1990 (ITS-90) directly within the industrial process. This is often lost when industrial thermometry is performed due to (for example) unknown sensor drift or (for radiation thermometers viewing through windows) unquantified transmission changes.

The following outstanding problems with high temperature measurement will be addressed by this research:

Non-contact thermometry:

[1] Information about the EMRP can be found at: http://www.euramet.org/index.php?id=publicity

- Emissivity and reflected radiation, with the target of achieving *in-situ* traceability

- Corrections for varying window/path transmission, to approximately 2500 ºC

- Real time traceable temperature measurement in laser materials processing

Contact thermometry:

- Lifetime assessment of base metal and drift measurements of base and noble metal thermocouples

- Self-validation and demonstrated *in-situ* validation for temperature sensors to at least 2000 ºC

- Facility for determination of reliable reference functions for high temperature non-standard thermocouples (demonstrated by determining a better reference function for the Ir-60%Rh/Ir thermocouple)

HiTeMS started in September 2011 and will conclude in Aug 2014. It is anticipated that by the end of the project many of the common problems encountered in industrial high temperature thermometry will have been addressed. More information about the project can be found at the HiTeMS website [20].

## VII. SUMMARY

An introduction to the international temperature scale of 1990 and temperature calibration has been given with special emphasis on traceability.

New high temperature fixed points were introduced and their significance in improving high temperature measurement described.

An overview of thermocouples has been given, how they work and their different types. This was followed by a description of inhomogeneity, the most common and least understood uncertainty, and the new thermocouple type based on pure thermoelements of Pt and Pd.

The use of the principle of self-validation in making step change improvements in high temperature thermometry was outlined.

Finally a brief description of the EMRP project HiTeMS was given and its chief objectives described.

In summary the developments reported here show that a step change improvement in the practice of high temperature and in particular thermocouple thermometry is underway.

## VIII. ACKNOWLEDGMENTS

The work described here was supported by the UK National Measurement Office, with additional support under a programme of, and funded by, the European Space Agency and the EMRP. The EMRP is jointly funded by the EMRP participating countries within EURAMET and the EU.

Thanks are given to Oijai Ongrai and Claire Elliott of NPL for providing some of the experimental results shown here, and Richard Rusby, of NPL, who gave valuable comments on this manuscript.

## IX. REFERENCES

[1] H. Preston-Thomas, "The International Temperature Scale of 1990, ITS90", Metrologia, 1990, **27**, pp. 3-10 and p. 127

[2] http://www.bipm.org/en/publications/its-90.html

[3] The International Vocabulary of Basic and General Terms in Metrology, ISO 1993

[4] G., Machin, "Twelve years of high temperature fixed point research: a review", Presented "The International Temperature Symposium (ITS9)", California, USA, Proceedings available from Autumn 2012.

[5] G. Machin, Dong Wei, M. J. Martín, D. Lowe, T. J. Wang, X. Lu, "A comparison of the ITS90 between, NPL, NIM and CEM above the silver point using High Temperature fixed points", Int. J. Thermophys., **31**, pp. 1466-1476, 2010

[6] G. Machin, "Realising the benefits in improvements in high temperature measurement", Acta Metrologica, Sinica, 2008, **29**, pp. 10-17

[7] J.V. Pearce, G. Machin, T. Ford, S. Wardle, "Optimising heat-treatment of gas turbine blades with a Co-C fixed-point for improved in-service thermocouples", Int. J. Thermophys., **29**, pp. 222-230, 2008

[8] Thermocouple Temperature Measurement, Kinzie, P.A., Wiley-Interscience, 1973

[9] NPL Temperature Course Notes, 2012

[10] Manual on The Use of Thermocouples in Temperature Measurement, 4th Edition, ASTM Manual Series: MNL 12, Revision of Special Technical Publication (STP) 470B (1993) , BS EN (IEC) 60584-2, *Specification for thermocouple tolerances*, BSI, 1993

[11] http://srdata.nist.gov/its90/useofdatabase/use_of_database.html

[12] W.G. Burns, D.C. Ripple, M. Battuello, "Platinum versus palladium thermocouples, an emf–temperature reference function from 0 °C to 1500 °C", Metrologia, **35**, pp. 761-780, 1998

[13] F. Edler, R. Morice, J. Pearce,, "Construction and investigation of Pt/Pd thermocouples in the framework of Euramet project 857", Int. J. Thermophys., **29**, pp. 199-209, 2008

[14] C. Elliott, J. Pearce, G. Machin, T. Ford, K. Hicks, "Pt/Pd thermocouple resilience over 327 operating hours in an industrial calibration laboratory"; Presented at "The International Temperature Symposium (ITS9)", California, USA, Proceedings available from Autumn 2012.

[15] O. Ongrai, J. Pearce, G. Machin, S. Sweene "Self-calibration of a W/Re thermocouple using a miniature Ru-C (1954 °C) eutectic cell"; Presented at "The International Temperature Symposium (ITS9)", California, USA, Proceedings available from Autumn 2012.

[16] S. Augustin, F. Bernhard, D. Boguhn, A. Donin, and H. Mammen, "Industrially applicable miniature fixed-point thermocouples," in 8th International Symposium on Temperature and Thermal Measurement in Industry and Science (TEMPMEKO 2001), Berlin, 2001, pp. 3-8.

[17] ASTM E988, "Standard temperature-electromotive force (EMF) tables for tungsten-rhenium thermocouples", ASTM International, USA

[18] J. Pearce, C. Elliott, G. Machin, O. Ongrai, "Self-validating Type C thermocouples to 2300 °C using high temperature fixed points", Presented at "The International Temperature Symposium (ITS9)", California, USA, Proceedings available from Autumn 2012.

[19] G. Machin,, K. Anhalt, F. Edler, J. Pearce, M. Sadli, R. Strnad, E. Vuelban; "HiTeMS: A project to solve high temperature measurement problems in industry", Presented at "The International Temperature Symposium (ITS9)", California, USA, Proceedings available from Autumn 2012.

[20] http://projects.npl.co.uk/hitems/

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# An Integrated Backstepping and Sliding Mode Tracking Control Algorithm for Unmanned Underwater Vehicles

Bing Sun, Daqi Zhu, Weichong Li

Laboratory of Underwater Vehicles and Intelligent Systems
Shanghai Maritime University
Shanghai, China
email:zdq367@yahoo.com.cn

*Abstract*—**In this paper, an integrated backstepping and sliding mode tracking control algorithm is developed for three-dimensional tracking control of unmanned underwater vehicles (UUV). The proposed control strategy combines with a kinematic controller and dynamic controller together. The kinematic controller integrates a bio-inspired model with the basktepping method while the dynamic controller uses robust sliding mode control. Unlike the traditional backstepping method suffering from the speed jump problem, the application of bio-inspired model can generate smooth and continuous velocity signal even in the large initial errors. Therefore, a smooth control signal can be obtained by dynamic controller without thruster control saturation. The effectiveness and efficiency of the proposed control strategy are demonstrated through simulations and comparison studies.**

***Keywords-Unmanned underwater vehicles, tracking control, backstepping, bio-inspired model.***

## I. INTRODUCTION

Unmanned underwater vehicles (UUV) have been wildly used as a platform employed in risky missions such as oceanographic observations, bathymetric surveys, ocean floor analysis, military applications, recovery of lost man-made objects, etc [1]. This requires a more precise control behavior which has motivated an intensive research in the last decade. As a consequence, several different control approaches have been applied to the UUV motion control such as the adaptive control techniques [2-3], sliding-mode control [4-6], backstepping control algorithms [7-8], fuzzy-logic and neural network methods [9-12], etc.

The underwater vehicle dynamics is strongly coupled and highly nonlinear. In order to deal with the uncertain nonlinear parts in the underwater vehicle's dynamics, many researchers concentrated their interests on the applications of sliding mode control. Sliding mode method [4-5] is usually used for dynamic tracking control for the outstanding characteristic including insensitivity to parameter variations, and good rejection of disturbances. So sliding mode control is extraordinary suitable for robust tracking control of underwater vehicle. However, one major drawback of the sliding-mode approach is the high frequency of control action (chattering). To eliminate/reduce chattering, various methods have been proposed to reach a continuous robust control. For example, S. Serdar proposed a chattering-free sliding-mode control method with an adaptive estimate term [6].

The backstepping control algorithms [7-8] is the commonly used approach for tracking control. However, the disadvantage for backstepping method is quite obvious. The velocity control law is directly related to the state errors, so large velocities will be generated in big initial error condition and sharp speed jump occurs while sudden tracking error happens. It means that the required acceleration and forces/moments exceed their control constraint even infinite values at the velocity jump points, which is practically impossible.

To resolve the impractical speed jump problem of large initial velocities resulted from the backstepping technique, some fuzzy control methods [9-10] and neural network control algorithms [11-12] are proposed. The fuzzy rules based tracking control approaches can solve the problem of large initial vehicle velocities, but it is very difficult to formulate the fuzzy rules, which are usually obtained by trial and error based human knowledge. As neural networks are characterized by flexibility and an aptitude for dealing with non-linear problems, they are envisaged to be beneficial when used on underwater vehicles. For examples, a neural network adaptive controller is proposed by Li [11] for autonomous diving control of an UUV using adaptive backstepping method when the smooth unknown dynamics of a vehicle is approximated by a neural network. Anyhow, the existing neural networks based tracking control algorithms for underwater vehicle require either on-line learning or off-line training procedures which could be computational complicated.

This paper focuses on the problem of speed jump and the thruster control constraints for UUV, and the a kinematics/dynamics cascaded control system integrating backstepping technology and sliding mode control with bio-inspired neural dynamics model (bio-inspired model) [13-14] is presented for three-dimensional (3D) tracking control of OUTLAND1000 UUV. Due to the shunting characteristics of bio-inspired model, the output of the bio-inspired model is bounded in a finite interval and smooth without any sharp jumps when inputs have sudden changes. In addition, the novel tracking controller proposed can meet on-line navigation of an UUV because no learning procedures are needed in the bio-inspired neural dynamics model.

This paper is organized as follows: In the second section, the basic kinematics and dynamics of the UUV. In the third section, an integrating backstepping and sliding mode tracking control algorithm based on bio-inspired model is presented. In the fourth section thruster configuration and force allocation for OUTLAND1000 UUV are introduced. To illustrate effectiveness and efficiency of the proposed method, simulation examples are given in the fifth section. Finally, some concluding remarks are made.

## II. Background

In this section, two coordinate systems for UUV control are first briefly presented. Then the kinematic and dynamic model of UUV is provided. Finally the tracking control problem is briefly stated.

### A. Kinematic model

In a three-dimensional (3D) Cartesian workspace shown in Fig. 1, two coordinate frame systems are defined: the inertial frame system $\{O-XYZ\}$ and the body-fixed frame system $\{O_0-X_0Y_0Z_0\}$. The coordinate systems illustrated in Fig. 1 obey the right-hand rule and the Z-axis points to the downward.

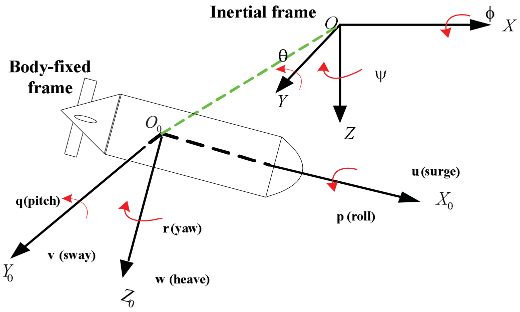


Fig. 1. Coordinate systems

The kinematic models are formulated as below. Let $\eta=[x \quad y \quad z \quad \phi \quad \theta \quad \psi]^T$ be the generalized coordinates representing the position $(x,y,z)$ and the orientation $(\phi,\theta,\psi)$ with respect to the inertial frame and $\boldsymbol{q}=[u \quad v \quad w \quad p \quad q \quad r]^T$ be the translational velocities $(u,v,w)$ and the rotational velocities $(p,q,r)$ with respect to the body frame attached at the vehicle. Then, the kinematic model relating the body-fixed frame to the inertial frame can be expressed in a compact vector form as follows[24]:

$$\dot{\eta}=J(\eta)\boldsymbol{q} \tag{1}$$

where $J\in R^{n\times n}$ is the spatial transformation matrix between the inertial frame and the UUV's body-fixed frame.

### B. Dynamic model

The dynamic equation of UUV can be presented as a compact vector form [15]:

$$\boldsymbol{M}\dot{\boldsymbol{q}}+\boldsymbol{C}(\boldsymbol{q})\boldsymbol{q}+\boldsymbol{D}(\boldsymbol{q})\boldsymbol{q}+\boldsymbol{g}(\eta)=\tau \tag{2}$$

where $\boldsymbol{M}\in R^{n\times n}$ -inertia matrix including the added mass effects ; $\boldsymbol{C}\in R^{n\times n}$ -the matrix of Coriolis and centrifugal terms (including added mass caused by hydrodynamic effect) ; $\boldsymbol{D}\in R^{n\times n}$ - the hydrodynamic damping matrix ; $\boldsymbol{g}\in R^{n}$ - gravity and buoyancy forces; $\tau\in R^{n}$ is the control forces and moments acting on the UUV centre of mass.

As mentioned earlier, the system dynamics are not exactly known. The system dynamics can be divided into two parts: estimated dynamics $\hat{\tau}$ and unknown dynamics $\tilde{\tau}$ :

$$\tau=\hat{\tau}+\tilde{\tau} \tag{3}$$

where $\hat{\tau}=\hat{\boldsymbol{M}}\dot{\boldsymbol{q}}+\hat{\boldsymbol{C}}\boldsymbol{q}+\hat{\boldsymbol{D}}\boldsymbol{q}+\hat{\boldsymbol{g}}$ , $\tilde{\tau}=\tilde{\boldsymbol{M}}\dot{\boldsymbol{q}}+\tilde{\boldsymbol{C}}\boldsymbol{q}+\tilde{\boldsymbol{D}}\boldsymbol{q}+\tilde{\boldsymbol{g}}$ , $\hat{\boldsymbol{M}},\hat{\boldsymbol{C}},\hat{\boldsymbol{D}},\hat{\boldsymbol{g}}$ are estimated terms, $\tilde{\boldsymbol{M}},\tilde{\boldsymbol{C}},\tilde{\boldsymbol{D}},\tilde{\boldsymbol{g}}$ are the unknown terms.

In this paper, the mainly research focuses on the model of OUTLAND1000 UUV. The thruster configuration and force allocation of OUTLAND1000 UUV will be described in detail in section 4. It has only four degrees of freedom (DOF) $(u,v,w,r)$ to conduct motion control, $p=q=0$ . So the simplified dynamic model in this paper's simulation is given as follows [16]:

$$(m-X_{\dot u})\dot u-mvr-X_u u-X_{uu}u|u|=\tau_X \tag{4}$$

$$(m-Y_{\dot v})\dot v-mur-Y_v v-Y_{vv}v|v|=\tau_Y \tag{5}$$

$$(m-Z_{\dot w})\dot w-Z_w w-Z_{ww}w|w|=\tau_Z \tag{6}$$

$$(I_z-N_{\dot r})\dot r+muv-mur-N_r r-N_{rr}r|r|=\tau_N \tag{7}$$

where $X_{\dot u},Y_{\dot v},Z_{\dot w},N_{\dot r}$ is the added mass effect, $X_u,Y_v,Z_w,N_r$ is the linear drag and $X_{uu},Y_{vv},Z_{ww},N_{rr}$ is the quadratic drag.

### C. Tracking control problem

The UUV is usually required to move at a low forward speed and a low rotational speed when it executes investigation tasks. This needs a precious tracking control. Consider that the UUV's major movement is in four degrees of freedom (DOF): surge, sway, heave, yaw, so in this paper, only the four DOF tracking control problem is represented. The controller design problem can be described as follows. The desired state of UUV is defined as

$$\eta_d=[x_d \quad y_d \quad z_d \quad \psi_d]^T \tag{8}$$

where $\eta_d=[x_d \quad y_d \quad z_d \quad \psi_d]^T$ is the desired state of UUV in the inertial frame, ( $x_d,y_d,z_d$ ) is coordinate of desired path in the inertial frame, $\psi_d$ is the counter-clockwise rotation angle of UUV along the Z-axis.

The desired forward and angular velocities can be deduced by

$$\begin{aligned} u_d &=\dot x_d\cos\psi_d+\dot y_d\sin\psi_d \\ v_d &=\dot x_d(-\sin\psi_d)+\dot y_d\cos\psi_d \\ w_d &=\dot z_d \\ r_d &=\dot\psi_d=\frac{\dot x_d\ddot y_d-\ddot x_d\dot y_d}{\dot x_d^2+\dot y_d^2} \end{aligned} \tag{9}$$

The actual state of UUV is represented by $\eta=[x \quad y \quad z \quad \psi]^{\mathrm{T}}$ , $\boldsymbol{q}=[u \quad v \quad w \quad r]^T$ . As the objective of the path tracking controllers is to make UUV follow the known path by controlling the velocity and angular velocities, so the tracking error $\boldsymbol{e}=\boldsymbol{\eta}_d-\boldsymbol{\eta}=[e_x \quad e_y \quad e_z \quad e_\psi]^T$ converges to zero. Here $\boldsymbol{e}$ is the tracking error in the inertial frame. A detailed model of tracking control problem is given in Fig. 2.

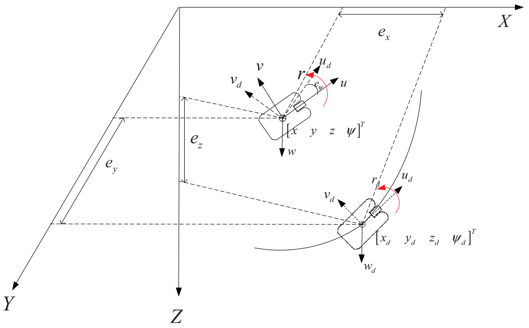

X
Y
Z
$e_x$
$e_y$
$e_z$
$v$
$v_d$
$r$
$u_d$
$u$
$e_\psi$
$[x \quad y \quad z \quad \psi]^T$
$w$
$r_d$
$u_d$
$v_d$
$[x_d \quad y_d \quad z_d \quad \psi_d]^T$
$w_d$

Fig. 2. Tracking control problem

## III. CONTROL ALGORITHMS

The basic control architecture of the system is illustrated in Fig. 3. The design of the hybrid control strategy consists of two parts: (1). an outer loop virtual velocity controller by using position and orientation state errors; (2). an inner loop sliding-mode controller by using velocity state vector.

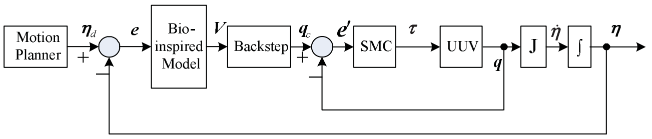


Fig. 3. The cascaded controller of UUV

### A. Virtual velocity controller

Backstepping method for nonholonomic mobile robot has been designed a lot for velocity tracking [12-13]. But the UUV in this study is a holonomic system, so the backstepping control law for the mobile robot is not fit for this control system. For this reason, a new backstepping control law is designed for UUV and makes it possible to follow a given reference posture with stability.

The virtual velocity controller based on the backstepping approach can be defined as:

$$\boldsymbol{q}_c=\begin{bmatrix}u_c\\ v_c\\ w_c\\ r_c\end{bmatrix}=\begin{bmatrix}k(e_x\cos\psi+e_y\sin\psi)+(u_d\cos e_\psi-v_d\sin e_\psi)\\ k(-e_x\sin\psi+e_y\cos\psi)+(u_d\sin e_\psi+v_d\cos e_\psi)\\ w_d+k_z e_z\\ r_d+k_\psi e_\psi\end{bmatrix} \quad (10)$$

where $k,k_z,k_\psi$ are constant coefficients, $\boldsymbol{q}_d=[u_d \quad v_d \quad w_d \quad r_d]^T$ is the desired velocity in the body-fixed frame, $u_d\cos e_\psi-v_d\sin e_\psi$ , $u_d\sin e_\psi+v_d\cos e_\psi$ represents the desired velocity frame transformed to the actual velocity frame seen in Fig. 2.

### B. Bio-inspired velocity controller

In-depth analysis of (10), the virtual speed can be found directly related to the tracking errors. In order to resolve the speed jump and control constraint problem, a bio-inspired model is added in the controller to design the virtual velocity. Bio-inspired neural dynamics model was first developed by Grossberg [17]. It can describe an on-line adaptive behavior of individuals. It was originally derived based on the membrane model proposed by Hodgkin and Huxley [18] for a patch of membrane using electrical elements. The dynamics of voltage across the membrane $V_m$ can be described in the membrane model using state equation technique as

$$C_m\frac{dV_m}{dt}=-(E_p+V_m)g_p+(E_{Na}-V_m)g_{Na}-(E_k+V_m)g_k \quad (11)$$

where $C_m$ is the membrane capacitance. The parameters $E_k$ , $E_{Na}$ and $E_p$ are the Nernst potentials for potassium ions, sodium ions, and passive leak current in the membrane, respectively. $g_k$ , $g_p$ and $g_{Na}$ are the conductance of potassium, sodium and the passive channels are functions of input signals that vary with time.

To simplify the equation, a shunting equation is obtained as

$$\dot{V}=-AV+(B-V)S(t)^+-(D+V)S(t)^- \quad (12)$$

where $V$ is the neural activity of the neuron. Parameters $A,B$ and $D$ are respectively the passive decay rate, the upper and lower bounds of the neural activity. The variables $S^+$ and $S^-$ represent the excitatory and inhibitory input, respectively. The shunting dynamic of an individual neuron can be modeled by this equation. The neutron dynamics are restricted to a bounded interval $[-D,B]$ and an automatic gain control. So we can infer the shunting equation to the following form

$$\dot{V}_i=-AV_i+(B-V_i)f(e_i)-(D+V_i)g(e_i) \quad (13)$$

where $f(e_i)=\max(e_i,0)$ , $g(e_i)=\max(-e_i,0)$ , $A,B,D$ are positive constants. For a system with appropriate chosen inputs, various desirable functional properties such as competitive, short memory, upper bound and lower bound can be derived from the model. The systems output $V$ is guaranteed to stay in a region $[-D,B]$ for any excitatory and inhibitory inputs.

So the proposed virtual velocity controller can be given as

$$\boldsymbol{q}_c=\begin{bmatrix}u_c\\ v_c\\ w_c\\ r_c\end{bmatrix}=\begin{bmatrix}k(V_x\cos\psi+V_y\sin\psi)+(u_d\cos e_\psi-v_d\sin e_\psi)\\ k(-V_x\sin\psi+V_y\cos\psi)+(u_d\sin e_\psi+v_d\cos e_\psi)\\ w_d+k_z V_z\\ r_d+k_\psi V_\psi\end{bmatrix} \quad (14)$$

where $k,k_z,k_\psi$ are the same parameters of (13). Due to the shunting characteristics of bio-inspired model, the output of the bio-inspired model is bounded in a finite interval and smooth without any sharp jumps when inputs have sudden changes. So the controller performance is significantly improved.

### C. The adaptive sliding-mode controller

After the velocity controller generates the virtual velocity of the underwater vehicles, a sliding-mode controller is used to generate the control forces and moments $\boldsymbol{\tau}=[\tau_X \quad \tau_Y \quad \tau_Z \quad \tau_N]^T$.

Then the control inputs $\boldsymbol{\tau}$ will be applied to the UUV dynamic model to produce the actual velocity in surge, sway, heave and yaw ($\boldsymbol{q}=[u \quad v \quad w \quad r]^T$) in the body-fixed frame respectively. So it will be easy to get the actual underwater vehicle's states ($\boldsymbol{\eta}=[x \quad y \quad z \quad \psi]^T$) in the inertial frame by $\dot{\boldsymbol{\eta}} = J\boldsymbol{q}$.

As a rule, sliding-mode control can be divided into two parts. First, define a sliding manifold $\boldsymbol{s}$. Second, find a control law to move toward the sliding manifold. The sliding manifold is defined as[8]:

$$\boldsymbol{s} = \dot{e}_c + 2\Lambda e_c + \Lambda^2 \int e_c \qquad (15)$$

where $\boldsymbol{e}_c = \boldsymbol{q}_c - \boldsymbol{q}$ is the velocity error between the virtual velocity and the actual velocity, $\Lambda$ represents a strictly positive constant, $\boldsymbol{s}$ is a $4\times1$ vector. Derivation of (15), then

$$\dot{\boldsymbol{s}} = \ddot{\boldsymbol{e}}_c + 2\Lambda\dot{\boldsymbol{e}}_c + \Lambda^2\boldsymbol{e}_c = \ddot{\boldsymbol{e}}_c + 2\Lambda(\dot{\boldsymbol{q}}_c - \dot{\boldsymbol{q}}) + \Lambda^2\boldsymbol{e}_c \qquad (16)$$

When the system is operating on the sliding surface, (11) equals zero, i.e.

$$\dot{\boldsymbol{s}} = \ddot{\boldsymbol{e}}_c + 2\Lambda\dot{\boldsymbol{e}}_c + \Lambda^2\boldsymbol{e}_c = \ddot{\boldsymbol{e}}_c + 2\Lambda(\dot{\boldsymbol{q}}_c - \dot{\boldsymbol{q}}) + \Lambda^2\boldsymbol{e}_c = 0 \qquad (17)$$

Substituting (2) into (17), then

$$\ddot{\boldsymbol{e}}_c + 2\Lambda(\dot{\boldsymbol{q}}_c - \boldsymbol{M}^{-1}(\boldsymbol{\tau} - \boldsymbol{Cq} - \boldsymbol{Dq} - \boldsymbol{g})) + \Lambda^2\boldsymbol{e} = 0 \qquad (18)$$

So the equivalent control law can be concluded as

$$\boldsymbol{\tau}_{eq} = \hat{\boldsymbol{M}}(\dot{\boldsymbol{q}}_c + \frac{\ddot{\boldsymbol{e}}_c}{2\Lambda} + \frac{\Lambda}{2}\boldsymbol{e}_c) + \hat{\boldsymbol{C}}\boldsymbol{q} + \hat{\boldsymbol{D}}\boldsymbol{q} + \hat{\boldsymbol{g}} \qquad (19)$$

where $\hat{\boldsymbol{M}}, \hat{\boldsymbol{C}}, \hat{\boldsymbol{D}}, \hat{\boldsymbol{g}}$ are estimated terms. Considering the difficulty of computing $\ddot{e}_c$ in (19), a feedback control input of acceleration error is introduced

$$\ddot{\boldsymbol{e}}_c = -k\dot{\boldsymbol{e}}_c \qquad (20)$$

where $k$ is a constant scalar representing the strictly positive constant that determines the rate of acceleration error.

The conventional sliding-mode can be designed as

$$\boldsymbol{\tau} = \boldsymbol{\tau}_{eq} + k\,\mathrm{sgn}(\boldsymbol{s}) \qquad (21)$$

To eliminate chattering problem caused by the discontinuous term, an adaptive term [8] is added in the control law to replace the switching term

$$\boldsymbol{\tau}_{ad} = \tilde{\boldsymbol{\tau}}_{est} + (K + \frac{\hat{\boldsymbol{C}}}{2\Lambda})\boldsymbol{s} \qquad (22)$$

where $\tilde{\boldsymbol{\tau}}_{est}$ is an adaptive term that estimates the lumped uncertainty vector defined in (3), K is also a constant scalar representing the strictly positive constant related to the convergence rate of the controller. The estimation of the lumped uncertainty vector is proposed to follow:

$$\dot{\tilde{\boldsymbol{\tau}}}_{est} = \Gamma\boldsymbol{s} \qquad (23)$$

where $\Gamma$ represent the strictly positive constant that determines the rate of adaption.

The total control law can be defined as

$$\boldsymbol{\tau} = \boldsymbol{\tau}_{eq} + \boldsymbol{\tau}_{ad} = \boldsymbol{\tau}_{eq} + \tilde{\boldsymbol{\tau}}_{est} + (K + \frac{\hat{C}}{2\Lambda})\boldsymbol{s} \qquad (24)$$

## IV. Formulation of The Control Allocation Problem

The OUTLAND1000 is used to demonstrate the performance of the proposed bio-inspired cascaded tracking control approach. It has four fixed direction thrusters; three horizontal thrusters denoted as $HT^i, i \in [1,3]$, and one vertical thruster denoted as $VT^1$. The OUTLAND1000 and the thruster's configuration enables direct control of surge, yaw and heave, as indicated in Fig. 4, 5.

In this paper the four degree motion control is discussed. Each thruster exerts thrust (force) **F** and torque **Q**. The position vector ${}^i\boldsymbol{r} = [{}^ir_x \quad {}^ir_y \quad {}^ir_z]^T$ determines the position of the point of attack of the force **F.** The orientation of the thruster is defined by the unit vector ${}^i\boldsymbol{e} = [{}^ie_x \quad {}^ie_y \quad {}^ie_z]^T$. The force **F** also generates the moment **Q =r×F.** The vector of forces and moments can be written as [19]:

$$F_i = k_1 n_i^2 = k u_i^2, n_i = k_2 u_i, Q_i = r_i \times k u_i^2, i = 1,2,3,4 \qquad (25)$$

where $\lambda = k_1 k_2^2$ is the thruster motor control parameter, $u_i$ is the control voltage of $ith$ thruster, $n_i$ is the rotational speed of $ith$ thruster.

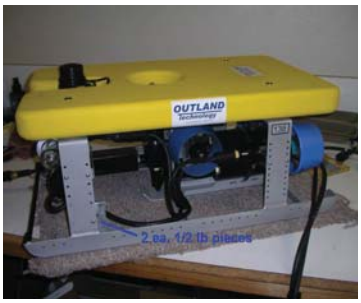


Fig. 4 OUTLAND1000

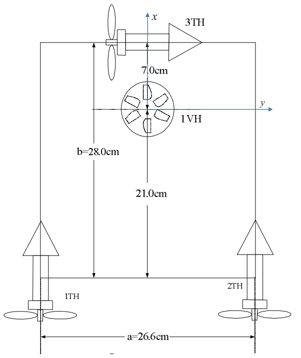


Fig. 5 Configuration of the thrusters

The total vector of propulsion forces and moments $\boldsymbol{\tau}$ in the horizontal plane are considered:

$$\tau = \begin{bmatrix} \tau_X \\ \tau_Y \\ \tau_Z \\ \tau_N \end{bmatrix} = \begin{bmatrix} \lambda & \lambda & 0 & 0 \\ 0 & 0 & \lambda & 0 \\ 0 & 0 & 0 & \lambda \\ a\lambda/2 & -a\lambda/2 & (b-21)\lambda & 0 \end{bmatrix} \begin{bmatrix} u_1^2 \\ u_2^2 \\ u_3^2 \\ u_4^2 \end{bmatrix} = B \cdot u \qquad (26)$$

where $\boldsymbol{\tau}$ are represented as the motions of surge, sway, heave and yaw respectively, and $a\lambda/2 = 13.3\lambda$ , $(b-21)\lambda = 7\lambda$ (from Fig. 5), Each component of the control vector is limited by constraint:

$$-u_m \le u_i \le u_m, -n_m \le n_i \le n_m, \; i = 1,2,3,4 \qquad (27)$$

where $u_m$ is the thruster maximum control variable, $n_m$ is the thruster maximum rotational speed . From (25) and (26):

$$\tau_{Xm} = 2F_m = 2\lambda u_m^2 \Rightarrow \lambda = 2\tau_{Ym} / u_m^2 \qquad (28)$$

$$\tau_{Ym} = F_m = \lambda u_m^2 \Rightarrow \lambda = \tau_{Ym} / u_m^2 \qquad (29)$$

$$\tau_{Zm} = F_m = \lambda u_m^2 \Rightarrow \lambda = \tau_{Ym} / u_m^2 \qquad (30)$$

$$\tau_{Nm} = a\lambda u_m^2 / 2 + (b-21)\lambda u_m^2 = 20.3\lambda u_m^2 \Rightarrow 20.3\lambda = \tau_{Nm} / u_m^2 \qquad (31)$$

Finally, the general constrained control allocation problem for the OUTLAND1000 can be formulated as: For given normalized $\overline{\tau}$ , find $\overline{u}$ feasibility such that

$$\overline{\tau} = \overline{B} \cdot \overline{u} \tag{32}$$

where $-1 \le \overline{u} = [\frac{u_1}{u_m} \quad \frac{u_2}{u_m} \quad \frac{u_3}{u_m} \quad \frac{u_4}{u_m}]^T \le 1$, $-1 \le \overline{\tau} \le 1$,

$$\overline{\tau} = [\frac{\tau_X}{\tau_{Xm}} \quad \frac{\tau_Y}{\tau_{Ym}} \quad \frac{\tau_Z}{\tau_{Zm}} \quad \frac{\tau_N}{\tau_{Nm}}]^T, \ \overline{B} = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0.655 & -0.655 & 0.345 & 0 \end{bmatrix}.$$

## V. SIMULATION

In this paper, the two methods were simulated for tracking control problem: the proposed backstepping method and the bio-inspired method. The backstepping method given in (10) was used as a case study to illustrate the performance of the proposed control strategies. From the simulation results, the proposed controller can reach a robust control while the speed jump and thruster saturation problem can be solved successfully.

The hydrodynamic parameter of OUTLAND1000 UUV is given in Table I by estimation through the comparison of the similar UUV model [16]. To reflect uncertainties of the vehicle dynamics, 20% model inaccuracies were incorporated into the controller's dynamic model. It means that in the following simulations, the parameters in Table I will be set as the actual value, while the estimated value is 80% of the actual value. The parameter setting of the cascaded controller is shown in Table II.

Table I. Hydrodynamic Parameters of OUTLAND1000 UUV

| | | |
|---|---|---|
| $X_{\dot{u}} = 34$ | $X_u$=6 | $X_{uu} = 18$ |
| $Y_{\dot{v}} = 75$ | $Y_v = 10$ | $Y_{vv} = 4$ |
| $Z_{\dot{w}} = 33$ | $Z_w = 7$ | $Z_{ww} = 4$ |
| $N_{\dot{r}} = 62$ | $N_r = 14$ | $N_r = 14$ |

A typical case to track a spiral line is studied. The UUV starts at Posture (0, 0, 1, 0), while the desired initial robot posture is (0, -2, 0, 0). Thus the initial posture error is (0, -2, -1, 0). Time varies from 0 to 20s. The desired state of UUV is $x_d(t) = 2\sin(0.5t)$ , $y_d(t) = -2\cos(0.5t)$ , $z_d(t) = 0.1t$ , $\psi_d(t) = 0.5t$ . The parameter setting of the hybrid controller is shown in Table II.

Table II. Controller Parameters

| $k_c$ | Γ | K | Λ | $k$ | $k_z$ | $k_\psi$ | A | B | D |
|---|---|---|---|---|---|---|---|---|---|
| 10 | 5 | 100 | 3 | 2 | 2 | 5 | 2 | 1 | 1 |

Fig. 6 ~ Fig. 8 shows the simulation results of the spiral line tracking. The red solid lines indicate the basktepping method results, and the blue solid lines are the bio-inspired method results. Fig. 6 gives the desired trajectory and tracking control results of backstepping method and bio-inspired method on the same condition. Fig. 7 shows the virtual velocity $\boldsymbol{q}_c$ by the backstepping method and bio-inspired method. The normalized thruster control variable $\overline{\boldsymbol{u}}$ of trajectory tracking is shown in Fig. 8. Table III is the maximum normalized control variable of each thruster.

In the simulation results of trajectory tracking in Fig. 6, it seems that for both two methods, the UUV can drive to reach and stay on the trajectory in quick response. In Fig. 7, for the virtual velocity controller based on the backstepping approach (non-biological inspired), this virtual velocity $q_c$ occurs the sharp speed jumps when tracking errors change suddenly at initial time. For example the virtual sway speed of the backstepping method jumps to -4m/s, but the bio-inspired method is limited in the range of -0.5m/s~0 m/s in Fig. 7.

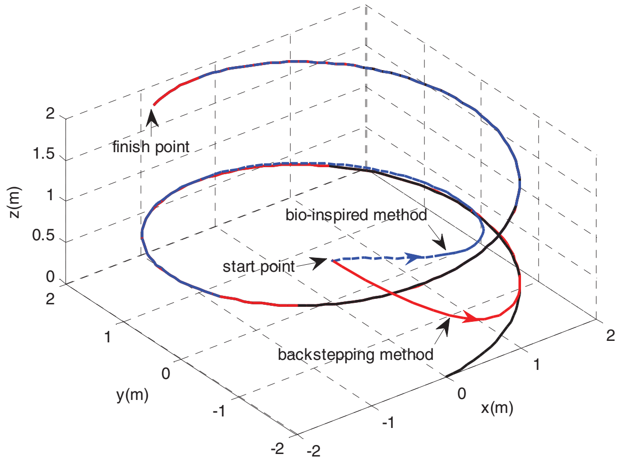


Fig. 6. Systems trajectories using bio-inspired model and the backstepping method

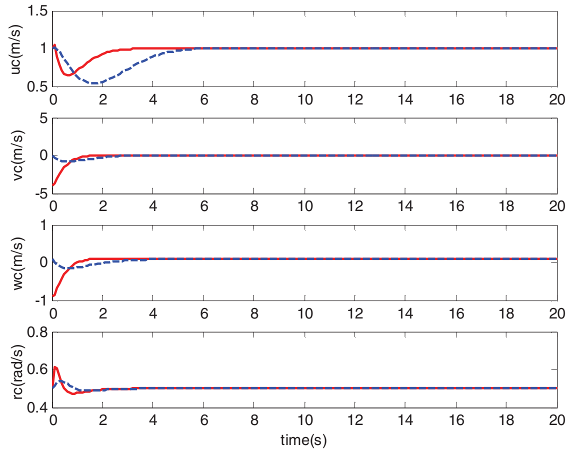


Fig. 7. Virtual velocity using bio-inspired model (dotted line) and the backstepping method (solid line)

In order to catch up the virtual velocities, the UUV accelerates both linear and angular velocities. So the UUV needs to generate quite large forces/moments to track the virtual velocity $q_c$ . It means that the required acceleration and forces/moments may exceed their control constraint values at the velocity jump points. In order to achieve such large speed, it can be seen that the maximum normalized thruster control variable of $\overline{u}_3$ is -5.6880 and the other maximum normalized thruster control variable are also greater than the limitation output ( $-1 \le \bar{\boldsymbol{u}} \le 1$ ) which is practically impossible. In addition, the control variable needs more time to adjust the overshoot process. By contrast, for the bio-inspired tracking control method, no any control value exceeds their maximum control values, and all control forces limited in the thruster control saturation point. This proves the efficiency of the proposed cascaded tracking method.

Table III. Maximum normalized thruster control variable

| Algorithm | $\bar{u}_1$ | $\bar{u}_2$ | $\bar{u}_3$ | $\bar{u}_4$ |
|---|---|---|---|---|
| Backstepping method | 2.4953 | -1.8335 | -5.6880 | -2.3782 |
| Bio-inspired method | 0.9055 | -0.6122 | -0.5290 | -0.4728 |

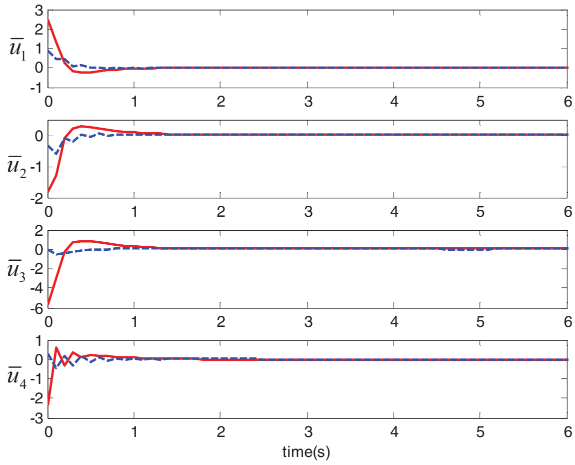


Fig. 8. Thruster forces using bio-inspired model (dotted line) and the backstepping method (solid line)

## VI. Conclusion

Background information about tracking control of unmanned underwater vehicles is firstly established in the paper. Then an integrated backstepping and sliding mode tracking control algorithm is proposed for three-dimensional tracking control problem. In the control system, there exist two closed loop systems: inner loop ensures the velocity tracking and the outer loop ensures the position and orientation tracking. In the traditional backstepping method, it always suffers from the sharp speed jump problem which will cause thruster saturation. Because of the smooth and bounded response properties, the proposed velocity controller uses the bio-inspired model to eliminate or inhibit the sharp speed jumps. From the simulation results, it is clearly to see bio-inspired method reduces the thrust output obviously without significant performance loss while the conventional backstepping method may cause thruster saturation.

In the dynamic controller, the sliding mode control method uses an adaptive term to compensate the nonlinear uncertainties part and the disturbance of the underwater vehicles dynamics. Simulation results showed that the designed controller performs well with stability and robustness. On the other hand, ocean currents should be added to the outside interference. This warrants further research.

## Acknowledgment

This project is supported by the National Natural Science Foundation of China (51075257) and the Creative Activity Plan for Science and Technology Commission of Shanghai (10550502700), the Project of Excellent Academic Leaders of Shanghai (11XD1402500) and the Yangtze River Delta Region scientific and technological project (10595812700).

## References

[1] J. Yuh, Design and Control of Autonomous Underwater Robots: A Survey, Autonomous Robots, vol. 8, no. 1, pp. 7-24, Jan. 2000.

[2] G. Antonelli, F. Caccavaleet, and S. Chiaverini, Adaptive tracking control of underwater vehicle-manipulator systems based on the virtual decomposition approach, *IEEE Trans. Robot. Autom.*, vol. 20, no. 3, pp. 594-602, Jun. 2004.

[3] D. A. Smallwood and L. L. Whitcomb, Model-based dynamic positioning of underwater robotic vehicles: Theory and experiment, *IEEE J. Ocean. Eng.*, vol. 29, no.1, pp. 169-186, Jan. 2004.

[4] V. Sankaranarayanan and A. D. Mahindrakar, Control of a Class of Underactuated Mechanical Systems Using Sliding Modes, *IEEE Trans. Robot*, vol. 25, no. 2, pp. 459-467, Apr. 2009.

[5] T. Nguyen, J. Leavitt, F. Jabbari, and J. E. Bobrow, Accurate sliding-mode control of pneumatic systems using low-cost solenoid valves, *IEEE/ASME Trans. Mechatronics*, vol. 12, no. 2, pp. 216-219, Apr. 2007.

[6] S. Soylu, B. J. Buckham, R. P. Podhorodeski, A chattering-free sliding-mode controller for underwater vehicles with fault-tolerant infinity-norm thrust allocation, *Ocean Engineering*, vol. 35, no. 16, pp. 1647-1659, Nov. 2008

[7] L. Lapierre and B. Jouvencel, Robust Nonlinear Path-Following Control of an AUV, *IEEE J Oceanic Eng*, vol. 33, no. 2, pp: 89-102, Apr. 2008.

[8] R. Fierro and F. L. Lewis, Control of a nonholonomic mobile robot: Backstepping kinematics into dynamics, *J. Robot. Syst.*, vol. 14, no. 3, pp. 149-163, 1997.

[9] K. Ishaque, S. S. Abdullah, and S. M. Ayob, A simplified approach to design fuzzy logic controller for an underwater vehicle, *Ocean Engineering*, vol. 38, pp. 271-284, Jan. 2011.

[10] G. Antonelli, S. Chiaverini, and G. Fusco, A fuzzy-logic-based approach for mobile robot path tracking, *IEEE Trans. Fuzzy Syst.,* vol. 15, no. 2, pp. 211-221, Apr. 2007.

[11] J. H. Li, P. M. Lee, and B. H. Jun, A neural network adaptive controller for autonomous diving control of an autonomous underwater vehicle, International Journal of Control Automation and Systems, vol. 2, no. 3, pp. 374-383, Sep. 2004.

[12] V. S. Kodogiannis, Neuro-control of unmanned underwater vehicles, *Int J of Systems Sc*, vol. 37, no. 3, pp. 149-162, Feb. 2006.

[13] C. Luo and S. X. Yang, A bio-inspired neural network for real-time concurrent map building and complete coverage robot navigation in unknown environment, *IEEE Trans. Neural Netw.,* vol. 19, no. 7, pp. 1279-1298, Jul. 2008.

[14] S. X. Yang and C. Luo, A neural network approach to complete coverage path planning, *IEEE Trans Syst Man Cybern B*, vol. 34, no. 1, pp. 718-725, Feb. 2004.

[15] T. I. Fossen, Guidance and Control of Ocean Vehicles, Wiley, New York, 1994.

[16] D. Steinke, Design and simulation of a Kalman filter for ROV navigation, Master Thesis, University of Victoria, 2003.

[17] S. Grossberg, Nonlinear neural networks: Principles, mechanisms, and architectures, *Neural Networks,* vol. 1, no. 1, pp. 17-61, 1988.

[18] A. L. Hodgkin and A. F. Huxley, A quantitative description of membrane current and its application to conduction and excitation in nerve, J. Physiol., vol. 117, no. 4, pp. 500-544, 1952.

[19] A.J. Sørensen and Y.N. Smogeli, Torque and power control of electrically driven marine propellers, Control Engineering Practice, vol. 17, no. 10, pp. Sep. 1053-1064, 2009.

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# Modelling of the pendulum-driven cart system with friction

Wenjuan Ren (MRes student), Sam Wane (PhD student), and Hongnian Yu
Faculty of Computing, Engineering and Technology, Staffordshire University, Stafford, UK
rwenjuan@126.com, {S.O.Wane, H.Yu}@staffs.ac.uk

**Abstract — This paper investigates several general friction models and applies different friction models to accurately simulate the propulsion of a capsule driven by an internally oscillating mass. The six published friction models are reviewed and are validated using the simulation studies. The simulation results are verified with a physical prototype in an experimentation test. The most accurate friction model can be chosen for a given application.**

***Keywords-Pendulum-driven cart, friction models, friction measurement experiment***

## I. INTRODUCTION

Friction is the resistance to motion that exists when a solid object is moved tangentially with respect to the surface of another that it touches, or when an attempt is made to produce such motion [1]. This is a common physical nonlinear phenomenon and exists in almost every kind of mechanical system and makes an effect on the system which cannot be ignored. Friction can lead to tracking errors, limit cycles and unsteady stick-slip movements [2]. If we do not minimize the friction force effectively, it may cause both economic losses and industrial losses. Friction is also a quite important factor to some processes, such as walking, starting and stopping a car, skating and so on.

Nowadays many focused researches are conducted on the 'Capsule Robot' for medical purposes, because the capsule robot can go into the human body to do work such as treatment, inspection, image acquisition and so on without hurting the patient [3], [4], [5]. A new tracking control issue via the pendulum-driven cart system was solved by the closed-loop control strategy [6]. The six-step optimization motion control approach was also improved by the pendulum-driven cart-pole system [7]. In [6] and [7], the velocity of the cart changed to negative value at some moment, so the displacement decreased. In the pendulum-driven cart system, there was no opposite force, the cart should not go backwards, and this paper is to do research on this problem. However, a simple friction model was used in [3]-[7]. The friction plays an active role in the capsule robot movement. This paper aims to investigate the realistic friction models which can be used in analysis and control of an active driving capsule robot.

The structure of the paper is as follows: in section II, the review of previous work on the pendulum-driven cart pole-system is presented. In section III, investigation on friction phenomena and friction models are shown. A simulation study is carried out in section IV along with the implementation of the different friction models mentioned above. The experiment on measurement of friction parameters is described and compared with the experimental results and the simulation results, where the most suitable friction model for this system is selected. In section V, the conclusions are presented.

## II. MODELLING OF THE PENDULUM-DRIVEN CART SYSTEM

Fig.1 shows the pendulum-driven cart system, the inverted pendulum is fixed on the cart. The cart has a relative smooth surface which makes it move horizontally on the ground. There is a torque motor mounted on the cart which generates the torque to swing the inverted pendulum. $M$ is the mass of the cart, m is the mass of the ball of the pendulum, $l$ is the length of the pendulum, $\mu$ is the friction coefficient between the ground and the cart, $\theta$ is the pendulum angle from the vertical direction, $x$ is the displacement of the cart, and $\tau$ is the torque from the torque motor.

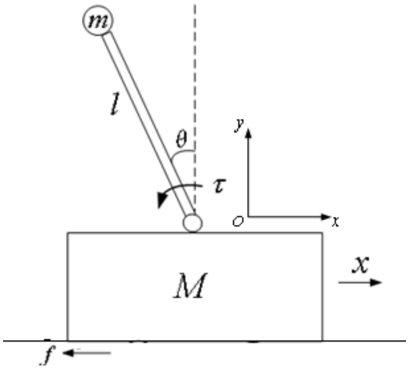


Figure 1. Pendulum-driven cart system

The Coulomb model of friction used in [6] is

$$F_f = -\mu N \operatorname{sgn}(\dot{x}),\ N = Mg + F_y \tag{1}$$

where $\mu$ is the friction coefficient between the cart and the surface, $N$ is the normal force, $F_y$ is the resultant force in the vertical direction.

The dynamic model of the system is described by

This work has been supported by the European Erasmus-Mundus Sustainable eTourism project 2010-2359, the EPSRC UK-Japan Network on Human Adaptive Mechatronics Project (EP/E025250/1) and EU Erasmus Mundus Project-ELINK (EM ECW-ref.149674-EM-1- 2008-1-UK-ERAMUNDUS).

$$(M+m)\ddot{x} - ml\ddot{\theta}(\mu \sin\theta \operatorname{sgn}\dot{x} + \cos\theta) + \\ ml\dot{\theta}^2(\sin\theta - \mu\cos\theta \operatorname{sgn}\dot{x}) + \mu(M+m)g \operatorname{sgn}\dot{x} = 0 \tag{2}$$

The improved friction model used in [7] is

$$F_f = -\mu N \operatorname{sgn}(\dot{x}),\ N = Mg + F_y \\ \mu = \begin{cases} sign(F_x)\mu_k & (\dot{x} = 0) \\ sign(\dot{x})\mu_k & (\dot{x} \neq 0 \end{cases} \tag{3}$$

where $\mu_k$ is the friction coefficient, $F_y$ is the resultant force in the vertical direction. The following dynamic model can be obtained

$$(M+m)\ddot{x} - ml\ddot{\theta}(\mu\sin\theta + \cos\theta) + ml\dot{\theta}^2 \\ (\sin\theta - \mu\cos\theta) + \mu(M+m)g\operatorname{sgn}\dot{x} = 0 \tag{4}$$

Based on the friction model (2) and the dynamic model (4), the simulation was carried out using Matlab/Simulink and the sample time is 0.01 second. When the inverted pendulum swings, the reaction force from the pendulum to the cart drives the cart to move in the horizontal direction, the displacement of the cart is shown in Fig. 2.

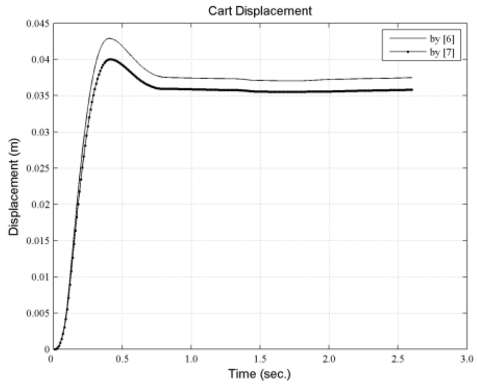


Figure 2. Cart displacements as found by [6]and [7]

From Fig. 2, there is decrease in the displacement curves, which means that the velocities of the cart change to negative values at the point that the displacement values decrease, which is not consistent with the reality situation, based on research of the reasons, improvement of the friction model can get the better result.

From the system configuration, the motor swings the pendulum to cause a lateral force on the cart to enable motion, which can be considered to be a spring. Assuming the cart to be a steel block on a smooth aluminium surface, the system can be simplified as shown in Fig. 5.

## III. FRICTION MODELS

Friction is the key factor to determine the progress of the pendulum-driven cart and the accuracy between the simulation and the physical system. In this paper, general friction phenomena and friction models have been investigated and the friction models have been applied to the dynamic model of the pendulum-driven cart system in Matlab\ Simulink [6], [7] and physical experiment has been carried out to prove the real friction model of the system via comparison of the simulation study and the experiment results.

Friction is the tangential reaction force between two surfaces in contact [8], is dependent on the physical characteristics of the contact surfaces, and displacement and the velocity of the contact body. There are many typical phenomena of friction, for instance, the Stribeck affect, hysteresis, spring-like characteristics for static friction, and varying break-away force [2]. Based on these friction phenomena, there are many friction models. The classical friction models are Coulomb friction model, Coulomb plus viscous friction model and Coulomb plus viscous plus static friction model [9].

Based on the experiment results, the Dahl friction model, LuGre friction model and so on [9], [10], [11] were proposed. In this paper, the Coulomb friction model, the Coulomb plus viscous friction model, the Coulomb plus viscous plus static friction model, the Stribeck friction model, the LuGre friction mode and the Dahl friction model are studied and implemented into the dynamic model of the pendulum-driven cart-pole system.

### *A. Modelling idea and evaluation criteria*

The conventional hypothesis to model the friction is that the direction of the friction force should be opposite to the direction of the relative velocity between the two contact surfaces. In [6], [7], the velocity of the cart changes to negative and the cart retreats at the end of the motion. The reason for this is that at the moment when the velocity of the cart changes to zero, friction changes its direction to the opposite at once. However in the simulation study, it cannot be achieved exactly even if the sample time is very small, to save the computing time and get the more exact result, modelling the friction force under the new hypothesis that the friction tries to stop the mass block [12]. This paper switches the dynamic and static friction model when the relative velocity between the steel mass block and the aluminium surface changes to zero. The reason for switching is that when a body is moving, the friction involved is the dynamic friction, when the velocity of the body changes to zero, the body stops, at this time, the friction involved is the static friction. Consequently, the friction model in this paper is dynamic model when the velocity of the block is not equal to zero, and uses the static model when the velocity of the block changes to zero.

To select the most accurate friction model for the mass block system, evaluation criteria used in this paper are that the velocity will not change to negative and the displacement of the mass block will not decrease, that is, the mass block will stop with the effect of friction and there is no reverse motion in the system, which is consistent with the physical reality.

### *B. Friction phenomena*

#### *1) Stick-slip motion*

Stick-slip motion is one of the typical friction behaviours in system. The reason for stick-slip motion is the friction is greater when the body keeps still than in motion, experiment [2] shows that the mass is still at first, and when there is a linearly increasing force generated by the spring, the mass moves to a very small displacement firstly and then starts to slide when the applied force reaches the break-away force, decrease the spring force and the mass slows down and stops, the friction force increases because the static friction is greater than the dynamic friction. In [18], the model for stick-slip friction is always the function of velocity. Outside the small region around velocity is zero, the stick-slip friction is an arbitrary function of velocity, and inside the region, friction is determined by other forces in the system.

*2) Stribeck effect*

The Stribeck effect is to describe the friction behaviour when the relative velocity between the two contact surfaces is very small. The Stribeck friction force is function of steady-state velocity, when the relative velocity is within very low range, the friction force will decrease as the relative velocity increases. The relationship of the Stribeck effect and the relative velocity is shown in Fig. 3.

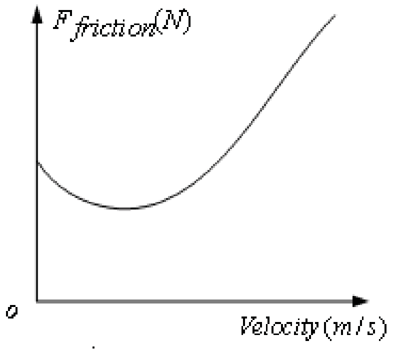


Figure 3. Stribeck effect

*3) Pre-sliding displacement*

When the applied force is smaller than the maximum static friction, the two contact bodies keep still, but from a micro point of view, there is a very tiny displacement on the asperities on the contact surfaces, this is called pre-sliding displacement, which is also known as the Dahl effect [9], [10]. During the pre-sliding stage of the contact bodies, the deflection of the asperities of the surfaces has the spring-like characteristic, and the friction force is the function of the displacement and is independent with the relative velocity between the contact surfaces. The graph of the Dahl effect and the displacement is shown as Fig. 4.

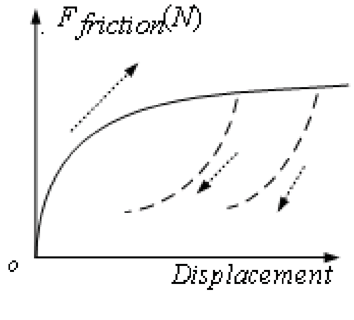


Figure 4. Pre-sliding displacement behaviour as found by [13]

*C. Friction models*

*1) Stribeck model*

The Stribeck model describes the friction phenomenon that the friction force does not decrease discontinuously when the contact surfaces starts to slip. The friction force is dependent on the velocity, when the relative velocity between the contact surfaces is very low, the friction force has a negative slope.

$$F = \begin{cases} F(v) & (\dot{x} \neq 0) \\ F_e = F_x & (\dot{x} = 0 \, \& \, |F_e| \prec |F_s|) \\ F_s \operatorname{sgn}(F_e) & (otherwise) \end{cases} \tag{5}$$

$$F_v = F_c + (F_s - F_c)e^{-\frac{\dot{x}}{\dot{x}_s}} + \mu\dot{x}$$

where $F_y$ is the Stribeck friction force, $F_s$ is the static friction force, $\dot{x}_s$ is called the Stribeck velocity, which means the low velocity as shown in Fig. 3[8], $F_c$ is the Coulomb friction force, $F_x$ is the resultant force in the horizontal direction.

*2) Dahl friction model*

The applied force is performed by the spring-like characteristics. There is a kind of viscous between the contact surfaces, if the applied force is larger than the defined range, and the viscous effect will be damaged.

$$\frac{dz}{dt} = \dot{x} \text{-} \sigma_0 \frac{|\dot{x}|}{F_c} Z$$

$$F_f = \begin{cases} F_d = \sigma_0 Z \operatorname{sgn}(\dot{x}) & (\dot{x} \neq 0) \\ F_x & (\dot{x} = 0) \end{cases} \tag{6}$$

where $F_d$ is the Dahl friction force, $F_c$ is the Coulomb friction force, $\sigma_0$ is the stiffness coefficient, and $Z$ is the bristle deflection. When the velocity changes to zero, the friction form changes to the resultant force of the horizontal direction, that is $F_x$.

*3) Coulomb friction model*

The Coulomb friction model was discovered by Leonardo da Vinci (1452-1519) and is commonly used in engineering [14]. The Coulomb friction is opposite to the motion direction, proportion to the normal pressure, and independent with displacement, but dependent on the direction of velocity.

$$F_f = \begin{cases} F_c = \mu N \operatorname{sgn}(\dot{x}) & (\dot{x} \neq 0) \\ F_x & (\dot{x} = 0) \end{cases} \tag{7}$$

where $\mu$ is the friction coefficient, $N$ is the normal force, $\dot{x}$ is the velocity and $F_x$ is the horizontal resultant force.

*4) Viscous friction model*

With the development of fluid dynamic technology, affect of viscosity lubrication and viscosity on force has been studied, and velocity has been considered as an important factor of friction, the viscous friction model has been proposed as *Fv* in (8) [15], where $\dot{x}$ is the sliding speed, and $\mu_v$ is the viscous coefficient. When the velocity changes to zero, the friction model switches to $F_x$ as shown in (8).

$$F_f = \begin{cases} F_V = \mu_V \dot{x} \operatorname{sgn}(\dot{x}) & (\dot{x} \neq 0) \\ F_x & (\dot{x} = 0) \end{cases} \tag{8}$$

where *Fv* is the viscous friction, is the viscous friction coefficient, $\mu_v$ is the resultant force in the horizontal direction, $\dot{x}$ is the velocity.

*5) LuGre friction model*

The LuGre friction models the average deflection force of elastic springs [16]. If the applied tangential force is large enough to damage the spring-like characteristic, the bristle starts to slip. The LuGre friction models the Stribeck effect and supposes that the contact surfaces are pushed apart by the lubricant.

$$\begin{gathered} \frac{dz}{dt} = \dot{x} - \frac{\sigma_1 |\dot{x}|}{F_V} Z \\ F_V = F_C + (F_S - F_C) e^{-(\frac{\dot{x}}{v_S})\alpha} \\ F_f = \begin{cases} F_V = (\sigma_1 Z + \sigma_2 \frac{dz}{dt} + \sigma_3 \dot{x}) \operatorname{sgn}(\dot{x}) & (\dot{x} \neq 0) \\ F_x & (\dot{x} = 0) \end{cases} \end{gathered} \tag{9}$$

where the state *Z* represents the average bristle deflection, which means the difference between the relative position of the bristle and the position where the bond was formed [8], $\sigma_1$ is the stiffness, $\sigma_2$ is the micro damping and $\sigma_3$ is the viscous friction coefficient, $F_v$ represents the Stribeck effect, $F_c$ is the Coulomb friction force and $F_s$ is the static friction force.

*6) Coulomb+viscous friction model*

Based on the development of fluid mechanics, it is found that there is viscosity between the contact surfaces, the viscous friction force is proportional to the velocity.

$$F_f = \begin{cases} F_c \operatorname{sgn}(\dot{x}) + \mu \dot{x} \operatorname{sgn}(\dot{x}) & (\dot{x} \neq 0) \\ F_x & (\dot{x} = 0) \end{cases} \tag{10}$$

where $F_c$ is the Coulomb friction, $\dot{x}$ is the velocity, $\mu$ is the viscous friction coefficient and $F_x$ is the resultant force in the horizontal direction.

## IV. SIMULATION STUDY OF THE SPRING-BLOCK SYSTEM

To get the real values of the parameters, for instance, the dynamic friction coefficient, the static friction coefficient, the maximum static friction force and so on, in this paper, a physical experiment was carried out to complete this work.

For the system as shown in Fig. 5, the steel block stands on the aluminium flat, compasses the spring one end of which is fixed to the left wall. When the spring is released, the steel block will move forward and will stop at one point because of the friction of the contact surfaces. Applying different friction models in simulation, different values of the mass block displacements can be obtained. The model under which the displacement value is closest to the experiment result can be chosen as the most accurate model.

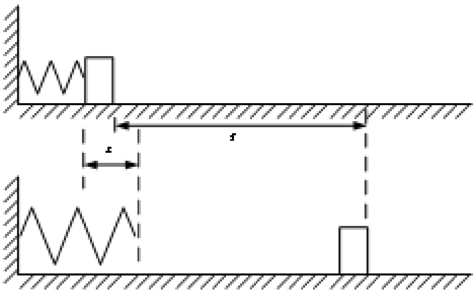


Figure 5. Spring-steel block system for experiment and simulation study

From the energy equation (11), the displacement of the system is unique for the determinant parameters. As a result, the value of displacement can be chosen as the key factor to select the friction model.

$$\frac{1}{2} kx^2 = \mu mgs \tag{11}$$

Both in the simulation study and in the experiment, the applied force is *F*=10N, the applied time is 0.1s.

TABLE I. STEEL BLOCK PARAMETERS

| Parameter | Value |
|---|---|
| *M* | 1kg |
| *K* | 100N/m |
| $\mu_d$ | 0.25[17] |
| $\mu_s$ | 0.35[17] |

Parameters used simulation study, the same with the real parameters used in experiment

### *A. Physical experiment to find the most accurate friction model*

As shown in Fig. 5, the spring is attached to the left wall with a compression length *x*. There are also a steel mass block and an aluminium platform. When the spring is released, the steel mass block will go to the right direction and will stop at a distance *s*. Values of parameters applied in the experiment and the simulation study are the same. In Fig. 5, *m* is the mass of the steel block, *k* the spring constant, and *F=Kx* the external applied force to the system. The task of the experiment is to measure the displacement of the block when the spring releases. Parameter values used in the experiment are as shown in Table I. Through the experiment, the displacement of the mass block is 0.20*m*.

### *B. Apply friction models to the spring-mass block system in simulink*

*1) Coulomb friction model result*

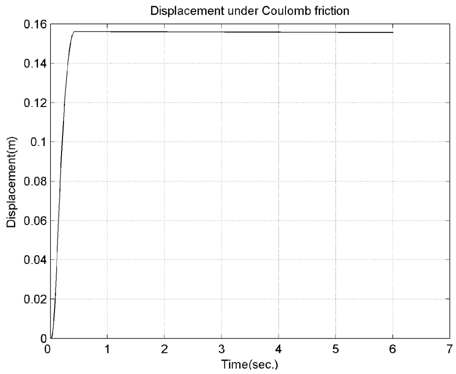


Figure 6. Displacement under Coulomb friction model

Figure 6. shows that the velocity of the mass block under the Coulomb friction model is always positive and the displacement keeps at 0.16$m$ when the mass block goes to still, without reverse motion.

*2) Dahl friction model result*

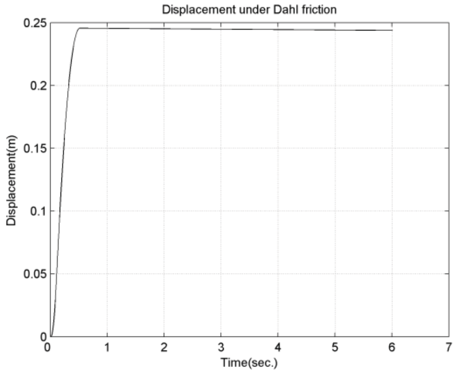


Figure 7. Displacement under Dahl friction model

Figure 7. illustrates that the relative velocity between the steel mass block and aluminium flat does not turn to negative, and the displacement is 0.25$m$ when the steel block stops without the inverse motion.

*3) Stribeck friction model result*

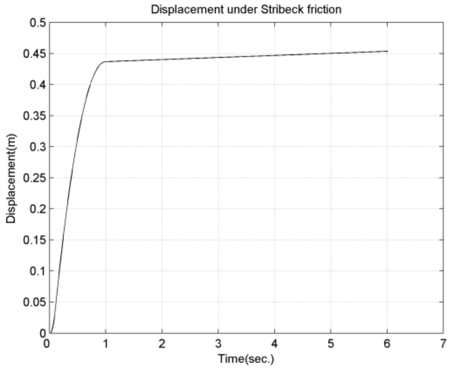


Figure 8. Displacement under Stribeck friction model

From Fig. 8, the velocity of the steel block under the Stribeck friction model does not change to zero exactly and the displacement is about 0.45$m$ when the simulation time reaches about 10$s$, the final result is not consistent with the physical reality.

*4) LuGre friction model result*

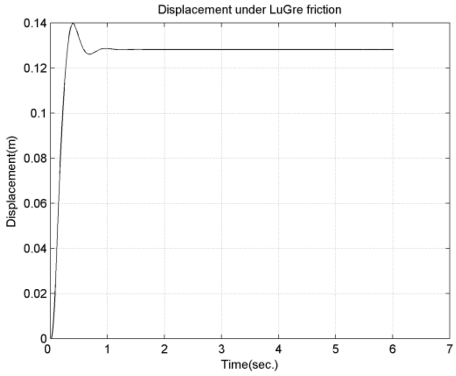


Figure 9. Displacement under LuGre friction model

Figure 9. demonstrates that the velocity of the steel block under the LuGre friction model firstly changes to negative and then returns to zero. And the displacement decreases before it keeps still. The final displacement is 0.13$m$.

*5) Coulomb plus viscous friction model result*

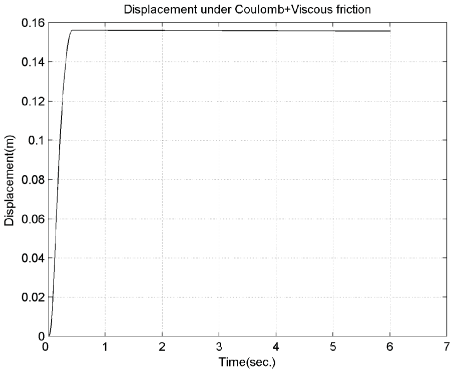


Figure 10. Displacement under Coulomb+Viscous friction model

From Fig. 10, the velocity does not change to negative and the displacement of the steel mass block stops at 0.16$m$ without inverse motion.

The research results from the simulation study are summarized in table II.

TABLE II. RESEARCH RESULTS

| **Friction model** | **Velocity*** | **Displacement** | **Relative error**** |
|---|---|---|---|
| Coulomb | Yes | 0.16$m$ | 20% |
| Dahl | Yes | 0.25$m$ | 25% |
| LuGre | No | 0.13$m$ | 35% |
| Stribeck | Yes | 0.45$m$ | 125% |
| Coulomb+Viscous | Yes | 0.16$m$ | 20% |

*velocity is always positive or not; **relative error is between the simulation result and the experiment result.

From Table II, the LuGre friction model falls into disuse firstly because of negative velocity. Displacement of mass block reaches different values under different friction models. The minimal relative error is 20% under the improved Coulomb friction model as shown in (7) and improved Coulomb plus viscous friction model as shown in (10). According to the selection criterion proposed previous, the most accurate friction model should be the improved Coulomb friction model or the Coulomb plus viscous friction model.

Friction property depends mostly on the contact surfaces. Because the spring-steel block and the pendulum-driven cart share the same physical condition, the selected friction model can be applied to the pendulum-driven cart system in the future research.

## V. CONCLUSIONS

Applying the selected friction model (take Coulomb plus viscous friction model for instance) to the pendulum-driven cart-pole system as shown in Fig. 1, the displacement and velocity of the cart under the Coulomb plus Viscous friction model are shown in Fig. 11.

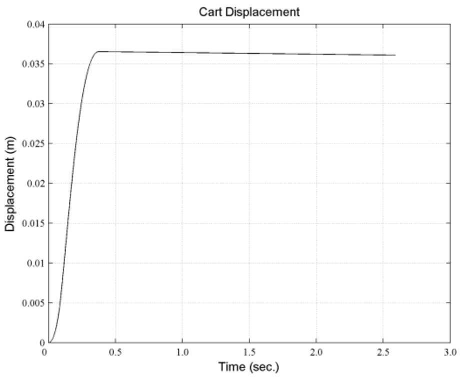


Figure 11. Cart displacement under Coulomb+viscous friction model

From Fig. 11, the cart velocity under the Coulomb plus viscous friction model does not change to negative. The cart displacement under the Coulomb plus viscous friction model does not decrease any more. At the same time, the improved friction model can also makes the cart drives to the same displacement with the previous friction models, without any inverse motion, which is consistent with the physical phenomena.

This paper has completed the following tasks

- Studied the friction phenomena and investigated switching between static and dynamic friction models.
- Proposed two selection criterions of the friction models.
- Simulated the spring-steel block system.
- Did the physical experiment.
- Got the most suitable friction model for the spring-steel mass block system as shown in Fig. 5 and the pendulum-driven cart system as shown in Fig. 1.

In conclusion, the selected friction model considers the switching between dynamic and static friction, and makes the pendulum-driven cart system move in the positive direction without any inverse motion.

## REFERENCES

[1] Ernest Rabinowicz, "Friction and wear of materials," 2nd ed. Wiley-Interscience, 1995, pp. 65–79.

[2] C. Canudas de Wit, H. Olsson, K. J. Astrom and P. Lischinsky, "A new model for control of systems with friction, " IEEE transactions on automatic control, vol. 40, No.3, March 1995, pp. 419-425.

[3] Hongyi Li, Katsuhisa Furuta, Chernousko, F.L." Motion generation of the capsubot using internal force and static friction," presented at the 2006 IEEE Conference, Decision and Control, San Diego, CA, Dec. 13-15, 2006, pp. 6575-6580.

[4] M. Nazmul Huda, Hong-Nian Yu and Samuel Olive Wane,"Self-contained capsubot propulsion mechanism," International Journal of Automation and Computing, Vol. 8, No. 3, August 2011, pp. 348-356.

[5] Y. Liu, H. Yu, and T. C. Yang, "Analysis and control of a capsubot", proceedings of the 17th world congress, the international federation of Automatic Control, Seoul, Korea, July 6-11, 2008.

[6] H Yu, Y Liu, and T Yang. "Closed –loop tracking control of a pendulum-driven cart-pole underactuated system," Journal of System and Control Engineering, vol. 222, no.2. pp. 109-125, 2008.

[7] Y Liu, H Yu, and B Burrows. "Optimization and control of a pendulum-driven cart-pole system," In proceedings of the IEEE International Conference on Networking, sensing and control, London. April 2007, pp. 151-156.

[8] H. Olsson, K.J. Astrom, C. C. de Wit, M. Gafvert, P. Lischinsky, Friction models and friction compensation. Eur. J. Control, Vol. 4, No. 3. 1998, pp. 176-195.

[9] Brain Armstrong-Helouvry, Pierre Dupont and Carlos Canudas de Wit, A survey of models, analysis tools and compensation methods for the control of machines with friction. Automatica, Vol. 30, No. 7,1994, pp. 1083-1138.

[10] P.R. Dahl, "A solid friction model," Belvoir Defense Technical Information Centre, May 1968.

[11] D. A. Haessig, Jr and B. Friedland, "On the modelling and simulation of friction," J. Dyn, Sys., Meas, Control, Vol. 113, Issue. 3, September 1991, pp. 354-362.

[12] Paul Breedveld, Annemarie Y. Diepenbroek, Ton van Lunteren. "Real-time simulation of friction in a flexible space manipulator," ICAR 1997, Monterey, CA, July 7-9, 1997.

[13] J. Courtney-Pratt and E. Eisner. "The effect of a tangential force on the contact of metallic bodies," in proceeding of the Royal Society, Vol. A238, 1957, pp. 529-550.

[14] Armstrong-Helouvry, Brian, "Control of machines with firction," The Springer International Series in Engineering and Computer Science, 1 Edition, 1991, pp. 10.

[15] O. Reynolds, "On the theory of lubrication and its application to Mr. Beauchamp Tower's experiments, including an experimental determination of the viscosity of olive oil,"transactions of the Royal Society of London Proceedings Series 1, 1886, pp. 157-234.

[16] K.J. Astrom and C. Canudas-de-Wit, "Revisiting the LuGre model," IEEE Control Systems Magazine 28, June 2006, pp. 101-114.

[17] American Society for Metals, ASM Handbook, Vol. 18, Lubrication, and Wear Technology.

[18] Karnopp, D. 1985, Computer simulation of stick-slip friction in mechanical dynamical systems. Journal of Dynamic Systems, Measurement, and Control, 107, 100–103.

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# A theoretical approach to the passive control of spiral vortex breakdown

Ubaid A. Qadri and Matthew P. Juniper
Department of Engineering
University of Cambridge
Cambridge, UK CB2 1PZ
Email:uaq20@cam.ac.uk

***Abstract*—Previous numerical simulations have shown that vortex breakdown starts with the formation of a steady axisymmetric bubble and that an unsteady spiralling mode then develops on top of this. We study how this spiral mode of vortex breakdown might be suppressed or promoted. We use a Lagrangian approach to identify regions of the flow which are sensitive to small open-loop steady and unsteady (harmonic) forces. We find these regions to be upstream of the vortex breakdown bubble. We investigate passive control using a small axisymmetric control ring. In this case, the steady and unsteady control forces are caused by the drag force on the control ring. We find a narrow region upstream of the bubble where the control ring will stabilise the flow and we verify this using numerical simulations.**

***Index Terms*—flow control, vortex breakdown, passive control, adjoint, sensitivity analysis**

## I. Introduction

Vortex breakdown has been observed in many practical flows, such as the flow over the leading edge of delta wings at high angles of attack, the injection of fuel and air into combustion chambers, and the intense rotating flow found in a tornado. In all these cases, when the fluid rotates with sufficient azimuthal velocity (swirl), a stagnation point and a recirculation bubble form within it. The transition from the flow without a breakdown bubble to the flow with a breakdown bubble is labelled *axisymmetric* vortex breakdown. In many cases, a spiral structure is seen to emanate and grow downstream of the breakdown bubble. This is labelled *spiral* vortex breakdown.

Vortex breakdown was first observed in the flow over gothic and delta wings at high angles of attack in 1957 [1]. Since then, several different forms of vortex breakdown have been observed in a variety of experimental settings such as tubes, nozzles, and combustion chambers. Investigators often observed the axisymmetric and spiral modes of breakdown to occur almost simultaneously. This led to disagreements over the nature of vortex breakdown. Recent numerical studies of vortex breakdown in an unconfined domain [2], [3], however, have confirmed that the basic form of vortex breakdown is axisymmetric and that the spiral mode is caused by the self-sustained growth of helical perturbations on top of the breakdown bubble. This is a global instability.

The importance of the vortex breakdown phenomenon means that there is a need to understand and control it. In the past, various open-loop control strategies have been attempted [4]. These include active flow control using blowing and suction, and passive flow control using mechanical devices in the flow. However, their success has always been limited due to insufficient knowledge of the physical mechanisms that are at work. To this end, numerical sensitivity analyses have been successful in predicting how one might control vortex shedding off cylinders and other blunt bodies at moderate Reynolds numbers [5]–[8]. These sensitivity analyses use adjoints to calculate the receptivity of the flow to external forcing and the sensitivity of the flow to internal feedback. They can provide information about the effect of steady and harmonic forces on the unstable mode. Hence they have been used to predict where a control device should be placed to either suppress or promote vortex shedding.

In this paper, we carry out a similar analysis around the axisymmetric vortex breakdown state and predict how the spiral mode of vortex breakdown might be suppressed or promoted. In Section II, we consider the stability of the axisymmetric breakdown state and show that spiral vortex breakdown is caused by a linear global instability. In Section III, we use adjoints to evaluate the effect of a small control force on the growth rate of the unstable mode and identify the regions of the flow that are most sensitive to a control force. In Section IV, we apply these results to the simple case of passive control using a small axisymmetric control ring. Finally, in Section V, we discuss how these techniques can be applied in practice.

This study is at $Re = 200$ and the primary motivation is scientific. There are important industrial motivations, however. Vortex breakdown occurs in wingtip vortices behind aircraft, in vacuum cleaners, and in gas turbine combustion chambers. In the case of combustion chambers, hydrodynamic instabilities in the flow can lock into acoustic resonances within the combustion chamber, causing high amplitude thermoacoustic instabilities, which can be catastropic. This fundamental study of spiral vortex breakdown will reveal the regions of the flow where control would be most effective. This could help designers devise effective control strategies.

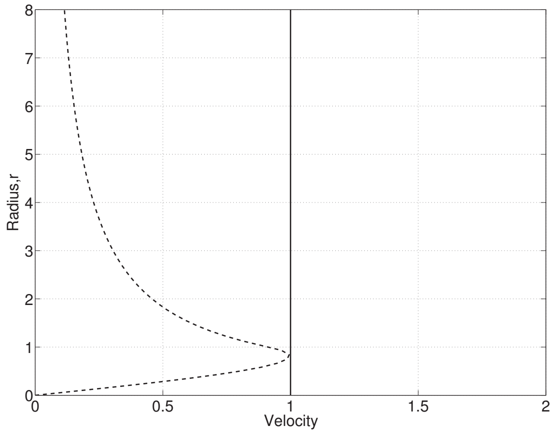


Fig. 1. Non-dimensional inlet velocity distributions for the Grabowski profile: the solid line represents the axial velocity, the dashed line represents the azimuthal velocity for a swirl value of $Sw = 0.915$.

## II. The flow configuration and the global stability analysis

We study the motion of a viscous fluid in a cylindrical domain with length $X_{max}$ and radius $R_{max}$, using cylindrical coordinates $(x, r, \theta)$. The flow has density $\rho$, pressure $p$, temperature $T$, and velocity $\mathbf{u} = (u_x, u_r, u_\theta)^T$. We describe the motion of the flow using the Navier–Stokes equations in the low Mach number limit. This allows for density variations in the flow but excludes acoustic waves. These equations can be expressed in terms of the momentum $\mathbf{m} = \rho\mathbf{u}$, temperature and pressure as

$$\frac{\partial \mathbf{q}}{\partial t} = \mathcal{N}\mathbf{q}, \tag{1}$$

where $\mathbf{q} \equiv (m_x,\, m_r,\, m_\theta,\, T, p)^T$ is the state vector and $\mathcal{N}$ is a nonlinear differential operator representing the action of the equations on the state vector. The density, $\rho$, is not included in the state vector because it can be derived from the temperature, $T$.

Along $x = X_{max}$ and $r = R_{max}$, we choose boundary conditions so that we model flow into a semi-infinite domain in the downstream and radial directions. At the inlet to the domain, we impose velocity profiles that have been used to study vortex breakdown numerically in the past [9]. This Grabowski profile, shown in Figure 1, has uniform density and temperature. The ratio of the azimuthal to axial velocities at $r = 1$ defines the swirl parameter, $Sw$. The Reynolds number is defined in terms of the nominal vortex core radius and uniform axial velocity. In this study, we keep $Re = 200$.

We obtain a steady axisymmetric laminar baseflow, $\bar{\mathbf{q}}$, or equilibrium point of the equations (1) such that

$$\mathcal{N}\bar{\mathbf{q}} = 0, \tag{2}$$

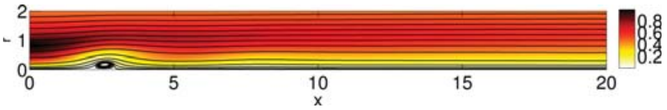


Fig. 2. Steady baseflow at $Sw = 0.915$ and $Re = 200$. The domain extends radially from $-8 \leq r \leq 8$ but only a portion is shown here. There is a small axisymmetric breakdown bubble around $x = 2.5$.

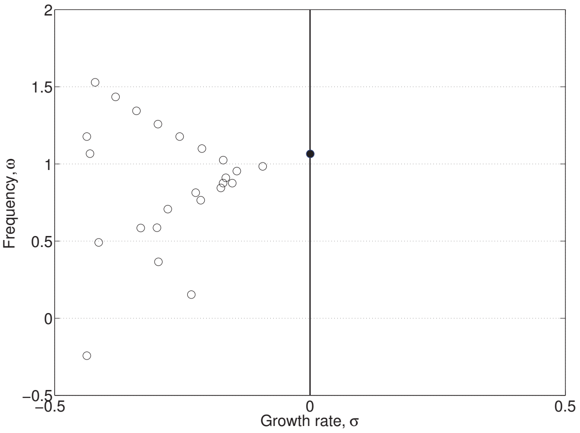


Fig. 3. Spectrum of the linear operator $\mathbf{L}$ for $m = -1$ for the baseflow in Figure 2, showing the 25 least stable eigenvalues. One mode, coloured black, is just unstable.

Figure 2 shows the steady baseflow at $Sw = 0.915$. There is a breakdown bubble around $x = 2.5$. The evolution of small perturbations $\mathbf{q}'$ around this field is governed by

$$\frac{\partial \mathbf{q}'}{\partial t} = \mathbf{L}\mathbf{q}', \tag{3}$$

where $\mathbf{L}$ represents the Navier–Stokes equations linearized about the base-flow $\bar{\mathbf{q}}$. We decompose the perturbations into Fourier modes in time and the azimuthal direction

$$\mathbf{q}'(x, r, \theta, t) = \hat{\mathbf{q}}(x, r)e^{\mathrm{i}m\theta + \lambda t}, \tag{4}$$

where $m$ (without a subscript) is the azimuthal wavenumber, and $\lambda \equiv \sigma + \mathrm{i}\omega$ contains the growth rate, $\sigma$, and frequency, $\omega$. We study the linear dynamics of the flow by analyzing the eigenvalues of $\mathbf{L}$. These are given by solving the matrix eigenvalue problem

$$\lambda\hat{\mathbf{q}} = \mathbf{L}_m\hat{\mathbf{q}}, \tag{5}$$

where $\mathbf{L}_m$ is the linear operator for the azimuthal wavenumber $m$. Each of these eigenvalues has a corresponding two-dimensional eigenfunction, $\hat{\mathbf{q}}(x, r)$. We label each eigenvalue/eigenfunction pair a *direct global mode*. If $\sigma > 0$, the mode is linearly globally unstable. In this linear analysis, the flow tends to the form of the global mode with highest $\sigma$ in the long-time limit and therefore this mode determines the system's overall stability.

Figure 3 shows the eigenvalue spectrum for $m = -1$ at $Sw = 0.915$, for which there is one unstable global mode. All

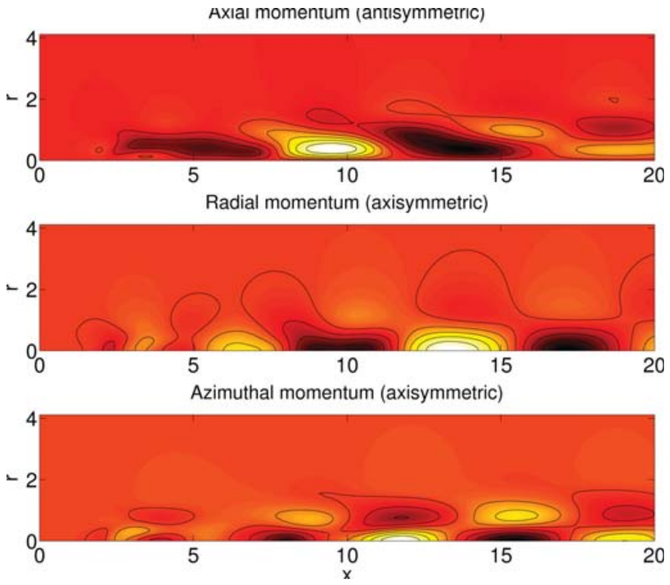


Fig. 4. Spatial structure of the most unstable eigenmode for $m = -1$, showing the real part of the axial, radial and azimuthal momentum in the top half of the domain.

other azimuthal wavenumbers are stable. The spatial structure of the unstable global mode is shown in Figure 4 using the real part of the axial, radial and azimuthal momentum. The imaginary part of the global mode is $1/4$ wavelength out of phase because this mode grows and advects downstream.

## III. Sensitivity to a control force

We now consider open-loop control of the unstable eigenvalue. Depending on the application, engineers might want either to suppress the unstable mode (to delay transition, for instance) or to promote it (to increase mixing, for instance).

To begin with, we evaluate the effect of a small control force on the unstable eigenvalue. We model the control force by adding mass, momentum and energy source terms to the right-hand side of equation (1):

$$\frac{\partial \mathbf{q}}{\partial t} = \mathcal{N}\mathbf{q} + \mathbf{F}. \qquad (6)$$

Here, the forcing terms have been grouped together as $\mathbf{F}$. The control force has a steady component ($\bar{\mathbf{F}}$) that acts on the base flow ($\bar{\mathbf{q}}$) and a linearized perturbation ($\mathbf{f}'$) that acts on the linear operator ($\mathbf{L}$) . We model the effects of these two components separately [10].

### *A. The sensitivity to steady forcing*

The eigenvalue of the global mode, $\lambda = \sigma + \mathrm{i}\omega$, is a function of the base flow fields ($\bar{\mathbf{q}}$) and these are, in turn, functions of the steady components of the forcing terms ($\bar{\mathbf{F}}$). The eigenvalue can, thus, be considered to be a function of the steady component of the forcing terms, $\lambda = f(\bar{\mathbf{F}})$. We wish to find the gradient of the functional $\lambda(\bar{\mathbf{F}})$ [8, Fig.9] for the unstable flow in §II. We investigate the variation of the eigenvalue, $\delta\lambda_{\bar{\mathbf{F}}}$, with respect to small variations of the steady forces, $\delta\bar{\mathbf{F}}$. The change in the eigenvalue is given by

$$\delta\lambda_{\bar{\mathbf{F}}} = \langle \nabla_{\bar{\mathbf{F}}}\lambda, \delta\bar{\mathbf{F}} \rangle, \qquad (7)$$

where $\nabla_{\bar{\mathbf{F}}}\lambda$ is a complex function that we call the *sensitivity of the eigenvalue to steady forcing*. The notation $\langle \mathbf{a}, \mathbf{b} \rangle$ denotes an inner product over a volume $V$,

$$\langle \mathbf{a}, \mathbf{b} \rangle = \frac{1}{V}\int_V \mathbf{a}^H \mathbf{b}\, dV, \qquad (8)$$

where $\mathbf{a}^H$ denotes the Hermitian (*i.e.* complex conjugate transpose) of $\mathbf{a}$.

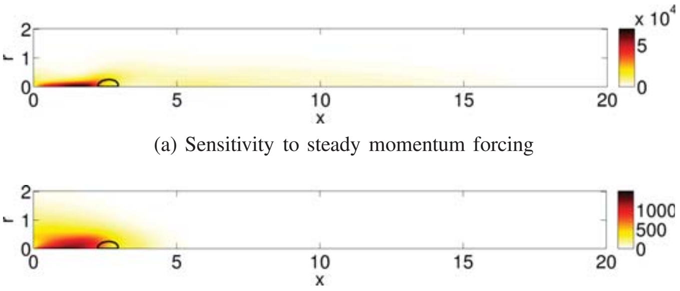
We calculate the sensitivity function by formulating a Lagrangian problem for $\lambda$. The nonlinear and linearised Navier–Stokes equations act as constraints in this problem,

$$\mathcal{L} = \lambda - \langle \bar{\mathbf{q}}^+, \mathcal{N}\bar{\mathbf{q}} - \bar{\mathbf{F}} \rangle - \langle \hat{\mathbf{q}}^+, \lambda\hat{\mathbf{q}} - \mathbf{L}_m\hat{\mathbf{q}} \rangle \qquad (9)$$

The Lagrange multipliers, $\bar{\mathbf{q}}^+$ and $\hat{\mathbf{q}}^+$, are the adjoint base flow and adjoint global mode fields respectively. We are interested in the functional derivative of $\mathcal{L}$ with respect to $\bar{\mathbf{F}}$. To find this, we first set the functional derivatives of $\mathcal{L}$ with respect to all other variables to zero. This leads to a set of equations that defines an eigenvalue problem for the adjoint global mode, a set of equations for the adjoint base flow fields and the normalization condition $\langle \hat{\mathbf{m}}^+, \hat{\mathbf{m}} \rangle + \langle \hat{T}^+, \hat{T} \rangle = 1$. The analysis then shows that the sensitivity of the eigenvalue to the steady forcing terms is given by the relevant adjoint baseflow field. For example, the sensitivity to momentum forcing is given by the adjoint baseflow momentum, $\bar{\mathbf{m}}^+$.

The adjoint base flow fields are complex valued. The real part represents the sensitivity of the growth rate, $\sigma$ while the imaginary part represents the sensitivity of the frequency $\omega$ to small changes in the steady forcing. They provide information about the most sensitive regions for control based on the physical mechanisms that cause the instability.

The adjoint base flow equations contain terms from the steady base flow, the direct global mode and the adjoint global mode. These need to be calculated before the adjoint base flow equations can be solved. The procedure for obtaining the sensitivity to steady forcing involves the following steps:
1. Obtain a steady base flow (Equation 2 and Figure 2).
2. Obtain the direct global mode by solving the direct eigenvalue problem (Equation 5 and Figure 4).
3. Obtain the adjoint global mode by solving the adjoint eigenvalue problem.
4. Normalize the adjoint global mode.
5. Solve the adjoint base flow equations to obtain the sensitivity of that global mode to steady forcing.



(a) Sensitivity to steady momentum forcing

(b) Sensitivity to unsteady momentum forcing

Fig. 5. Sensitivity to steady and unsteady forcing for the marginally unstable mode. Darker regions are more sensitive. The thick black line indicates the vortex breakdown bubble. The quantities plotted are (a) The adjoint baseflow momentum $||\bar{\mathbf{m}}^{+}||$ and (b) the adjoint global mode momentum $||\hat{\mathbf{m}}^{+}||$

### B. *Sensitivity to unsteady forcing*

The change in the eigenvalue due to a small linearized force ($\delta\mathbf{f}'$) that acts on the linear operator $\mathbf{L}$ is given by

$$\delta\lambda_{\mathbf{f}'} = \langle \nabla_{\mathbf{f}'}\lambda, \delta\mathbf{f}' \rangle, \tag{10}$$

where $\nabla_{\mathbf{f}'}\lambda$ is labelled the *sensitivity of the eigenvalue to harmonic forcing*.

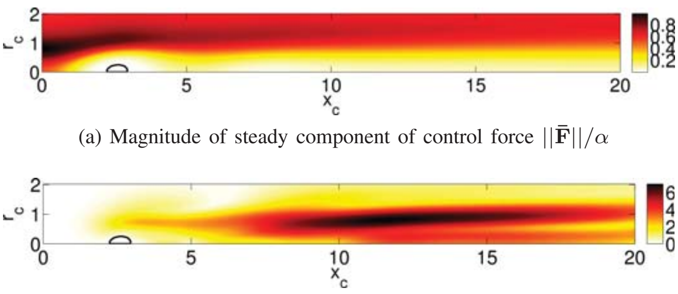
As for the steady forcing terms, the sensitivity of the eigenvalue to harmonic forcing is given by the relevant adjoint global mode fields. The change in the eigenvalue is, thus, simply obtained by projecting the linearized force onto the adjoint global mode fields. The largest change is obtained when the forcing frequency is equal to the frequency of the linear global mode [8, §4.1].

### C. *Results of sensitivity analysis*

Figures 5a and 5b show the sensitivity of the marginally unstable eigenvalue to steady and harmonic forcing respectively. We notice that, for both types of forcing, the flow is most sensitive just upstream of the breakdown bubble. This shows where a control force will have the greatest effect. The scales in the plot indicate that the sensitivity to steady forcing is almost an order of magnitude greater than the sensitivity to harmonic forcing.

## IV. Passive control using a small control ring

In this section, we extend the results from the previous section to the specific case of a thin axisymmetric control ring that is placed in the flow. The force on the flow is equal and opposite to the drag force that the control ring experiences. In a cylindrical co-ordinate system, this force can be modelled by the force on a small circular cylinder. As a simple model, the steady and unsteady components of the force due to a cylinder placed at $(x_c, r_c)$ are given by



(a) Magnitude of steady component of control force $||\bar{\mathbf{F}}||/\alpha$

(b) Magnitude of unsteady component of control force $||\hat{\mathbf{f}}||/\alpha$

Fig. 6. Magnitude of the steady and unsteady force from a control ring placed in the flow. Darker regions have higher magnitude. The thick black line indicates the vortex breakdown bubble.

$$\begin{aligned}
\bar{\mathbf{F}}(x,r) &= -\alpha||\bar{\mathbf{u}}||\bar{\mathbf{u}}\,\delta(x-x_c, r-r_c), && (11)\\
\mathbf{f}'(x,r,t) &= \hat{\mathbf{f}}(x,r)e^{\lambda t}, && (12)\\
\hat{\mathbf{f}}(x,r) &= -\alpha\left(\frac{\bar{\mathbf{u}}\cdot\hat{\mathbf{u}}}{||\bar{\mathbf{u}}||}\bar{\mathbf{u}} + ||\bar{\mathbf{u}}||\hat{\mathbf{u}}\right)\delta(x-x_c, r-r_c) && (13)
\end{aligned}$$

where $\alpha$ is a measure of the magnitude of the force.

Figures 6a and 6b show the magnitude of the steady and unsteady force as a function of the location of the control ring. The steady component is largest near the inlet because the baseflow velocities are high there, whereas the unsteady component is largest further downstream because the amplitude of the global mode is maximum there. We also notice that the unsteady component of the force is about an order of magnitude greater than the steady component. However, we know from equations (7) and (10) that the effect of the control ring on the unstable eigenvalue depends on the overlap of figures 5 and 6. We substitute the expressions in equations (11) and (13) into equations (7) and (10). The total change in the eigenvalue is given by the sum of the contributions from the steady and unsteady components

$$\delta\lambda_{total} = \delta\lambda_{\bar{\mathbf{F}}} + \delta\lambda_{\mathbf{f}'}. \tag{14}$$

In Figure 7, we plot the change in the eigenvalue as a function of the location of the control ring showing the total change as well as the contributions from the steady and unsteady components of the force on the same color scale. These figures identify the locations where passive control using the control ring would be most effective. There is a narrow region upstream of the bubble where the control ring will stabilise the flow and a much larger region downstream of the bubble where the control ring will destabilise the flow. These figures also show that the contribution from the steady component is significantly larger than the contribution from the unsteady component.

## V. Dynamics of the controlled system

We now verify whether the behaviour predicted by our linear sensitivity analysis actually occurs. We model the

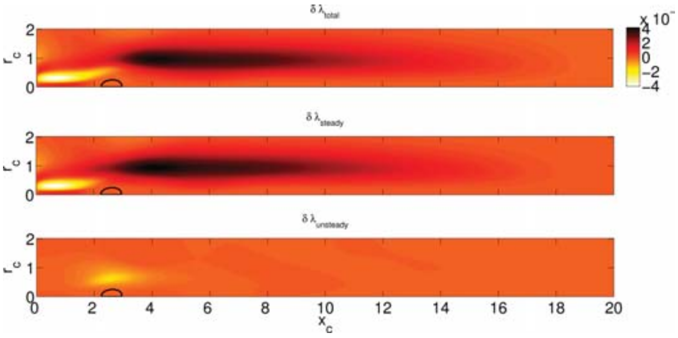


Fig. 7. The change in the eigenvalue as a function of the location of the control ring, from top (a) Total change, (b) Contribution from the steady component of the force and (c) Contribution from the unsteady component of the force. The figures have the same color scale. Light regions indicate regions of stabilisation, whereas dark regions indicate regions of destabilisation. The thick black line indicates the vortex breakdown bubble.

presence of a small control ring at $(x_c, r_c) = (0.66, 0.33)$. This corresponds to the centre of the region of stabilisation in Figure 7(a).

We choose a value of $\alpha = 0.1$ and add the forcing term in equation (11) to our nonlinear equations. We obtain a new steady baseflow and study the linear dynamics of the controlled system. Figure 8 compares the spectrum of the controlled system with that of the uncontrolled system. We notice that the unstable eigenvalue has been stabilised. Our linear sensitivity analysis predicts the eigenvalue of the system with a control ring at $(x_c, r_c) = (0.66, 0.33)$ to be $-0.021 + 2.120\text{i}$ which agrees well with the value obtained from the stability analysis of the controlled system, $-0.018 + 2.119\text{i}$.

The dynamics of the controlled and uncontrolled system can also be seen in Figure 9. We superpose small amplitude random-noise perturbations on the steady baseflows for the controlled and uncontrolled systems and monitor the energy of these perturbations over time. In the uncontrolled system, following some initial transient phase, these perturbations grow linearly. In the controlled system, these perturbations decay linearly.

## VI. Applications and further work

In the previous section, we demonstrated that this approach to the control of spiral vortex breakdown works in theory. We now consider some of the practical issues related to this approach.

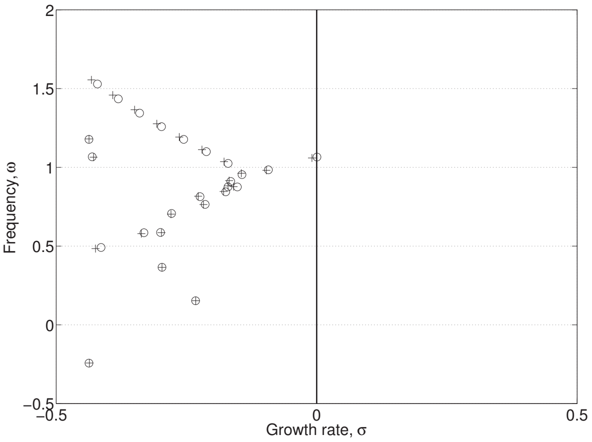


Fig. 8. Spectrum of the linear operator $\mathbf{L}$ for $m = -1$ for the uncontrolled (o) and controlled (+) baseflows, showing the 25 least stable eigenvalues. The marginally unstable eigenvalue in the uncontrolled flow is stable in the controlled flow.

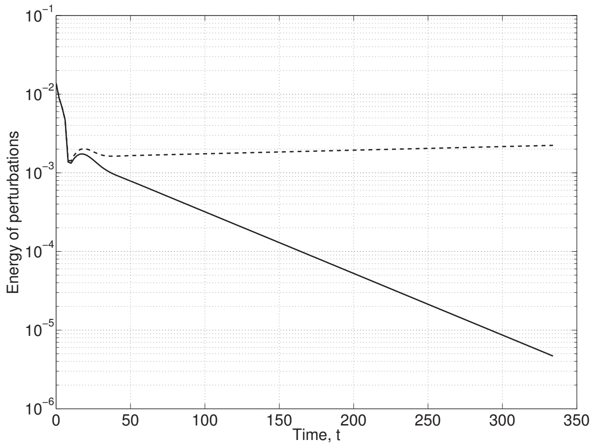


Fig. 9. The evolution of the energy of perturbations for $m = -1$ on top of the baseflows for the uncontrolled (dashed) and controlled (solid) system.

Practical engineering applications feature flows at much higher Reynolds numbers than that considered in this paper. In such cases, obtaining a steady baseflow would be very difficult and it might be easier and more relevant to carry out a sensitivity analysis around the mean flow. This would not be mathematically rigorous but could still provide valuable practical information for designers.

The control ring concept that we have used here is a simple model for a control device. An example of something similar that could be used in practice is a hot wire that forms an axisymmetric ring centred on the axis. The sensitivity of the location of the hot wire to the temperature and velocity can be obtained from the relevant adjoint fields, (namely, the adjoint temperature $\bar{T}^+$ and the adjoint momentum $\bar{\mathbf{m}}^+$. We have chosen to demonstrate this technique for vortex breakdown in an open domain. However, it can just as easily be applied to the control of vortex breakdown in a closed vessel (such as a combustion chamber). In this case, the adjoint pressure $\bar{p}^+$ can be used to obtain the sensitivity to mass injection. This would offer insight into how active control techniques such as blowing or suction could be used to control spiral vortex breakdown.

The tools that have been developed here are now being applied to study swirling flows with variable density and combustion. The overall aim of this project is to develop the capability of producing sensitivity maps of real fuel injectors in combustion chambers. These sensitivity maps would give information about where the design should be changed to promote or suppress certain flow behaviours, taking into account the physical mechanisms that act in the flow.

## VII. Conclusions

In this paper, we have considered a theoretical approach to the control of spiral vortex breakdown. We have shown that spiral vortex breakdown is caused by an unstable eigenmode that grows on top of the steady axisymmetric vortex breakdown bubble. We have used a Lagrangian approach to identify regions where the flow is most sensitive to small steady and unsteady (harmonic) forces. We have found that the regions upstream of the vortex breakdown bubble are most sensitive to steady and unsteady forces.

We then considered passive control using a small control ring placed in the flow. We have identified regions where this control device should be placed to stabilise or destabilise the flow. We have found that there is a narrow region upstream of the bubble where the control ring will stabilise the flow. We have verified this using numerical simulations.

The results from this study have shown that a linear sensitivity analysis can provide useful information for control, based on the underlying physics of the flow. This technique can be easily extended to control spiral vortex breakdown in more complicated geometries.

## Acknowledgement

U. A. Qadri is grateful to Trinity College, Cambridge for financial support. The numerical computations were performed using the Darwin Supercomputer of the University of Cambridge High Performance Computing Service (http://www.hpc.cam.ac.uk/), provided by Dell Inc. using Strategic Research Infrastructure Funding from the Higher Education Funding Council for England.

## References

[1] D. H. Peckham and S. A. Atkinson, “Preliminary results of low speed wind tunnel tests on a gothic wing of aspect ratio 1.0,” Aeronautical Research Council, Tech. Rep., 1957.

[2] M. Ruith, P. Chen, E. Meiburg, and T. Maxworthy, “Three-dimensional vortex breakdown in swirling jets and wakes: direct numerical simulation,” *Journal of Fluid Mechanics*, vol. 486, pp. 331–378, JUL 10 2003.

[3] F. Gallaire, M. Ruith, E. Meiburg, J. Chomaz, and P. Huerre, “Spiral vortex breakdown as a global mode,” *Journal of Fluid Mechanics*, vol. 549, pp. 71–80, FEB 25 2006.

[4] A. M. Mitchell and J. Délery, “Research into vortex breakdown control,” *Progress in Aerospace Sciences*, vol. 37, no. 4, pp. 385–418, 2001.

[5] D. Hill, “A theoretical approach for analyzing the re-stabilization of wakes,” *AIAA Paper 92-0067*, 1992.

[6] O. Marquet, D. Sipp, and L. Jacquin, “Sensitivity analysis and passive control of cylinder flow,” *Journal of Fluid Mechanics*, vol. 615, pp. 221–252, NOV 25 2008.

[7] P. Meliga, D. Sipp, and J.-M. Chomaz, “Open-loop control of compressible afterbody flows using adjoint methods,” *Physics of Fluids*, vol. 22, no. 5, MAY 2010.

[8] D. Sipp, O. Marquet, P. Meliga, and A. Barbagallo, “Dynamics and control of global instabilities in open flows: a linearized approach,” *Applied Mechanics Reviews*, vol. 63, no. 030801, 2010.

[9] W. J. Grabowski and S. A. Berger, “Solutions of Navier-Stokes equations for vortex breakdown,” *Journal of Fluid Mechanics*, vol. 75, no. JUN11, p. 525, 1976.

[10] O. Marquet, D. Sipp, L. Jacquin, and J.-M. Chomaz, “Multiple timescale and sensitivity analysis for the passive control of the cylinder flow,” in *AIAA Paper 2008-4228*, 2008.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Stochastic Neural Network Control of Rigid Robot Manipulator with Passive Last Joint

Jing Li
Department of Mathematics
Xidian University
Xi'an 710071, China
Email: xidianjing@gmail.com

Chenguang Yang
and Phil Culverhouse
School of Computing and Mathematics
Plymouth University
PL4 8AA, UK
Email: {chenguang.yang; P.Culverhouse}@plymouth.ac.uk

Hongbin Ma
School of Automation
Beijing Institute of Technology
Beijing 100081, China
Email: mathmhb@bit.edu.cn

***Abstract*—Stochastic adaptive control of a manipulator with a passive joint which has neither an actuator nor a holding brake is investigated. Aiming at shaping the controlled manipulators dynamics to be of minimized motion tracking errors and joint accelerations, we employ the linear quadratic regulation (LQR) optimization technique to obtain an optimal reference model. Adaptive neural network (NN) control has been developed to ensure the reference model can be matched in finite time, in the presence of various uncertainties and stochastic noise. In addition, due to the stochastic noise, we transform the system equation to the Ito stochastic differential equation (SDE) form and then use the Ito formula to deal with the stochastic terms of the systems. Simulation studies show the effectiveness of the planned trajectory and the feedback control laws.**

*Key Words* – Stochastic NN control, optimization, LQR, model reference control

## I. Introduction

Underactuated robots have received considerable research attention in the last two decades ([1]-[7]). In contrast to conventional robot for which each joint has one actuator and its degree of freedom equals the number of actuators, an underactuated robot has passive joints equipped with no actuators. The underactuation structure make possible for the robots to reduce the weight, energy consumption, and cost of manipulators, which can be applied to the tasks involving an impact, e.g., hitting or hammering an object, will be useful since the impact causes no damage to the joint actuators. It can also contribute to fault tolerance of fully-actuated manipulators in case some of the joint actuators fail.

Though the passive joints are not actuated but they can be controlled by using the dynamic coupling with the active joints, i.e., these passive joints can be indirectly driven by other active joints. The zero torque at the passive joints results in a second-order nonholonomic constraint. This method allows the control of more joints than actuators. In robotics, nonholonomic constraints formulated as nonintegrable differential equations containing time-derivatives of generalized coordinates (velocity, acceleration etc.) are mainly studied. Such constraints include the following: 1) Kinematic constraints which geometrically restrict the direction of mobility; 2) Dynamic constraints due to dynamic balance at passive degrees of freedom where no force or torque is applied. Wheeled vehicles [1], rolling contact between objects [2], trailers [3], [4], and manipulators with nonholonomic gears [5] are mechanical systems which have constraints of the former type. Constraints on space robots [6], [7] belong to the latter. These systems commonly have fewer control inputs than the number of generalized coordinates. Therefore, it is necessary to combine the limited number of inputs skillfully in order to control all the coordinates. So how to efficiently control this kind of nonholonomic systems becomes an interesting research area.

In this paper we consider a $n$-joints under-actuated system with passive last joint and use the LQR optimization approach to derive a reference model for the first $n-1$ joints subsystems, which guarantees motion tracking and achieves the minimized moving accelerations. High order neural networks (HONNs) have been employed to design the adaptive control in order to make the controlled dynamics to match the reference model dynamics in finite time. Instead of leaving the unactuated joint dynamics uncontrolled, a reference trajectory for the last joint is designed to indirectly affect the movements such that the desired motion can be achieved. HONNs also have been employed to construct a reference trajectory generator of the last joint.

## II. Stochastic finite-time attractiveness

Consider the following stochastic nonlinear system with the Ito SDE form

$$dx = f(x)dt + g(x)dw \tag{1}$$

where $x \in \mathbf{R}^n$ is the state, $w$ is an independent $r$-dimensional standard Wiener process defined on the complete probability space $(\Omega, \mathscr{F}, \mathbf{P})$, the Borel measurable functions $f : \mathbf{R}^n \to \mathbf{R}^n$ and $g : \mathbf{R}^n \to \mathbf{R}^{n\times r}$ are locally Lipschitz continuous with $f(0) = 0$ and $g(0) = 0$. Without loss of generality, we use 0 and $x_0$ to denote the initial time and the initial state of the system. The solution of system (1) with the initial state $x_0$ is denoted by $x(t;x_0)$.

The following two definitions come from [8], which will be used to express our system stability.

*Definition 1:* For system (1), define $T(x_0,w) = \inf\{T \geq 0 : x(t;x_0) = 0, \forall t \geq T\}$, which is called the stochastic settling time function.

*Definition 2:* The equilibrium $x = 0$ of system (1) is globally stochastically finite-time attractive, if for $x_0 \in \mathbf{R}^n$, the following conditions hold.

(i) Stochastic settling time function $T_0(x_0,w)$ exists with probability one.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

(ii) Provided that $T_0(x_0,w)$ exists, then $E[T_0(x_0,w)] < \infty$.

For a given $V(x) \in \mathbf{C}^2$, the infinitesimal generator $\mathscr{L}$ with regard to (1) is defined by

$$\mathscr{L}V(x) = \frac{\partial V}{\partial x} f(x) + \frac{1}{2} Tr\left\{ g^T(x) \frac{\partial^2 V}{\partial x^2} g(x) \right\}. \tag{2}$$

We now state the stochastic Lyapunov lemma for stochastic finite-time attractiveness, which is a combination of Corollary 1 in [8], Theorem 3.1 in [9] and Revision of Corollary 1 in [10], and the proof is omitted here for simplicity.

*Lemma 1:* Assume that system (1) admits a unique solution. If there exists a $\mathbf{C}^2$ function $V : \mathbf{R}^n \to \mathbf{R}_+$ and class $\mathbf{K}_\infty$ function $\alpha_1, \alpha_2$, positive numbers $c > 0$ and $0 < \gamma < 1$, such that for $\forall x \in \mathbf{R}^n$ and $t \geq 0$,

$$\begin{aligned} &\alpha_1(\|x\|) \leq V(x) \leq \alpha_2(\|x\|), \\ &\mathscr{L}V(x) \leq -c(V(x))^\gamma, \end{aligned}$$

then the equilibrium $x = 0$ of system (1) is stochastically finite-time attractive, and $E[T_0(x_0,w)] \leq \frac{(V(x_0))^{1-\gamma}}{c(1-\gamma)}$, which implies $T_0(x_0,w) < +\infty$ a.s.

## III. Dynamics of Under-actuated Robot Manipulator

Partition of generalized coordinate vector $q$ as $q = [q_a^T,\ q_b^T]^T$ with $q_a = [q_1, q_2, \ldots, q_{n-1}]^T$ and $q_b = q_n$ such that

$$q_a = I_0 q, \quad I_0 = [I_{[n-1,n-1]}, 0_{[n-1,1]}] \in R^{(n-1)\times n} \tag{3}$$

and $\tau_a = [\tau_1, \tau_2, \ldots, \tau_{n-1}]^T$. The dynamics model of robot manipulator with passive last joint is described as follows:

$$\begin{aligned} &\begin{bmatrix} M_a & M_{ab} \\ M_{ba} & M_b \end{bmatrix} \begin{bmatrix} \ddot{q}_a \\ \ddot{q}_b \end{bmatrix} + \begin{bmatrix} C_a & C_{ab} \\ C_{ba} & C_b \end{bmatrix} \begin{bmatrix} \dot{q}_a \\ \dot{q}_b \end{bmatrix} \\ &\qquad + \begin{bmatrix} g_a \\ g_b \end{bmatrix} + \begin{bmatrix} \dot{w}_a \\ \dot{w}_b \end{bmatrix} = \begin{bmatrix} \tau_a \\ 0 \end{bmatrix} \end{aligned} \tag{4}$$

Define

$$M = \begin{bmatrix} M_a & M_{ab} \\ M_{ba} & M_b \end{bmatrix}, g = \begin{bmatrix} g_a \\ g_b \end{bmatrix},$$
$$C = \begin{bmatrix} C_a & C_{ab} \\ C_{ba} & C_b \end{bmatrix}, \dot{w} = \begin{bmatrix} \dot{w}_a \\ \dot{w}_b \end{bmatrix}$$

Then (4) can be written in a compact form as

$$M\ddot{q} + C\dot{q} + g + \dot{w} = I_0^T \tau_a \tag{5}$$

where $w$ is an $n$-dimensional independent standard Wiener process. The following property are well known for the Lagrange-Euler formulation of robotic dynamics:

*Property 1:* The matrix $M$ is symmetric and positive definite.

Therefore, the blocks $M_a$ and $M_b$ are also invertable and the inverse of matrix $M$ exist and is

$$M^{-1} = \begin{bmatrix} S_b^{-1} & -M_a^{-1} M_{ab} S_a^{-1} \\ M_b^{-1} M_{ba} S_b^{-1} & S_a^{-1} \end{bmatrix} \tag{6}$$

where $S_a$ and $S_b$ are Schur complements of $M_a$ and $M_b$, respectively, defined as $S_a = M_b - M_{ba} M_a^{-1} M_{ab}, S_b = M_a - M_{ab} M_b^{-1} M_{ba}$. Multiplying $I_0 M^{-1}$ on both sides of (5) gives us

$$\ddot{q}_a + I_0 M^{-1} C\dot{q} + I_0 M^{-1} g + I_0 M^{-1} \dot{w} = I_0 M^{-1} I_0^T \tau_a = S_b^{-1} \tau_a \tag{7}$$

Then, multiplying $S_b$ on both sides of the above equation, we have

$$S_b \ddot{q}_a + S_b I_0 M^{-1} C\dot{q} + S_b I_0 M^{-1} g + S_b I_0 M^{-1} \dot{w}_a = \tau_a \tag{8}$$

Define $\mathscr{M} \triangleq S_b \in \mathbf{R}^{(n-1)\times(n-1)}$, $\mathscr{C} \triangleq S_b I_0 M^{-1} C = [\mathscr{C}_a, \mathscr{C}_b] \in \mathbf{R}^{(n-1)\times n}$ with $\mathscr{C}_a \in \mathbf{R}^{(n-1)\times(n-1)}$, $\mathscr{C}_b \in \mathbf{R}^{(n-1)\times 1}$, and $\mathscr{G} = S_b I_0 M^{-1} g, \mathscr{S} = S_b I_0 M^{-1}$, then, we have $q_a$-subsystems as follows

$$\Sigma_{q_a} : \mathscr{M}\ddot{q}_a + \mathscr{C}_a \dot{q}_a + \mathscr{C}_b \dot{q}_b = \tau_a - \mathscr{G} - \mathscr{S}\dot{w}_a \tag{9}$$

At the same time, we obtain the $q_b$-subsystem as follows:

$$\Sigma_{q_b} : M_b \ddot{q}_b + C_b \dot{q}_b + g_b + \dot{w}_b + M_{ba} \ddot{q}_a + C_{ba} \dot{q}_a = 0 \tag{10}$$

*Remark 1:* It should be mentioned that due to the unknown system parameters in the above dynamics formulation, the dynamics matrices $\mathscr{M}, \mathscr{S}$ are actually unknown for control design.However, we still can estimate their regions. So we assume there exist the positive constants $\underline{m}$ and $\bar{m}$ such that $\underline{m} \leq |\mathscr{M}^{-1}| \leq \bar{m}$. But it should be mentioned that these bounds are only used in the stability analysis, and their exact values need not to be known in our controller design.

## IV. Control of subsystem $\Sigma_a$

### A. Subsystem dynamics and optimal reference model

For convenience, defining $\bar{q}_a = [q_a^T,\ \dot{q}_a^T]^T$, $\bar{q}_b = [q_b,\ \dot{q}_b]^T$, and $\bar{q} = [q_a^T,\ q_b,\ \dot{q}_a^T,\ \dot{q}_b]^T$, we rewrite (9) as

$$\dot{\bar{q}}_a = A_a \bar{q}_a + A_b \bar{q}_b + B\mathscr{M}^{-1}(\tau_a - \mathscr{G}) - B\mathscr{M}^{-1}\mathscr{S}\dot{w}_a \tag{11}$$

where

$$\begin{aligned} A_a &= \begin{bmatrix} 0_{[n-1,n-1]} & I_{[n-1,n-1]} \\ 0_{[n-1,n-1]} & -\mathscr{M}^{-1}\mathscr{C}_a \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ A_{a1} & A_{a2} \end{bmatrix} \\ A_b &= \begin{bmatrix} 0_{[n-1,1]} & 0_{[n-1,1]} \\ 0_{[n-1,1]} & -\mathscr{M}^{-1}\mathscr{C}_b \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ A_{b1} & A_{b2} \end{bmatrix} \\ B &= \begin{bmatrix} 0_{[n-1,n-1]}, I_{[n-1,n-1]} \end{bmatrix}^T \end{aligned} \tag{12}$$

For clarity, here and hereafter, the argument $\bar{q}$ of $A_a, A_b, \mathscr{M}, \mathscr{S}$ and $\mathscr{G}$ is omitted.

The control objective is to control the subsystem dynamics (11) to follow a given reference model

$$\dot{\bar{q}}_m = A_m \bar{q}_m + B M_d^{-1} r_m \tag{13}$$

where $\bar{q}_m \in R^{2(n-1)\times 2(n-1)}$ is the desired response of the system,

$$\begin{aligned} A_m &= \begin{bmatrix} 0_{[n-1,n-1]} & I_{[n-1,n-1]} \\ A_{m1}(\bar{q}_m) & A_{m2}(\bar{q}_m) \end{bmatrix} \\ &= \begin{bmatrix} 0_{[n-1,n-1]} & I_{[n-1,n-1]} \\ -M_d^{-1} K_d & -M_d^{-1} C_d \end{bmatrix} \in \mathbf{R}^{2(n-1)\times 2(n-1)} \\ r_m &= -F_\eta(q_d, \dot{q}_d), \quad \bar{q}_m = [q_m^T, \dot{q}_m^T]^T \end{aligned} \tag{14}$$

In order to choose the optimal values of the reference model parameters, we introduce the following performance index:

$$P_I = \int_{t_0}^{t_f} (e_m^T Q e_m + \ddot{q}_m^T M_d \ddot{q}_m) dt. \tag{15}$$

where $Q = \begin{bmatrix} q_1 & 0 \\ & \ddots \\ 0 & q_{n-1} \end{bmatrix}$, which minimizes both the motion tracking error $e_m = q_m - q_d$ and the joints' angular accelerations. In order to apply the LQR optimization technique[11], we rewrite the reference model (13) as

$$\dot{\bar{q}}_m = A_d \bar{q}_m + Bu \tag{16}$$

with

$$\begin{aligned} A_d &= \begin{bmatrix} 0_{[n-1,n-1]} & I_{[n-1,n-1]} \\ 0_{[n-1,n-1]} & 0_{[n-1,n-1]} \end{bmatrix}, \\ u &= -M_d^{-1}[K_d,\ C_d]\bar{q}_m - M_d^{-1} F_\eta(q_d, \dot{q}_d) \end{aligned} \tag{17}$$

Noting that $u = \ddot{q}_m$ and introducing $\bar{Q}$ defined as

$$\bar{Q} = \begin{bmatrix} Q & 0_{[n-1,n-1]} \\ 0_{[n-1,n-1]} & 0_{[n-1,n-1]} \end{bmatrix} \tag{18}$$

we can then rewrite the performance index (15) as

$$P_{\bar{I}} = \int_{t_0}^{t_f} \left( (\bar{q}_m - \bar{q}_d)^T \bar{Q} (\bar{q}_m - \bar{q}_d) + u^T M_d u \right) dt, \tag{19}$$

where $\bar{q}_d = [q_d^T,\ \dot{q}_d^T]^T$. If we regard $u$ as the control input to system (16), then the minimization of (19) subject to dynamics constraint (16) becomes a typical LQR control design problem, where the solution of $u$ that minimizes (19) is

$$u = -M_d^{-1} B^T P \bar{q}_m - M_d^{-1} B^T s \tag{20}$$

where $P$ is the solution of the following differential equation

$$-\dot{P} = PA_d + A_d^T P - PBM_d^{-1} B^T P + \bar{Q}, \quad P(t_f) = 0_{[2(n-1),2(n-1)]}$$

and $s$ is the solution of the following differential equation

$$-\dot{s} = (A_d - BM_d^{-1} B^T P)^T s + \bar{Q}\bar{q}_d, \quad s(t_f) = 0_{[2(n-1)]} \tag{21}$$

Comparing equations (17) and (20), we can see that the matrices $K_d$ and $C_d$ can be calculated in the following manner:

$$[K_d, C_d] = B^T P, \quad F_\eta = B^T s \tag{22}$$

### *B. NN control and model matching*

According to $\mathscr{M}$'s nonsingularity and from the state feedback control for linear systems, we conclude that there exist $K(\bar{q}) \in \mathbf{R}^{(n-1)\times 2(n-1)}, L(\bar{q}) \in \mathbf{R}^{(n-1)\times 2}, T(\bar{q}) \in \mathbf{R}^{(n-1)\times(n-1)}, G(\bar{q}) \in \mathbf{R}^{(n-1)\times 1}$ such that for the control law chosen by

$$\tau_a = G(\bar{q}) + K(\bar{q})\bar{q}_a + L(\bar{q})\bar{q}_b + T(\bar{q}) r_m, \tag{23}$$

the closed-loop system is the same as the reference model (13). By substituting the control law (23) into the system equation (11), the closed-loop system is given by

$$\begin{aligned} d\bar{q}_a = &\{[A_a + B\mathscr{M}^{-1}K(\bar{q})]\bar{q}_a + [A_b + B\mathscr{M}^{-1}L(\bar{q})]\bar{q}_b \\ &+ B\mathscr{M}^{-1}T(\bar{q})r_m + B\mathscr{M}^{-1}[G(\bar{q}) - \mathscr{G}]\}dt \\ &- B\mathscr{M}^{-1}\mathscr{S} dw_a \end{aligned} \tag{24}$$

Comparing it to match the reference model (13), we obtain

$$\begin{aligned} A_a + B\mathscr{M}^{-1}K(\bar{q}) = A_m, \qquad & \mathscr{M}^{-1}T(\bar{q}) = M_d^{-1} \\ A_b + B\mathscr{M}^{-1}L(\bar{q}) = 0_{[2(n-1),2]}, \quad & G(\bar{q}) - \mathscr{G} = 0_{[n-1]} \end{aligned} \tag{25}$$

Then we have

$$K(\bar{q}) = \mathscr{M}([A_{m1}\ A_{m2}] - [A_{a1}\ \ A_{a2}]),\ G(\bar{q}) = \mathscr{G}, \tag{26}$$

$$L(\bar{q}) = -\mathscr{M}[A_{b1}\ A_{b2}],\ T(\bar{q}) = \mathscr{M}M_d^{-1} \tag{27}$$

Unfortunately, according to Remark 1, the dynamic matrices $A_a, A_b, \mathscr{M}, \mathscr{G}$ are not available during practical implementation, and then the exact values of the desired gains $K(\bar{q}), L(\bar{q}), T(\bar{q})$ and $G(\bar{q})$ are also unknown.

We can employ the HONNs [12] to approximate the controller gains as follows

$$K(\bar{q}) = K^*(\bar{q}) + \varepsilon_K,\ L(\bar{q}) = L^*(\bar{q}) + \varepsilon_L \tag{28}$$

$$T(\bar{q}) = T^*(\bar{q}) + \varepsilon_T,\ G(\bar{q}) = G^*(\bar{q}) + \varepsilon_G \tag{29}$$

with

$$K^*(\bar{q}) = [W_K^{\langle T \rangle} \langle\cdot\rangle S_K(\bar{q})], T^*(\bar{q}) = [W_T^{\langle T \rangle} \langle\cdot\rangle S_T(\bar{q})] \tag{30}$$

$$L^*(\bar{q}) = [W_L^{\langle T \rangle} \langle\cdot\rangle S_L(\bar{q})], G^*(\bar{q}) = [W_G^{\langle T \rangle} \langle\cdot\rangle S_G(\bar{q})] \tag{31}$$

where $\langle\rangle$ expresses matrix block-wise operator, defined in [12]. $W_{Ki,j}, W_{Li,k}, W_{Ti,s}, W_{Gi} \in \mathbf{R}^{l\times 1}$ are the NN ideal weights for $K_{i,j}(\bar{q}), L_{i,k}(\bar{q}), T_{i,s}(\bar{q}), G_i(\bar{q})$, respectively $(i = 1, \cdots, n-1; j = 1, \cdots, 2(n-1); k = 1, 2; s = 1, \cdots, n-1)$, $l$ is the number of the neurons. $S_K(\bar{q}), S_L(\bar{q}), S_T(\bar{q}), S_G(\bar{q})$ are the outputs of the bounded basis functions, and $\varepsilon_K, \varepsilon_L, \varepsilon_T, \varepsilon_G$ are the NN approximation errors. For a fixed number of nodes, we know that $\|\varepsilon_K\|, \|\varepsilon_L\|, \|\varepsilon_T\|, \|\varepsilon_G\|$ are bounded, $W_K, W_L, W_T, W_G$ are unknown constant parameters.

Consider the following NN based control law

$$\begin{aligned} \tau_a &= \hat{K}(\bar{q})\bar{q}_a + \hat{L}(\bar{q})\bar{q}_b + \hat{T}(\bar{q})r_m + \hat{G}(\bar{q}) + \tau_r \\ &= [\hat{W}_K^{\langle T \rangle} \langle\cdot\rangle S_K(\bar{q})]\bar{q}_a + [\hat{W}_L^{\langle T \rangle} \langle\cdot\rangle S_L(\bar{q})]\bar{q}_b \\ &\quad + [W_T^{\langle T \rangle} \langle\cdot\rangle S_T(\bar{q})]r_m + [W_G^{\langle T \rangle} \langle\cdot\rangle S_G(\bar{q})] + \tau_r \end{aligned} \tag{32}$$

where $\tau_r$ is a robust control term for closed-loop stability which will be defined later to compensate for the approximation errors of the NNs and to suppress the disturbaces.

Define

$$\begin{aligned} &e = \bar{q}_a - \bar{q}_{am}, \quad \tilde{W}_K = \hat{W}_K - W_K, \\ &\tilde{W}_L = \hat{W}_L - W_L,\ \tilde{W}_T = \hat{W}_T - W_T,\ \tilde{W}_G = \hat{W}_G - W_G \end{aligned} \tag{33}$$

Substituting the control law (32) into the subsystem dynamics (11), using (33), applying the NN approximations (28)-(29), and recalling (25), we obtain the following error equation

$$\begin{aligned} de = &\Big\{ A_m e + B\mathscr{M}^{-1}(\tau_r - \varepsilon_K \bar{q}_a - \varepsilon_L \bar{q}_b - \varepsilon_T r_m - \varepsilon_G) \qquad (34) \\ &+ B\mathscr{M}^{-1}[\tilde{W}_K^{\langle T \rangle} \langle\cdot\rangle S_K(\bar{q})]\bar{q}_a + B\mathscr{M}^{-1}[\tilde{W}_L^{\langle T \rangle} \langle\cdot\rangle S_L(\bar{q})]\bar{q}_b \\ &+ B\mathscr{M}^{-1}[\tilde{W}_T^{\langle T \rangle} \langle\cdot\rangle S_T(\bar{q})]r_m + B\mathscr{M}^{-1}[\tilde{W}_G^{\langle T \rangle} \langle\cdot\rangle S_G(\bar{q})] \Big\} dt \\ &- B\mathscr{M}^{-1}\mathscr{S} dw_a \end{aligned}$$

For stable $A_m$ of the reference model, let $P_m$ be the symmetric positive definite solution of the Lyapunov equation

$$P_m A_m + A_m^T P_m = -Q_m \tag{35}$$

where $Q_m$ is symmetric positive definite.

The following theorem states the stability of the adaptive NN control.

*Theorem 1:* For the system (11), consider the NN based control laws (32). If the updating laws of the weights of the adaptive NNs are given by

$$\begin{aligned}(\dot{\hat{W}}_{Ki}^{\langle T\rangle})^T &= -\Gamma_{Ki}\langle\cdot\rangle S_{Ki}(\bar{q})\bar{q}_a(e^TP_me)e^TP_m(B)_i\\ (\dot{\hat{W}}_{Ti}^{\langle T\rangle})^T &= -\Gamma_{Ti}\langle\cdot\rangle S_{Ti}(\bar{q})r_m(e^TP_me)e^TP_m(B)_i\\ (\dot{\hat{W}}_{Li}^{\langle T\rangle})^T &= -\Gamma_{Li}\langle\cdot\rangle S_{Li}(\bar{q})\bar{q}_b(e^TP_me)e^TP_m(B)_i\\ \dot{\hat{W}}_{Gi} &= -\Gamma_{Gi}S_{Gi}(\bar{q})\bar{q}_b(e^TP_me)e^TP_m(B)_i\end{aligned} \tag{36}$$

and

$$\begin{aligned}\tau_r &= -k_r(e^TP_me)\,\mathrm{sgn}(B^TP_me)-k_2e-k_3(\|B\|\|P_m\|\|e\|)^2\\ k_r &= k_1+k_{r1}+k_{r2}\end{aligned} \tag{37}$$

where $(B)_i$ stands for the $i-$th column of $B$, $k_1,k_2$ are the positive constants, $k_3 \geq 6\bar{m}\bar{\mathscr{S}}^2, k_{r1} \geq \|\varepsilon_K\bar{q}_a+\varepsilon_L\bar{q}_b+\varepsilon_Tr_m+\varepsilon_G\|, k_{r2} \geq \|[\tilde{W}_K^{\langle T\rangle}\langle\cdot\rangle S_K(\bar{q})]\|\|\bar{q}_a\|+\|[\tilde{W}_L^{\langle T\rangle}\langle\cdot\rangle S_L(\bar{q})]\|\|\bar{q}_b\|+\|[\tilde{W}_T^{\langle T\rangle}\langle\cdot\rangle S_T(\bar{q})]\|\|r_m\|+\|[\tilde{W}_G^{\langle T\rangle}\langle\cdot\rangle S_G(\bar{q})]\|$, $\Gamma_{Ki}\in\mathbf{R}^{(2(n-1)\cdot l)\times(2(n-1)\cdot l)}$, $\Gamma_{Ti}\in\mathbf{R}^{((n-1)\cdot l)\times((n-1)\cdot l)}$, $\Gamma_{Li}\in\mathbf{R}^{(2l)\times(2l)}$,$\Gamma_{Gi}\in\mathbf{R}^{l\times l}$ are the symmetric positive definite matrices, then the adaptive NN controller ensures that the closed-loop system are stochastically finite-time attractive, and for each bounded initial condition, and the parameter estimates $\hat{W}_K,\hat{W}_L,\hat{W}_T,\hat{W}_G$ satisfy

$$\begin{aligned}&\mathbf{P}\{\lim_{t\to\infty}\|\hat{W}_K\|,\lim_{t\to\infty}\|\hat{W}_L\|,\lim_{t\to\infty}\|\hat{W}_T\|\text{and}\\ &\lim_{t\to\infty}\|\hat{W}_G\| \text{ exist and are finite}\}=1.\end{aligned} \tag{38}$$

*Remark 2:* In $\tau_r$, we introduced a positive constant $k_2$, which is only used in the practical controller design to improve the controller's smoothness and doesn't affect our stability proof.

*Proof:* Choose the following Lyapunov function

$$V_1 = U_1+U_2,\ U_1=\frac{1}{2\bar{m}}(e^TP_me)^2,$$

$$\begin{aligned}U_2 = &\sum_{i=1}^{n-1}\tilde{W}_{Ki}^{\langle T\rangle}\Gamma_{Ki}^{-1}(\tilde{W}_{Ki}^{\langle T\rangle})^T+\sum_{i=1}^{n-1}\tilde{W}_{Li}^{\langle T\rangle}\Gamma_{Ki}^{-1}(\tilde{W}_{Li}^{\langle T\rangle})^T\\ &+\sum_{i=1}^{n-1}\tilde{W}_{Ti}^{\langle T\rangle}\Gamma_{Ti}^{-1}(\tilde{W}_{Ti}^{\langle T\rangle})^T+\sum_{i=1}^{n-1}\tilde{W}_{Gi}^{T}\Gamma_{Gi}^{-1}\tilde{W}_{Gi}.\end{aligned} \tag{39}$$

Bearing in mind $\underline{m} \leq |\mathscr{M}^{-1}| \leq \bar{m}$ and applying the Ito formula to $V_1$ yield

$$\begin{aligned}\mathscr{L}V_1 \leq\ & 2(e^TP_me)e^TP_mB(\tau_r-\varepsilon_K\bar{q}_a-\varepsilon_L\bar{q}_b-\varepsilon_Tr_m\\ &-\varepsilon_G)+6\bar{m}\bar{\mathscr{S}}(\|B\|\|P_m\|\|e\|)^2\\ &+2\sum_{i=1}^{n-1}[\tilde{W}_{Ki}^{\langle T\rangle}\langle\cdot\rangle S_{Ki}(\bar{q})]\bar{q}_a(e^TP_me)e^TP_m(B)_i\\ &+2\sum_{i=1}^{n-1}[\tilde{W}_{Ti}^{\langle T\rangle}\langle\cdot\rangle S_{Ti}(\bar{q})]r_m(e^TP_me)e^TP_m(B)_i\\ &+2\sum_{i=1}^{n-1}[\tilde{W}_{Li}^{\langle T\rangle}\langle\cdot\rangle S_{Li}(\bar{q})]\bar{q}_b(e^TP_me)e^TP_m(B)_i\\ &+2\sum_{i=1}^{n-1}\tilde{W}_{Gi}^{T}S_{Gi}(\bar{q})(e^TP_me)e^TP_m(B)_i\\ &+2\sum_{i=1}^{n-1}\tilde{W}_{Ki}^{\langle T\rangle}\Gamma_{Ki}^{-1}(\dot{\tilde{W}}_{Ki}^{\langle T\rangle})^T+2\sum_{i=1}^{n-1}\tilde{W}_{Ti}^{\langle T\rangle}\Gamma_{Ti}^{-1}(\dot{\tilde{W}}_{Ti}^{\langle T\rangle})^T\\ &+2\sum_{i=1}^{n-1}\tilde{W}_{Li}^{\langle T\rangle}\Gamma_{Li}^{-1}(\dot{\tilde{W}}_{Li}^{\langle T\rangle})^T+2\sum_{i=1}^{n-1}\tilde{W}_{Gi}^{T}\Gamma_{Ti}^{-1}\dot{\tilde{W}}_{Gi}\end{aligned} \tag{40}$$

Substituting the adaptive laws (36) to (40), and further substituting $\tau_r$ from (37), leads to

$$\begin{aligned}\mathscr{L}V_1 =\ & -2(k_1+k_{r2})(e^TP_me)e^TP_mB\,\mathrm{sgn}(B^TP_me)\\ &-2k_2(e^TP_me)e^TP_mBe\\ \leq\ & -k_0\|e\|^3<0,\|e\|\neq 0,\end{aligned} \tag{41}$$

with $k_0 = 2k_1\|P_m\|^2\|B\| > 0$. According to Theorem 1 [14], expression (41) means both $U_1$ and $U_2$ are bounded in probability and consequently $\|e\|,\|\tilde{W}_K\|,\|\tilde{W}_T\|$, $\|\tilde{W}_L\|,\|\tilde{W}_G\|$ are bounded in probability, i.e., $\mathbf{P}\{\lim_{t\to\infty}\|\hat{W}_K\|$, $\lim_{t\to\infty}\|\hat{W}_L\|$, $\lim_{t\to\infty}\|\hat{W}_T\|$, and $\lim_{t\to\infty}\|\hat{W}_G\|$ exist and are finite$\}=1$.

On the other hand, according to (40), (41) and (37), we have

$$\begin{aligned}\mathscr{L}U_1 \leq\ & -\frac{1}{\bar{m}}(e^TP_me)(e^TQ_me)-k_0\|e\|^3\leq -k_0\|e\|^3\\ =\ & = -k_0\left(\frac{1}{\|P_m\|^2}(\|e\|\|P_m\|\|e\|)^2\right)^{\frac{3}{4}}\\ \leq\ & -k_0\left(\frac{1}{\|P_m\|^2}\right)^{\frac{3}{4}}\left((e^TP_me)^2\right)^{\frac{3}{4}}=-c(U_1)^{\frac{3}{4}}\end{aligned} \tag{42}$$

where $c = k_0\left(\frac{2\bar{m}}{\|P_m\|^2}\right)^{\frac{3}{4}}$. Thus, by Lemma 1, the closed-loop system (34) achieves the finite-time attractiveness, i.e., the matching error $e$ will reach the origin in finite time with probability one. This completes the proof. ■

## V. Reference trajectory generator for $q_b$ subsystem

For the motion of the passive joint, however, to our best knowledge, there has been very little study to discuss the automatic control of its movement until now. In this work, we attempt to set up a framework to design the reference trajectory for manipulating the last joint to track the desired trajectory. As above discussed, after finite time, $q_a$ will exactly track $q_{ad}$, such that the dynamics (10) becomes as follows:

$$\begin{aligned}\ddot{q}_b =\ & -M_b^{-1}C_b\dot{q}_b-M_b^{-1}g_b-M_b^{-1}\dot{w}_b-M_b^{-1}M_{ba}\ddot{q}_{ad}\\ &-M_b^{-1}C_{ba}\dot{q}_{ad}\end{aligned} \tag{43}$$

Let $\varphi = [\varphi_1, \varphi_2]^T = [q_b, \dot{q}_b]^T$, $\phi = [\phi_1^T, \phi_2^T]^T = [q_{ad}^T, \dot{q}_{ad}^T]^T$ and $v = \ddot{q}_{ad}$. Then, equation (43) can be rewritten as

$$\dot{\varphi}_1 = \varphi_2, \dot{\varphi}_2 = f(\varphi, \phi, v) - M_b^{-1}\dot{w}_b \tag{44}$$

with $f(\varphi, \phi, v) = -M_b^{-1}M_{ba}v - M_b^{-1}(C_b\varphi_2 + g_b + C_{ba}\phi_2)$ and $\dot{\phi}_1 = \phi_2, \dot{\phi}_2 = v$.

Consider the desired forward position and forward velocity of the manipulator as $q_{bd}$ and $\dot{q}_{bd}$, respectively. Then, our design objective is to construct a $v$ (subsequently $\phi_1$ and $\phi_2$) such that $\varphi_1$ and $\varphi_2$ of system (44) follow $\varphi_{1m}$ and $\varphi_{2m}$ generated from the following reference model

$$\dot{\varphi}_{1m} = \varphi_{2m}, \dot{\varphi}_{2m} = f_m(q_{bd}, \dot{q}_{bd}, \varphi_m) \tag{45}$$

where $\varphi_m = [\varphi_{1m}, \varphi_{2m}]^T$ and $f_m(q_{bd}, \dot{q}_{bd}, \varphi_m) = -k_1(\varphi_{1m} - q_{bd}) - k_2(\varphi_{2m} - \dot{q}_{bd}) + \ddot{q}_{bd}$. It can be easily checked that the reference model (45) ensures that $\varphi_{1m} \to q_{bd}$ and $\varphi_{2m} \to \dot{q}_{bd}$. According to implicit function theorem based neural network design [13], there must exist a function $f_v : v^* = f_v(q_{bd}, \dot{q}_{bd}, \varphi, \phi)$ such that $f(\varphi, \phi, v^*) = f_m(q_{bd}, \dot{q}_{bd}, \varphi)$, i.e., there exists the ideal HONNs weight vectors such that

$$v^* = [W_v^{*\langle T\rangle}\langle\cdot\rangle S_v(z)] + \varepsilon_v, \quad z = [q_{bd}, \dot{q}_{bd}, \varphi^T, \phi^T]^T \tag{46}$$

where $\varepsilon_v \in \mathbf{R}^{(n-1)\times 1}$ is the neural network approximation error vector. Let us employ HONNs to approximate $v^*$ as follows:

$$\hat{v} = \hat{W}_v^{\langle T\rangle}\langle\cdot\rangle S_v(z) \tag{47}$$

with $\hat{W}_v^{\langle T\rangle} = [\hat{W}_{v1}^T, \hat{W}_{v2}^T, \cdots, \hat{W}_{v(n-1)}^T]$, where $\hat{W}_{vi}^T \in \mathbf{R}^{l\times 1}(i = 1, 2, \cdots, n-1)$ are the neural network weight vectors. Substituting $\hat{v}$ into (44) and using $f(\phi, \varphi, v^*) = f_m(q_{bd}, \dot{q}_{bd}, \varphi)$, we have

$$\begin{aligned} \dot{\varphi}_1 &= \varphi_2 & (48)\\ \dot{\varphi}_2 &= f_m(q_{bd}, \dot{q}_{bd}, \varphi) - M_b^{-1}(\dot{w}_b + M_{ba}([\tilde{W}_v^{\langle T\rangle}\langle\cdot\rangle S_v(z)] - \varepsilon_v)) \end{aligned}$$

where $\tilde{W} = \hat{W} - W^*$. Define $\tilde{\varphi}_1 = \varphi_1 - \varphi_{1m}$ and $\tilde{\varphi}_2 = \varphi_2 - \varphi_{2m}$ such that $\tilde{\varphi} = \hat{\varphi} - \varphi$. Then, the comparison between (45) and (48) yields

$$\begin{aligned} \dot{\tilde{\varphi}}_1 &= \tilde{\varphi}_2 & (49)\\ \dot{\tilde{\varphi}}_2 &= -k_1\tilde{\varphi}_1 - k_2\tilde{\varphi}_2 - M_b^{-1}(\dot{w}_b + M_{ba}([\tilde{W}_v^{\langle T\rangle}\langle\cdot\rangle S_v(z)] - \varepsilon_v)) \end{aligned}$$

*Theorem 2:* Consider the following weight adaptation law for HONN employed in (47)

$$\dot{\hat{W}}_{vi} = \Gamma_{vi}S_{vi}(z)\tilde{\varphi}^T P_W[0 \quad 1]^T - \sigma\Gamma_{vi}\hat{W}_{vi} \tag{50}$$

where $\Gamma_{vi} \in \mathbf{R}^{l\times l}$ and $\sigma$ are suitably chosen as a symmetric positive definite matrix and a positive scalar, respectively. Then, the tracking errors $\tilde{\varphi}_1$ and $\tilde{\varphi}_2$ in (49) will be eventually bounded into a small neighborhood around zero.

*Proof:* Let us rewrite the error dynamics (49) as the form of Ito SDE

$$\begin{aligned} d\tilde{\varphi} &= \left[A_W\tilde{\varphi} - [0 \ 1]^T M_b^{-1}\sum_{i=1}^{n-1} M_{bai}(\tilde{W}_{vi}^T S_{vi}(z) - \varepsilon_{vi})\right]dt\\ &\quad -[0 \ 1]^T M_b^{-1}dw_b \end{aligned} \tag{51}$$

where $M_{bai}$ represents the $i$-th element of vector $M_{ba}$, $A_W = \begin{bmatrix} 0 & 1 \\ -k_1 & -k_2 \end{bmatrix}$ satisfies the Lyapunov equation $A_W^T P_W + P_W A_W = -Q_W$, i.e., for any symmetric positive definite matrix $Q_W$, there exists a symmetric positive definite $P_W$ satisfying the above equation.

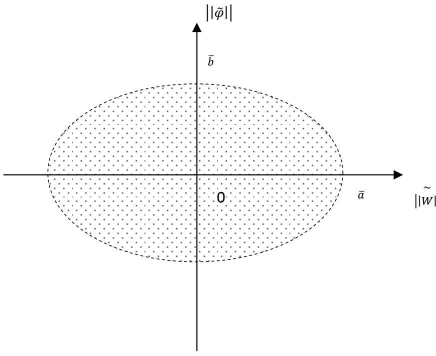


Fig. 1. Bounding set of $\|\tilde{W}_v^{\langle T\rangle}\|$ and $\|\tilde{\varphi}\|$.

Considering the following Lyapunov function

$$V_2(t) = \tilde{\varphi}^T P_W\tilde{\varphi} + M_b^{-1}\sum_{i=1}^{n-1} M_{bai}\tilde{W}_{vi}^T\Gamma_{vi}^{-1}\tilde{W}_{vi}^T \tag{52}$$

and the closed-loop dynamics (51) with the update law (50), we obtain

$$\begin{aligned} \mathscr{L}V_2(t) \leq\ & -\lambda_{Q_W}\|\tilde{\varphi}\|^2 - 2\sigma M_b^{-1}|M_{ba}|\|\tilde{W}_v^{\langle T\rangle}\|^2 + \varepsilon^2\|\tilde{\varphi}\|^2\\ & +\varepsilon^2\|\tilde{W}_v^{\langle T\rangle}\|^2 + \frac{1}{\varepsilon^2}\varepsilon_0^2 M_b^{-1}|M_{ba}|^2\|P_W[0 \ 1]^T\|^2\\ & +\frac{1}{\varepsilon^2}\sigma^2 M_b^{-1}|M_{ba}|^2\|W_v^{*\langle T\rangle}\|^2 + Tr(P_W)(M_b^{-1})^2 \end{aligned}$$

where $|\varepsilon_v| \leq \varepsilon_0$, $\lambda_{Q_W}$ is the minimum eigenvalue of $Q_W$, $\varepsilon$ is any given positive constant and we can choose it sufficiently small. Furthermore, we can choose the suitable $Q_W$ and $\sigma$ making $\lambda_{Q_W} \geq \varepsilon^2, 2\sigma M_b^{-1}|M_{ba}| \geq \varepsilon^2$, and it follows that $\dot{V}_2(t) \leq 0$ in the complementary set of a set $S_b$ defined as

$$S_b \triangleq \left\{(\tilde{\varphi}, \tilde{W}) \left| \frac{\|\tilde{W}_v^{\langle T\rangle}\|^2}{\bar{a}^2} + \frac{\|\tilde{\varphi}\|^2}{\bar{b}^2} - 1 \leq 0 \right.\right\}$$

with $\bar{a} = \dfrac{\bar{c}}{\sqrt{\lambda_{Q_W} - \varepsilon^2}}, \bar{b} = \dfrac{\bar{c}}{\sqrt{2\sigma M_b^{-1}|M_{ba}| - \varepsilon^2}}$, $\bar{c} = \sqrt{\frac{1}{\varepsilon^2}M_b^{-1}|M_{ba}|^2(\varepsilon_0^2\|P_W[0 \ 1]^T\|^2 + \sigma^2\|W_v^{*\langle T\rangle}\|^2) + Tr(P_W)(M_b^{-1})^2}$. Obviously, the set $S_b$ defined above is compact. Hence, by Theorem 1 in [14], it follows that all the solutions of (51) are bounded in probability. The set $S_b$ is shown in Fig. 1 and consists of the closed region bounded by the closed oval arc defined by $\frac{\|\tilde{W}_v^{\langle T\rangle}\|^2}{\bar{a}^2} + \frac{\|\tilde{\varphi}\|^2}{\bar{b}^2} = 1$. Thus, the proof is completed. ■

## VI. Simulation Studies

In this section, the developed trajectory generator and controller will be applied to the cart-pendulum system as shown in Fig. 2 [15]. Let $q_1 = x$ and $q_2 = \theta$, then the dynamics can be described as a fully control subsystem of $q_1$:

$$\Sigma_a : \ddot{q}_1 = \frac{4ml\dot{q}_2^2\sin q_2 - 3mg\sin q_2\cos q_2 + 4F}{4(M+m) - 3m\cos^2 q_2} \tag{53}$$

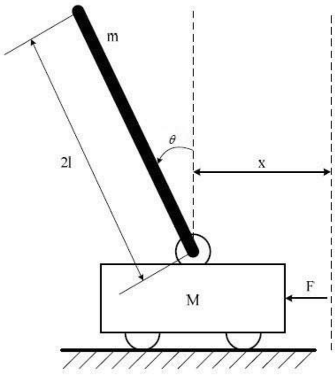


Fig. 2. The cart-pendulum system

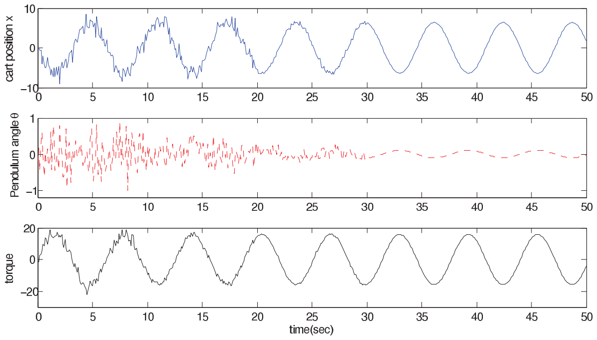


Fig. 3. The simulation results

and an uncontrolled subsystem of $q_2$

$$\Sigma_b : \ddot{q}_2 = \frac{3(M+m)g\sin q_2 - 3m\dot{q}_2^2\sin q_2\cos q_2}{4(M+m)l - 3ml\cos^2 q_2} + \frac{3mg\sin q_2\cos q_2}{4(M+m) - 3m\cos^2 q_2} - \frac{3}{4}\cos q_2\ddot{q}_1 \quad (54)$$

where the mass of cart is 2.4kg; the mass of pendulum is 0.23kg; the length of the pendulum (2l) is 0.36m, $w_1, w_2$ are indepent standard Weiner progresses. The control objective is to make $q_2$ to track $\frac{\pi}{30}\sin(t)$.

The simulation results are shown in Fig. 3. As clearly shown by the simulation results, in the presence of unknown system parameters and external disturbances, the proposed adaptive NN controller is able to guarantee the pendulum's exact tracking of the given trajectory.

## VII. Conclusion

In this paper, adaptive NN control has been designed on the stochastic under-actuated systems for dynamic balance and motion tracking of desired trajectories. The dynamics of the the actuated subsystem has been shaped to follow a reference model, which is derived by using the LQR optimization technique to minimize both the motion tracking error and the transient acceleration. The unactuated subsystem is discussed by suitably generating a reference trajectory. Simulation results have demonstrated the efficiency of the proposed method.

## Acknowledgements

This work is supported by the Marie Curie International Incoming Fellowship H2R Project (FP7-PEOPLE-2010-IIF-275078), the Natural Science Foundation of China under Grants (60804003, 61174045, 61111130208), the International Science and Technology Cooperation Program of China (0102011DFA10950), and the Fundamental Research Funds for the Central Universities (K50510700002, 2011ZZ0104).

## References

[1] J. P. Laumond. Feasible trajectories for mobile robots with kinematic and environment constraints. *Proceeding in Intelligent Autonomous Systems*, Amsterdam, The Netherlands: North Holland Publishing Co., 1987.

[2] Z. Li, J. Canny, Motion of two rigid bodies with roll on constraint. *IEEE Transactions on Robototics and Automattion*, vol. 6, pp. 62-72, 1990.

[3] M. Sampei, T. Tamura, T. Kobayashi, N. Shibui. Arbitrary path tracking control of articulated vehicles using nonlinear control theory. *IEEE Transactions on Control Systems Technology*, vol. 3, no. 1, pp. 125-131, 1995.

[4] O. J. Sodalen. Conversion of the kinematics of a car with n trailers into a chained form. *Proc. IEEE Int. Conf. Robot. Automat.* , pp. 382-387, 1993.

[5] O. J. Sodalen, Y. Nakamura, W. J. Chung. Design of a nonholonomic manipulator. *Proc. IEEE Int. Conf. Robot. Automat.*, pp. 8-13, 1994.

[6] Y. Nakamura and R. Mukherjee. Nonholonomic path planning of space robots via a bidirectional approach. *IEEE Trans. Robot. Automat.*, vol.7, pp. 500-514, 1991.

[7] E. Papadopoulous. Path planning for space manipulators exhibiting nonholonomic behavior. *Proc. IEEE/RSJ Int.Workshop Intell. Robots Syst.*, pp. 669-675, 1992.

[8] W. Chen, L. C. Jiao. Finite-time stability theorem of stochastic nonlinear systems. *Automatica*, vol. 46, pp. 2105-2108, 2010.

[9] J. Yin, S. Khoo, Z. Man, X. Yu. Finite-time stability and instability of stochastic nonlinear systems. *Automatica*, vol. 47, pp. 2671-2677, 2011.

[10] W. Chen, L. C. Jiao. Authors' reply to "Comments on 'Finite-time stability theorem of stochastic nonlinear systems [Automatica 46(2010)2105-2108'". *Automatica*, vol. 47, pp. 1544-1545, 2011.

[11] B. D. O. Anderson and J. B. Moore. *Optimal Control*. London: Prentice Hall, 1989.

[12] S. S. Ge, T. H. Lee, and C. J. Harris. *Adaptive Neural Network Control of Robotic Manipulators*. World Scientific Series in Robotics and Intelligent Systems, Vol. 19, London: World Scientific, 1998.

[13] C. Yang, S. S. Ge, C. Xiang, T. Chai and T. H. Lee. Output Feedback NN Control for two Classes of Discrete-time Systems with Unknown Control Directions in a Unified Approach. *IEEE Transactions on Neural Networks*, vol. 19, no. 11, pp. 1873-1886, 2008.

[14] H. Deng, M. Krstic, R. J. Williams. Stabilization of stochastic nonlinear systems driven by noise of unknown covariance. *IEEE Transactions on Automatic Control*, vol.46, no. 8, pp. 1237-1253, 2001.

[15] D. Chatterjee, A. Patra, and H. K. Joglekar, Swing-up and stabilization of a cart-pendulum system under restricted cart track length. *Systems & Control Letters*, vol. 47, no. 4, pp. 353-362, 2002.

# Wireless Event-driven Networked Predictive Control Over Internet

Wenshan Hu, Hong Zhou, and Qijun Deng

*Abstract*—**In networked control systems, the network uncertainties can degrade the control performance and even result in instability, especially for the Internet-based control system with wireless communication in which the transmission delay could be many times of the sampling period of control systems. NPC (Networked Predictive Control) is an active method to compensate for the effects of network uncertainty. However, control systems need to make long time prediction due to the long transmission distance. Without an accurate mathematical model, the control performance can not be guaranteed. In order to address this problem, a new NPC scheme designed for wireless NPC with long time delay is introduced in this paper. It uses the information of the previous prediction errors to correct the future predictions. A correction algorithm is designed to reduce the predictive errors. To validate the new control schemes, both simulations and experiments have been conducted. The results show that even with "not so" accurate mathematical models, the new schemes can still maintain good control performance in Internet-based NPC with wireless communication.**

## I. Introduction

In the last decade, network technology has been developed rapidly. More and more network technologies have been applied to control systems [1-8]. With the emergence of high speed network technology, a cluster of devices can be linked together economically to form distributed networks. Due to the use of network, especially Internet, the complexity and costs of distributed control systems are reduced greatly and the maintenance of the systems becomes much easier [9-10]. Control systems with devices from different locations can be integrated together using the existing Internet infrastructure which provides a cheap solution for remote data transmission and data exchanges. Internet based control systems allow remote monitoring and adjustment of plants over the Internet around world.

Recently, wireless networked control systems become a popular research area in control theory and industrial applications. Without requiring network cables, devices can be connected into networks using wireless communication, which is promising for remote industrial controls and factory automations. For example, distributed power generation and microgrids [11-12] play more and more important roles in new energy research area. In these systems, unlike big power stations in conventional grids, small and medium size power generation units are located diversely. Wireless wide area networks (WWAN) such as WiMAX, CDMA, GPRS allow rapid development, flexible installation, fully mobile operation which are ideally suitable for these distributed industrial applications.

The technologies of the Internet of Things are developed rapidly in recent years. The research of NCSs is a key part in Internet of Things. In the future Internet of Things, diversely located objects and devices are connected to Internet using both wired and wireless communication. Information such as control commands and measurement data is transmitted through networks.

In wireless networks, the random time delay and data dropouts induced by the data transmission and traffic congestion are even worse than wired ones. These network uncertainties disturb the control performance and even result in instability. Therefore, a wireless networked predictive control systems are studied in this paper. The networked predictive control consists of the control prediction generator and network delay compensator [13-18]. The control predictive generator provides a sequence of future control predictions and the network delay pick up the appropriate control signal from the sequence to eliminate the effects of network transmission delay. Because of the long time delay in wireless networks, the model based prediction may not be accurate due to the model uncertainties. To tackle this problem, NPC is modified and the prediction errors are corrected using the previous predictive errors in this paper.

## II. Design of Wireless Wide Area NPC

### A. Structure of Wireless Wide Area NPC

Fig. 1 is the structure of the proposed Internet-based NPC with wireless communication. It can be separated into two sides: the controller side and the plant side. Both sides are connected to the Internet using wireless connections.

Most of the control calculations are implemented on the controller side. It could be a powerful mainframe computer which has the capability to serve many control loops. The tasks on the plant side are simple. This part can be achieved using a low-cost solution with limited computing capability such as MCU or ARM embedded control board. The diagram of this kind of system is shown in Fig. 2.

Manuscript received September 15, 2011. This work was supported by the National Science Foundation of China under Grant 61004030.

Wenshan Hu Hong Zhou and Qijun Deng are with Department of Automation, School of Mechanical and Power Engineering, Wuhan University, Wuhan, 430072, China (e-mail: wenshan.hu@whu.edu.cn)

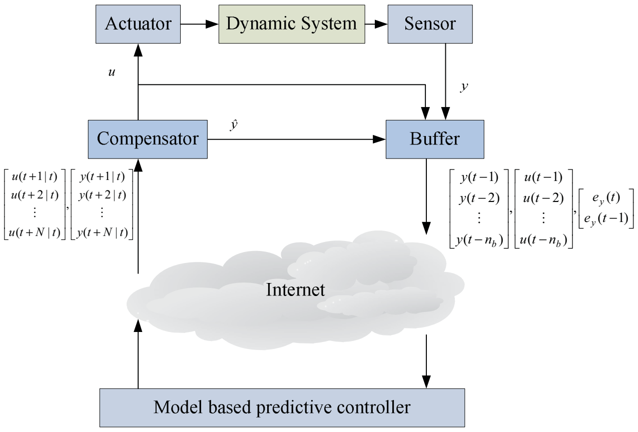


Fig. 1. Structure of Internet-based NPC

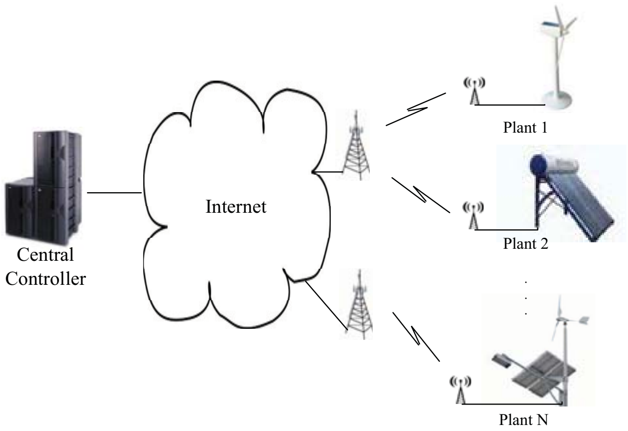


Fig. 2. Diagram of wireless wide area networked control systems

### B. Plant Model and Local Control

It is considered that the dynamic process can be described by the autoregressive moving average model as

$$A(z^{-1})y(t) = B(z^{-1})u(t) \tag{1}$$

where $u(t)$ and $y(t)$ are the open loop input and output of the plant. $A(z^{-1}) \in \Re[z^{-1}, n_a]$ with $a_0 = 1$ and $B(z^{-1}) \in \Re[z^{-1}, n_b]$ with $b_0 = 0$ are the system polynomials.

For the local control without network transmission delay and data dropout, a controller is designed as

$$C(z^{-1})u(t) = D(z^{-1})e(t) \tag{2}$$

where polynomial $C(z^{-1}) \in \Re[z^{-1}, n_c]$ and $c_0 = 0$ and $D(z^{-1}) \in \Re[z^{-1}, n_d]$ and

$$e(t) = r(t) - y(t) \tag{3}$$

where $r(t)$ is the reference input and $y(t)$ is the system output.

### C. Start of a Control Cycle

In this method, a control cycle is initiated by the plant side rather than controller side, which is similar to the method in [14-15]. At each sampling period, the previous plant output sequence $y(t)$, previous control sequence $u(t)$, previous prediction error sequence $e_y(t)$ and a timestamp $t$ which indicates the current plant side time are packed together and sent out in a single packet to the controller side, which initiates a control cycle. The three sequences in the packet are

$$\begin{bmatrix} y(t-1) \\ y(t-2) \\ \vdots \\ y(t-n_l) \end{bmatrix}, \begin{bmatrix} u(t-1) \\ u(t-2) \\ \vdots \\ u(t-n_l) \end{bmatrix}, \begin{bmatrix} e_y(t) \\ e_y(t-1) \end{bmatrix} \tag{4}$$

where $n_l$=max($n_a$, $n_b$, $n_c$, $n_d$).

### D. Design of Controller

The controller is event driven rather than time driven. It is only when the plant side receives the data from the feedback channel does the controller generates control predictions. Without receiving new data from the plant side, it is in an "idle state."

It is assumed that at a time instance, the controller receives a packet from the plant side. Inside the packet, the sequence of plant outputs $y$ (including $y(t), y(t-1), \cdots, y(t-n_a)$ ), previous control sequence $u$ (including $u(t), u(t-1), \cdots, u(t-n_a)$ ) and a time stamp which indicates the time at which the packet is packed and sent out. Because the controller is event-driven and the time delay compensation is based on the round trip time delay measurement, it is noted that variable $t$ indicates the plant side time instance at which the packet is packed and sent out. It has nothing to do with the time instance at which the controller receives the packet. Using the round trip delay prediction, the future control sequence can be generated using the algorithm described below without knowing any information about the controller time at all.

For the sake of simplicity of analysis, it is assumed that the maximum time delay is bounded within $N$ steps. The following defines the operation on the predictions:

$$x(t+i\,|\,t) = q^{-1}x(t+i+1\,|\,t) \quad \text{for } i = 0,1,\cdots \tag{5}$$

$$x(t-1) = q^{-1}x(t\,|\,t) \tag{6}$$

$$x(t-i-1) = q^{-1}x(t-i) \quad \text{for } i = 1,2,\cdots \tag{7}$$

where represents $x(.)$, $y(.)$ or $u(.)$, and $x(t+i|t)$ denotes the $i$-th step-ahead prediction of $x(t)$ based on the previous data up to time $t$.

Based on the data up to plant side time $t$, and the mathematical model of the plant, the one-step plant output prediction with model uncertainty correction can be calculated as

$$y(t+1\,|\,t) = (1 - A(q^{-1}))y(t+1\,|\,t) + B(q^{-1})u(t+1) + c_m \tag{8}$$

where $c_m$ is the correction factor, and $m$ is a counter on the controller side. $c_m$ is calculated using the information of previous prediction errors. The initial value of $m$ is 1 and each time the controller process a packet from the plant side, $m$ is increased by 1. The details of the calculation of $c_m$ are introduced in Section *2.E*.

The corresponding one-step prediction for control signal is

$$\begin{aligned} u(t+1\mid t) &= (1-C(q^{-1}))u(t+1\mid t) \\ &\quad + D(q^{-1})(r(t+1)-y(t+1\mid t)) \end{aligned} \tag{9}$$

Using the same method recursively, the future plant output predictions can be obtained based on the data calculated at the previous step.

$$y(t+k\mid t) = (1-A(q^{-1}))y(t+k\mid t) + B(q^{-1})u(t+k) + c_m \tag{10}$$

where $k$=1,2,…,$N$. The future control predictions are

$$\begin{aligned} u(t+k\mid t) &= (1-C(q^{-1}))u(t+k\mid t) \\ &\quad + D(q^{-1})(r(t+k)-y(t+k\mid t)) \end{aligned} \tag{11}$$

where $k$=1,2,…,$N$.

After $N$-step calculations, the future control sequence $U(t|t)$ and future plant output $Y(t|t)$ are obtained, where

$$U(t\mid t) = [u(t+1\mid t), u(t+2\mid t), \cdots, u(t+N\mid t)]^T \tag{12}$$

$$Y(t\mid t) = [y(t+1\mid t), y(t+2\mid t), \cdots, y(t+N\mid t)]^T \tag{13}$$

They are packed together with the timestamp $t$ and sent back to the plant side.

### *E. Calculation of Model Uncertainty Correction Factor*

In conventional NPC, if the mathematical model of the plant is not accurate enough, the control performance would be degraded greatly. It normally results in big static errors in some practical applications. Because the prediction is not accurate due to the model uncertainty, the controller "thought" the plant "had" reached the target position, so it stops the adjustment of the control signals. However, the plant doesn't follow the controller's prediction. Therefore, the static errors can not be corrected.

In order to cope with this problem, a model uncertainty correction algorithm is introduced in this paper. Based on the history prediction error sequence obtained from the plant side, the controller estimates the correction values for the future prediction. The estimation method is similar to the idea of PI control. Different from the PI control algorithm, the target of the method is to minimize the prediction errors rather than the control residuals. Therefore, the input is the prediction error rather than residual. The transfer function between prediction error $e_y(t)$ and the correction factor $c(t)$ is

$$c(t) = (k_p + k_i \frac{Tz^{-1}}{1-z^{-1}})e(t) = \frac{k_p - (k_p - k_i T)z^{-1}}{1-z^{-1}} e_y(t) \tag{14}$$

where the proportional gain $k_p$ and integral gain $k_i$ are the two parameters of the algorithm. Written in the incremental form, (14) is described as

$$\Delta c(t) = (1-z^{-1})c(t) = (k_p - (k_p - k_i T)z^{-1})e_y(t) \tag{15}$$

where $\Delta c(t)$ is the increase of $c(t)$.

With $\Delta c(t)$ calculated by using (15), the correction factor can be obtained,

$$c(t) = c(t-1) + \Delta c(t) \tag{16}$$

However, due to the stochastic nature of the Internet, the time delay is random. Therefore, the previous correction value $c(t\text{-}1)$ may not be available on the controller side. In order to simplify the analysis, the correction factor on the controller side $c_m$ is used to replace $c(t)$. Equation (16) can be rewritten as

$$c_m = c_{m-1} + \Delta c(t) \tag{17}$$

Combine with (16), the model uncertainty correction factor is obtained as

$$c_m = c_{m-1} + k_p e_y(t) - (k_p - k_i T)e_y(t-1) \tag{18}$$

With the previous correction factor $c_{m\text{-}1}$ and previous prediction errors $e_y(t)$ and $e_y(t\text{-}1)$ available, it is straightforward to calculate the current correction factor $c_m$.

If the transmission delay between the plant side and the controller is constant and there is no data dropout, $c_m$ is equal to the corresponding $c(t)$. However, in real network environment, the packet with plant side timestamp $t$ may arrive earlier than the packet with timestamp $t$-1. In that case, $c_m$ is not $c(t)$ but it is close to $c(t)$.

### *F. Design of Plant Side*

The plant side receives the packet from the controller, in which the timestamp $t$ the control sequence $U(t|t)$ and plant output sequence $Y(t|t)$ are packed. To calculate the RTT delay, the timestamp $t$ is picked up and compared with the current plant side time $t_c$

$$t_d = t_c - t \tag{19}$$

where $t_d$ is the RTT delay.

According to the time delay measurement $t_d$, the $t_d$-th value in the sequence $U(t|t)$ is picked up and applied to the plant actuator.

$$u(t+t_d) = u(t+t_d \mid t) \tag{20}$$

Similarly, the $t_d$-th value in sequence $Y(t|t)$ is also picked up as the predicted plant output.

$$\hat{y}(t+t_d) = y(t+t_d \mid t) \tag{21}$$

The predicted plant output $\hat{y}(t+t_d)$ is then compared the real plant output $y(t+t_d)$. The prediction error $e_y(t+t_d)$ is obtained as

$$e_y(t+t_d) = y(t+t_d) - \hat{y}(t+t_d) \tag{22}$$

The history data of the plant output $y$, control signal $u$ and prediction error $e_y$ are buffered as below

$$\begin{bmatrix} y(t_c-1) \\ y(t_c-2) \\ \vdots \\ y(t_c-n_b) \end{bmatrix}, \begin{bmatrix} u(t_c-1) \\ u(t_c-2) \\ \vdots \\ u(t_c-n_b) \end{bmatrix}, \begin{bmatrix} e_y(t_c) \\ e_y(t_c-1) \end{bmatrix} \qquad (23)$$

These data are packed with the current time stamp $t_c$ and sent out the controller side, which initiates another control cycle.

## III. Simulation Results

In order to validate the proposed method, a speed control system for a cooling fan is considered. The control system is designed to drive a fan to the target speed. The mathematical model is identified as

$$G(z^{-1}) = \frac{B(z^{-1})}{A(z^{-1})} = \frac{0.4662z^{-1} - 0.2843z^{-2}}{1 - 1.396z^{-1} + 0.4681z^{-2}} \qquad (24)$$

where the input is the PWM duty cycle (ranged from 0 to 1) applied to the fan motor and the output is the voltage reading from the speed sensor (ranged is from 0v to 3v). The power supply voltage is 12v and the sampling time is 0.1s.

The following Proportional-integral Controller is designed when the network transmission delay and data dropout is not considered.

$$G_c(z^{-1}) = \frac{D(z^{-1})}{C(z^{-1})} = \frac{0.33 - 0.23z^{-1}}{1 - z^{-1}} \qquad (25)$$

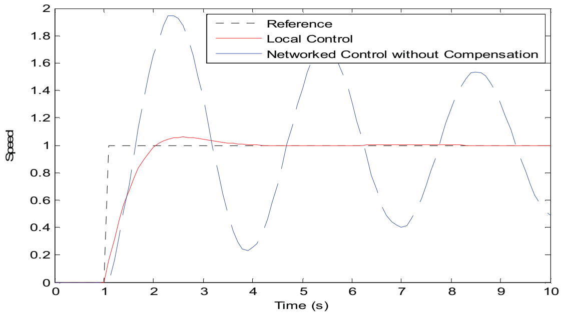


Fig. 3. Simulation of networked control without compensation

The unit step response of local control without communication delay and the networked control without any network uncertainty compensation are shown in Fig. 3. It indicates that the performance of local control is quite good. However, in the networked environment, the round trip time delay varies from 4 to 6 steps (0.4s to 0.6s), the control performance degrades greatly with huge overshoot and very long settling time, which is unacceptable in practical applications.

In order to compensate for the effects of the network uncertainty, the NPC methods proposed in this paper has been adopted. Fig. 4. shows the simulation results. For comparison, two simulations have been conducted. One is the conventional NPC, the other is the NPC with model uncertainty correction designed for NCS on WWANs. Because the mathematical model in simulation is perfectly accurate, the NPC methods can fully compensate for the effects of the network uncertainty. Therefore, there is no model prediction error at all. The results of local control and NPC with and without model uncertainty are exactly the same.

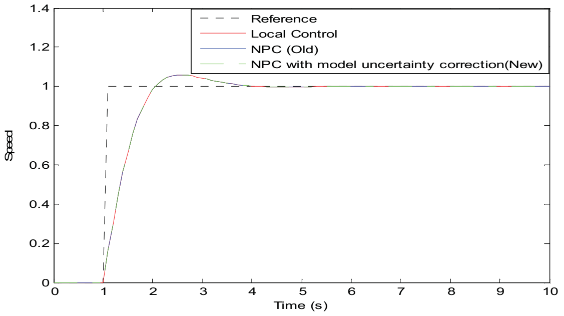


Fig. 4. Simulation of long –distance WNPC without model uncertainties

The mathematical model can not always reflect the dynamical behaviors of the real plant in practical applications. If the power supply voltage of the cooling fan changes from 12v to 10v, correspondingly, the mathematical model (24) will be change to

$$G(z^{-1}) = \frac{B(z^{-1})}{A(z^{-1})} = \frac{0.3885z^{-1} - 0.2369z^{-2}}{1 - 1.396z^{-1} + 0.4681z^{-2}} \qquad (26)$$

If the NPC algorithm still uses the old mathematical model to calculate the predictive control sequence, the prediction errors will be inevitable, which results in big static error shown in Fig. 5. The target output should be 1 for the unit step response, but the real output finally settles down at 0.84. The control performance has been degraded greatly. It shows the prediction errors can result in big static error if the model is not accurate.

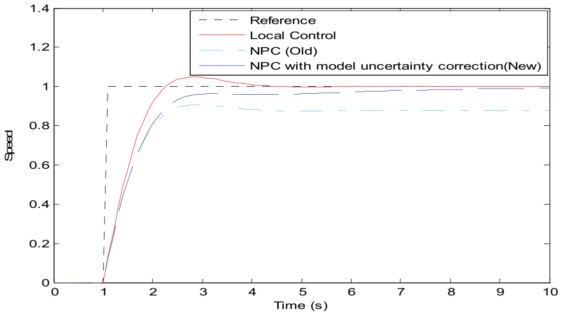


Fig. 5. Simulation of long-distance WNPC with model uncertainties

To cope with this problem, the model uncertainty compensation method proposed in this paper is applied. The previous prediction errors are used to correct the future predictions, and the "PI" correction algorithm tries decreasing the prediction errors as the control process going on. In this case, the both parameters $k_p$ and $k_i$ are 0.1. With the prediction becoming more accuracy, the plant output is quickly approaching the target, as shown in Fig. 5. The control performance has been improved significantly comparing with the conventional NPC.

## IV. Experimental Results

### A. *Test Rig*

A networked cooling fan speed control test rig has been setup to validate the effectiveness of the proposed control algorithm. The whole system consists of three parts: a PC based remote controller on the controller side and a networked module made by Chinese Academy of Sciences (CAS) and a cooling fan on the plant side. The PC and the networked module are connected via Internet and long-distance wireless network. The picture of the test rig is shown in Fig. 6. It can be seen that the cooling fan is placed on the left hand side and the networked module is on the right hand side.

The PC controller with the IP address 202.114.106.29 is located in the campus of Wuhan University. The networked module is connected to the GPRS wireless network provided by China Mobile. The PC controller works as the server side and the networked module works as the client side, so the communication channel between them can be established. The communication protocol adopted in the experiments is UDP. The diagram of the whole wireless networked control system is shown in Fig. 7.

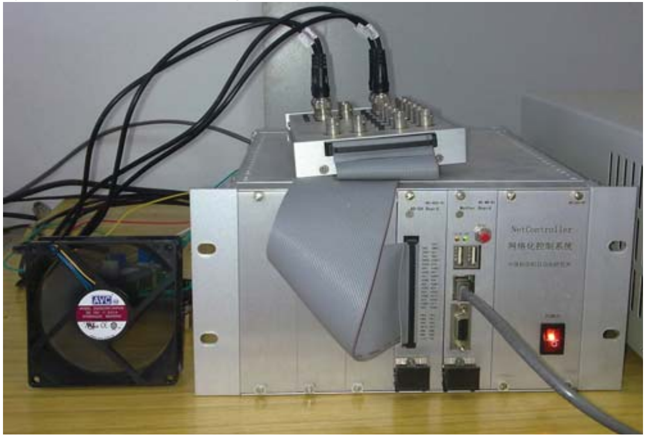


Fig. 6. Cooling fan control test rig

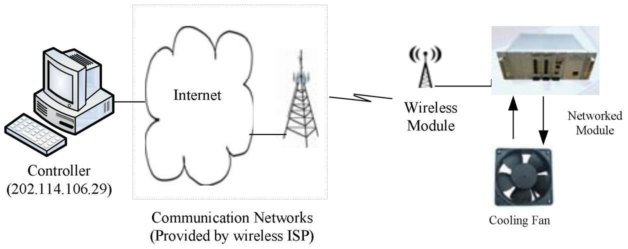


Fig. 7. Experimental diagram

### B. *Experimental Results*

The round trip time delays are measured during the experiments, which are shown in Fig. 8. It can be seen the round trip time delay varies from 4 to 7 steps (0.4 to 0.7s).

Fig 9. shows the experimental results of local control without network and networked control without time delay compensation. The target speed jumps from 1000RPM to 2200 RPM at 1s. It can be seen that in local control, the control performance is good, but network uncertainty degrades the control performance greatly.

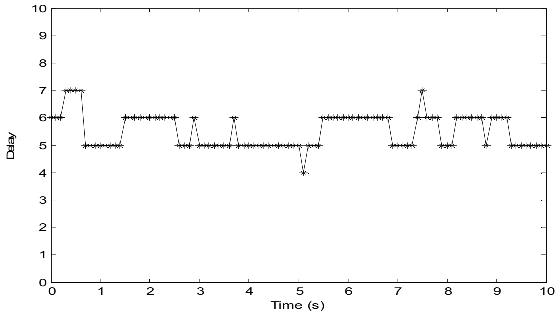


Fig. 8. Round trip time delay measurement

In order to compensate for the network uncertainty, NPC algorithm is adopted. The experiments of both the conventional NPC and NPC with model uncertainty correction proposed in this paper have been conducted. Because the mathematical model identified is quite accurate, both the old and new methods are able to maintain good control performance in the networked environment, which is shown in Fig. 10.

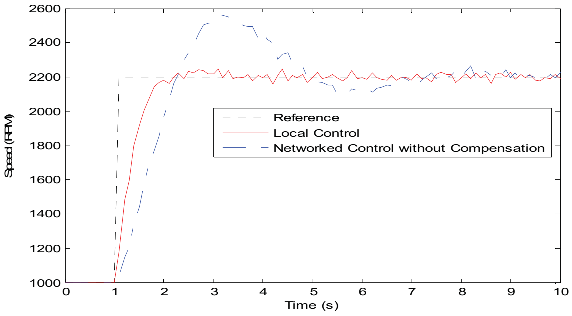


Fig. 9. Results of local control and networked control without delay compensation

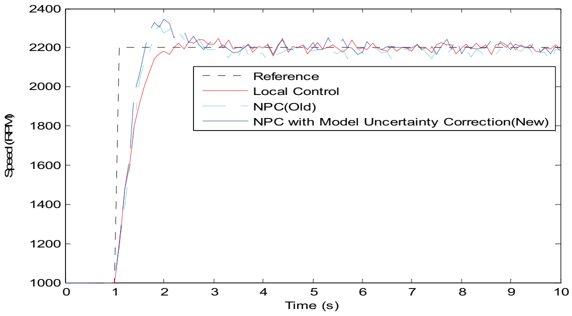


Fig. 10. Results of networked predictive control

When the working condition of the cooling fan is changed, the old method no longer works properly. For example, if the power supply voltage decreases from 12V to 10V, the old method results in a big static error, which is similar to the simulation results. As shown in Fig. 11, the static error is around 200RPM which is not acceptable in practical applications. However, if the new method proposed in this paper has been adopted, the result is very close to the local control even with a not so accurate mathematical model. Same with the simulations, both parameters $k_p$ and $k_i$ for the model correction algorithm are 0.1. The results prove that the proposed methods can compensate for the effects of both network uncertainties and model uncertainties in the NPC system effectively.

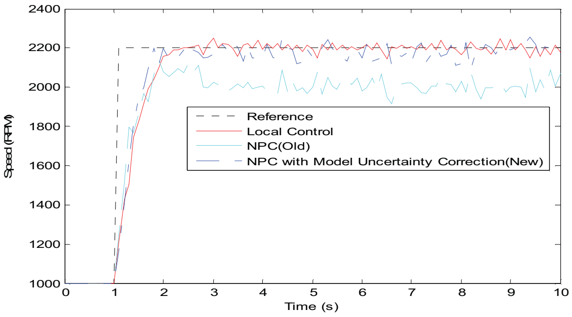


Fig. 11. Results of networked predictive control with 10v power supply

## V. Conclusions

In this paper, a new wireless Internet-based NPC with model uncertainty compensation has been introduced. It is designed for the networked control systems on with wireless communication. In these systems, because of the long transmission delay, the control systems have to cope with the cases that the time delay is many times of the sampling period. The NPC method is able to cope with that situation very well, but it requires very accurate mathematical models. With imperfect models, the control performances would be affected. In this paper, based on the NPC, a model uncertainty correction algorithm is designed to tackle this problem. In order to validate the proposed method, both simulations and experiments have been conducts. The results show that even not so accurate models, the new NPC scheme can still maintain good control performance in WWAN based networked control systems.

## References

[1] H. B. Song, L. Yu, W.A. Zhang, "Networked H1 filtering for linear discrete-time systems, Information Sciences," vol. 181, pp. 686–696, 2011

[2] F. W. Yang, Z. D. Wang, Y. S. Hung, M. Gani, "H1 control for networked systems with random communication delays," IEEE Transactions on Automatic Control, vol. 51, pp.511–518, 2006.

[3] L.Q. Zhang, Y. Shi, T.W. Chen, B. Huang, "A new method for stabilization of networked control systems with random delays," IEEE Transactions on Automatic Control, vol. 50, pp. 1177–1181, 2005.

[4] W. A. Zhang, L. Yu, H. B. Song, "H1 filtering of networked discrete-time systems with random packet losses, Information Sciences," vol. 179, pp. 3944–3955, 2009

[5] D. Yue, Q. L. Han, J. Lam, "Network-based robust H1 control of systems with uncertainty," Automatica vol. 41, pp. 999–1007, 2005...

[6] W. Xu, Z. Zhou, Q. Liu, "Hybrid one-way delay estimation for networked control system," *Advances in Engineering Software*, Vol. 41, 2010, pp. 705-177.

[7] Y. Tipsuwan, and M. Chow, "Gain scheduler middleware: a methodology to enable existing controllers for networked control and teleoperation-part II: teleoperation.," *IEEE Transactions on Industrial Electronics*, Vol. 51, Dec. 2004, pp. 1228 – 1237.

[8] W. A. Zhang and L. Yu. "A robust control approach to stabilization of networked control systems with time-varying delays," *Automatica*, Vol. 45, 2009, pp. 2440-2445

[9] Y. Xia, M. Fu, P. Shi, Analysis and Synthesis of Dynamic Systems with Time-delays, Springer, 2009.

[10] Y. Xia, M. Fu, G.-P. Liu, Analysis and Synthesis of Networked Control Systems, Springer, 2011.

[11] M. Prodanovic and T. C. Green. High-quality Power Generation through Distributed Control of a Power Park Micro Grid. *IEEE Transactions on Industrial Electronics,* vol. 53, pp. 1471-1482, 2006.

[12] F. Katiraei and M. R. Iravani. Power Management Strategies for a Microgrid with Multiple Distributed Generation Units. *IEEE Transactions on Power Systems*, 2006, 21(4): 1821-1831.

[13] J. X. Mu, D. Rees and G. P. Liu, "Design and stability analysis of networked predictive control systems," UKACC International Conference on Control, Bath, 2004

[14] G. P. Liu, S. C. Chai and D. Rees, "Networked Predictive Control of Internet/Intranet Based Systems," Chinese Control Conference, pp. 2024-2029, 2006

[15] G. P. Liu, Y. Xia, D. Rees and W. S. Hu. "Design and stability criteria of networked predictive control systems with random network delay in the feedback channel.," *IEEE Transactions on Systems, Man, and Cybernetics- Part C*, Vol. 37, pp. 173-184

[16] G. P. Liu, Y. Xia, D. Rees J. Chen and W. S. Hu. "Networked predictive control of systems with random network delays in both forward and feedback channels," *IEEE Transactions on Industrial Electronics*, Vol. 54, 2007, pp. 1282-1297

[17] W. S. Hu, G. P. Liu and D. Rees. "Event-driven networked predictive control systems," *IEEE Transactions on Industrial Electronics*, Vol. 54, 2007, pp. 1603-1613

[18] W. S. Hu, G. P. Liu and D. Rees. "Networked predictive control over the Internet using round-trip delay measurement," *IEEE Transactions on Instrumentation and Measurement*, Vol. 57, 2008, pp. 2231-2241

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# Bio-inspired rate control scheme for IEEE 802.11e WLANs

Xin-Wei Yao
College of Computer Science & Technology
Zhejiang University of Technology
Hangzhou, China 310023
Email: yxw_zjut@hotmail.com

Wan-Liang Wang
College of Computer Science & Technology
Zhejiang University of Technology
Hangzhou, China 310023
Email: wwl@zjut.edu.cn

Shuang-Hua Yang
Department of Computer Science
Loughborough University
leicestershire, UK LE11-3TU
Email:S.H.Yang@lboro.ac.uk

Jian-Wei Zheng
College of Computer Science & Technology
Zhejiang University of Technology
Hangzhou, China 310023
Email: zjw@zjut.edu.cn

Yue-Feng Cen
College of Computer Science & Technology
Zhejiang University of Technology
Hangzhou, China 310023
Email: cyf_zjut@hotmail.com

Yan-Wei Zhao
College of Mechanical Engineering
Zhejiang University of Technology
Hangzhou, China 310023
Email: zyw@zjut.edu.cn

***Abstract*—The uncontrolled use of limited resources in conjunction with unpredictable nature of traffic load injection in Wireless Local Area Networks (WLANs) may lead to congestion, and cannot guarantee strict Quality of Service (QoS) required by real-time service. Rate control is an important mechanism for the provisioning of QoS in the IEEE 802.11e WLANs. In this paper, a bio-inspired rate control scheme is proposed based on the extended Lotka-Volterra model, which considers the effects of arrival traffic on the system stability according to the limited network resources and competitions with others traffic flows, and ensures that the network works in unsaturated case and rapidly converge to a global stable equilibrium point (EP). Moreover, all traffic flows are of peaceful coexistence and QoS differentiation. Extensive simulations are conducted to illustrate the performance of the proposed rate control scheme.**

***Index Terms*—Rate control, QoS Differentiation, Bio-inspired model, 802.11e EDCA**

## I. Introduction

WLANs based on the IEEE 802.11 Distributed Coordination Function (DCF) have been widely used in recent years due to their simple deployment and low cost. Since the current DCF can only support best effort traffic, whilst the growing popularity of real-time services and multimedia based applications, it has recently become more critical to tailor IEEE 802.11 Medium Access Control (MAC) protocol to meet the stringent requirements of such services. The IEEE 802.11e Enhanced Distributed Channel Access (EDCA) is proposed to support prioritized QoS[1]. It provides priority-based medium access mechanism with different Arbitration Inter Frame Space (AIFS), initial and maximum Contention Window (CW) size, and the limit of consecutive Transmission Opportunity (TXOP).

Meanwhile, considerable effort was devoted to theoretical analysis of the performance of the 802.11 EDCA [2-5]. However, it cannot guarantee strict QoS required by real-time services such as voice and video without proper network rate control mechanisms. It has been proven that the QoS requirements of the real-time traffic can be satisfied if the input traffic is properly regulated[6]. Meanwhile, it is found that in unsaturated case the 802.11e EDCA achieves the maximum throughput and small delay because of the low collision probability[7,8]. To keep the network operating in unsaturated case, it is crucial to regulate total input traffic. In [6], Chen integrated two admission control schemes and a rate control scheme relying on the average delay estimate and the channel busyness ratio. Lee proposed a bandwidth control scheme by combing the IEEE 802.11e EDCA protocol to overcome the guaranteed bandwidth issue in multi-rate environments [9]. Antoniou proposed a Lotka-Volterra-based congestion control (LVCC) scheme to regulate the input flows without considering the QoS differentiation [10]. Yaghmaee [11] and Chen [12] proposed a priority-based rate control algorithm for service differentiation.

The transmission rate of the traffic can be controlled based on two criteria. First, the injected traffic should not break off the original transmitting traffic, and all the traffic would be co-existing and served. Second, the regulated traffic should be able to promptly access the whole bandwidth in order to utilize the channel efficiently. One may argue that this can be easily achieved if the channel access parameters such as AIFS and CW are set much larger than those for the real-time traffic. However, this approach is problematic in that it will unnecessarily impede the best effort traffic from accessing the channel even when there is no heavy real-time traffic in the network, leading to channel under-utilization and unreasonably large delay for the best effort traffic.

In this paper, to guarantee the strict QoS differentiation and maximize the utility of limited network sources, a bio-inspired rate control scheme is proposed to optimize the transmission rate of each data flow. The stability of the proposed control scheme is analyzed under different network conditions.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

TABLE I
RECOMMENDED IEEE 802.11E EDCA PARAMETER SETTING

| Prio_ | AC | Designation | AIFS | CWmin | CWmax | TXOP |
|---|---|---|---|---|---|---|
| 3 | AC(3) | Voice | 2 | 7 | 15 | 0.003 |
| 2 | AC(2) | Video | 2 | 15 | 31 | 0.006 |
| 1 | AC(1) | Best effort | 3 | 31 | 1023 | 0 |
| 0 | AC(0) | Background | 7 | 31 | 1023 | 0 |

The remainder of this paper is organized as follows: Section 2 introduces the protocol of IEEE 802.11e EDCA. The extended Lotka-Volterra model is introduced in Section 3. In Section 4, we introduce the bio-inspired rate control scheme. Section 5 presents the performance of bio-inspired rate control scheme. The paper is concluded in Section 6.

## II. IEEE 802.11E EDCA PROTOCOL

Since the demands of multimedia applications over WLANs increase tremendously in recent years, which requires the wireless network paradigm to be rethought in the view of need for mechanisms to deliver multimedia content with a certain level of quality of service (QoS). The IEEE 802.11 Task Group E has developed a new standard known as the IEEE 802.11e to provide the QoS support.

According to the IEEE 802.11e standard, the EDCA mechanism extends the DCF mechanism of traditional IEEE 802.11 protocol to enhance the QoS support in the MAC-layer by introducing four access categories $AC[n]$ $(n = 0, 1, 2, 3)$ to serve different types of traffic, which includes AC_VO (for voice traffic), AC_VI (for video traffic), AC_BE (for best effort traffic) and AC_BK (for background traffic). To simplify the notations, we rename four ACs as $AC[3]$, $AC[2]$, $AC[1]$ and $AC[0]$ from the highest priority to the lowest priority in the rest of this paper. Each queue $AC[n]$ transmits packets with an independent channel access parameters including: Minimal Contention Window Size ($CW_{\min}[n]$), Maximal Contention Window Size ($CW_{\max}[n]$), Arbitration Inter-Frame Space Number ($AIFS[n]$), and the limit of consecutive Transmission Opportunity ($TXOP[n]$). The recommended value of each parameter is shown in Table I.

To achieve service differentiation among four ACs, instead of using fixed DCF Interframe Space (DIFS) as in IEEE 802.11 DCF mechanism, EDCA assigns $CW_{\min}$, $CW_{\max}$, $AIFS$ and $TXOP$ with different values to manipulate the successful transmission probability of different types of frames. If one AC has a smaller value of $AIFS$ or $CW_{\min}$ or $CW_{\max}$, then this AC has more chances to access the wireless medium and transmit the waiting or arriving frames. For $AC[i]$ and $AC[j]$ $(0 \leq i < j \leq 3)$, then $CW_{min}[i] \geq CW_{min}[j]$, $CW_{max}[i] \geq CW_{max}[j]$, and $AIFS_i \geq AIFS_j$. Note that in the above inequalities, at least one must be strictly "not equal to" as shown in Table I. Each queue within a station is treated as an individual virtual station, and the backoff procedure of each AC is the same as that of DCF. When a collision occurs among different ACs within the same station, the AC with a higher priority is granted the opportunity to transmit, while the AC with a lower priority is kept waiting, i.e., suffers from a virtual collision.

## III. EXTENDED LOTKA-VOLTERRA MODEL

In order to achieve the better performance and QoS differentiation for IEEE 802.11e EDCA WLANs. We use the bio-inspired model to regulate the input data flows. The nature of the world shows that the dynamics of many biological systems and laws governing them are based on a surprisingly small number of simple generic rules which yield collaborative and effective patterns for resource management, task allocation and social differentiation without the need of any externally controlling entity[13]. For example, population dynamics has traditionally been the dominant branch of mathematical biology which studies how populations of species change in time and space as well as the process that cause these changes.

Population dynamics can be modeled with a simple balance equation that describes how the overall population size of species changes over time as a result of species interaction with each other as well as with non-living parts of their surroundings (i.e. resources). Proposed by Lotka and Volterra, the well-known Lotka-Volterra models concerning ecological population modeling have been extensively investigated in the literature[10,14]. When two or more species live in proximity and share the same basic requirements, they usually compete for resources, food, habit, or territory. A deterministic, competitive Lotka-Volterra system with $n$ species is given by [10,14]

$$\frac{dx_i}{dt} = x_i \left[ r_i - \sum_{j=1}^{n} a_{ij} x_j \right], \quad i = 1, 2, ..., n \tag{1}$$

where $x_i$ represents the population size of species $i$ at time $t$, the constant $r_i$ is the growth rate of species $i$, $n$ is the number of species in an ecosystem, and $a_{ij}$ represents the effect of inter-specific (if $i \neq j$) or intra-specific (if $i = j$) competition. The quotient $r_i / a_{ii}$ is the carrying capacity of the $i$th species in absence of other species. In a vector form, we can rewrite (1) as

$$dx/dt = diag\left(x_1, x_2, ..., x_n\right)\left[B - Ax)\right]$$

where $x = (x_1, x_2, ..., x_n)^{\mathrm{T}}$ is an $n$-dimensional species state vector, $B = (r_1, r_2, ..., r_n)^{\mathrm{T}}$ is the set of growth rate of each species, $A = (a_{ij})_{n \times n}$ is an $n \times n$ matrix, known as the community matrix, and superscript $^{\mathrm{T}}$ denotes the transpose operation.

In this paper, we borrow the idea of population dynamics to analyze the optimal sending rate of each traffic flow under changing network conditions. A WLANs can be considered as analogous to an ecosystem. An ecosystem comprises of multiple species that live together and interact with resources and competitors to meet their needs for survival and coexist. Similarly, a WLANs involves a number of wireless stations. Each station has a limited buffer size to store packets and is able to initiate a traffic flow. Traffic flows can be seen as species that compete with each other for available network resources while traversing the access point to the user. The number of bytes per traffic flow corresponds to the population size of each species. Moreover, the species of an ecosystem have different positions in its biological chain, i.e. some species are much stronger and powerful than others, they will consume more resources from the surrounding environment. This phenomenon can also been seen in WLANs, that the traffic flows with a higher priority are much more important than those with a lower priority, and bandwidth of the network will be priorly allocated to the traffic flows with the highest priority, which means that the packets of the flow with the highest priority have more opportunities to be transmitted. In analogy with ecosystems, the goal of the WLANs is expected to be the coexistence of all traffic flows and achieve QoS differentiation.

## IV. Proposed Bio-inspired rate control scheme

According to the above extended Lotka-Volterra competitive model, there are five correspondences between a WLANs and an ecosystem, i.e. 1) the traffic flows $n$ initiated by each node play the role of competing species; 2) the number of bytes $x_i$ sent by a traffic flow within a given period refers to the population size of a species; 3) the transmission rate of each traffic flow is affected by inter-actions among competing flows as well as the available bandwidth, named inter-specific or intra-specific coefficient $a_{ij}$; 4) the growth rate $r_i$ of each flow refers to the growth rate of each species; 5) the limited bandwidth $N$ of a WLANs can be seen as the resource in the ecosystem.

### A. Considering the QoS Differentiation

The transmission rate evolution of each flow will be driven by variations in available bandwidth of source/relay nodes and by the injection of unpredictable traffic load along the network path towards the user. Each station is expected to initiate a traffic flow when triggered by a specific event or a periodic sensing task. In order to analyze the difference of competition from inter-specific or intra-specific data flows, we redefine the inter-specific competition coefficient of different flows as $a_{ij}$ and the intra-specific competition coefficient of the same type flows as $\beta_i$. To support the service differentiation between the above four ACs in the IEEE 802.11e EDCA as described in Section II, we give two assumptions as follows:

1. The effect of inter-specific competition coming from flow $j$ to flow $i$ can be neglected if the priority of $AC[j]$ is lower than that of $AC[i]$, i.e. The higher priority $AC[i]$ will be granted to transmit frames when colliding with $AC[j]$. However, according to the access procedure of IEEE 802.11e EDCA protocol as shown in Fig.1, (where Short Interframe Space (SIFS) is the small time interval between the data frame and its acknowledgement; PCF Interframe Space (PIFS) is one of the interframe space used in IEEE 802.11 based WLANs.) the $AIFS$ values of $AC[2]$ and $AC[3]$ are identical as listed in Table I, so in this paper, we consider the competition from $AC[2]$ to $AC[3]$. Then the inter-specific competition coefficients are

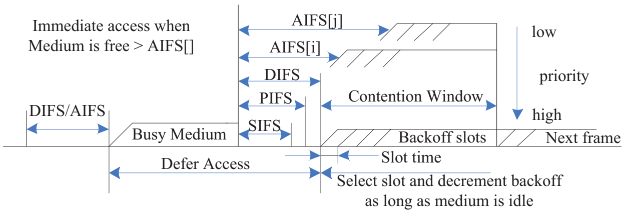


Fig. 1. IEEE 802.11e EDCA access procedure

$$\begin{cases} a_{ij} > 0, \ 0 \leq i < j \leq 3 \\ a_{32} > 0, \quad i = 3, j = 2 \\ a_{ij} = 0, \quad otherwise \end{cases}$$

2. The effect of intra-specific competition of any traffic flow $i$ with its number increasing or decreasing is a constant, i.e. the coefficient $a_{ii} = a_{jj} = \beta$.

Then the proposed bio-inspired rate control scheme based on the extended Lotka-Volterra model for WLANs with considering the QoS differentiation can be written as:

$$\begin{cases} \dot{x}_3(t) = r_3 x_3 \left[ \left(1 - \frac{\beta x_3}{N_3}\right) - a_{32} \frac{x_2}{N_2} \right] \\ \dot{x}_2(t) = r_2 x_2 \left[ \left(1 - \frac{\beta x_2}{N_2}\right) - a_{23} \frac{x_3}{N_3} \right] \\ \dot{x}_1(t) = r_1 x_1 \left[ \left(1 - \frac{\beta x_1}{N_1}\right) - a_{13} \frac{x_3}{N_3} - a_{12} \frac{x_2}{N_2} \right] \\ \dot{x}_0(t) = r_0 x_0 \left[ \left(1 - \frac{\beta x_0}{N_0}\right) - a_{03} \frac{x_3}{N_3} - a_{02} \frac{x_2}{N_2} - a_{01} \frac{x_1}{N_1} \right] \end{cases} \tag{2}$$

where $x_3$, $x_2$, $x_1$ and $x_0$ means the transmission rate of priority queue $AC[3]$, $AC[2]$, $AC[1]$ and $AC[0]$, respectively.

### B. Stability Analysis

The proposed Bio-inspired competitive rate control scheme is adaptive to bursty traffic flows. For example, the transmission rate will reach to the optimal maximal value when other data flows are removed; However, while some higher priority flows are injected into the network unpredictably, the transmission rates of the former flows are reduced smoothly according to the competitions and available bandwidth. So the real-time transmission rate is very important to show the validation of the proposed control scheme and its performance. Based on the nonlinear ordinary differential Equation (2), the real-time transmission rate can be derived:

$$x_3(t) = \frac{N_3 w_3 x_3(0)}{\beta N_2 x_3(0) + [N_3 w_3 - \beta N_2 x_3(0)] e^{-\frac{w_3 r_3}{N_2} t}}$$

$$x_2(t) = \frac{N_2 w_2 x_2(0)}{\beta N_3 x_2(0) + [N_2 w_2 - \beta N_3 x_2(0)] e^{-\frac{w_2 r_2}{N_3} t}}$$

$$x_1(t) = \frac{N_1 (N_2 N_3 - w_1) x_1(0)}{N_2 N_3 \beta x_1(0) + \varphi_1 e^{-\frac{N_2 N_3 - w_1}{N_2 N_3} \cdot r_1 t}}$$

$$x_0(t) = \frac{N_0 (N_1 N_2 N_3 - w_0) \cdot x_0(0)}{\beta N_1 N_2 N_3 x_0(0) + \varphi_0 e^{-\frac{N_1 N_2 N_3 - w_0}{N_1 N_2 N_3} \cdot r_0 t}}$$

where $w_3 = N_2 - a_{32} x_2$, $w_2 = N_3 - a_{23} x_3$, $w_1 = N_2 a_{13} x_3 + N_3 a_{12} x_2$, $w_0 = N_1 N_2 a_{03} x_3 + N_1 N_3 a_{02} x_2 +$

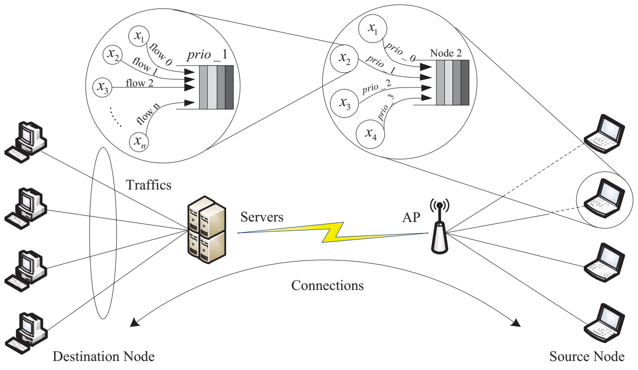


Fig. 2. Network topology

$N_2N_3a_{01}x_1$, $\varphi_1{=}N_1\left(N_2N_3-w_1\right)-\beta N_2N_3x_1\left(0\right)$, $\varphi_0 = N_0\left(N_1N_2N_3-w_0\right)-\beta N_1N_2N_3x_0\left(0\right)$.

In order to guarantee the stability of WLANs, the proposed rate control scheme should be sure that the real-time transmission rate can converge to a stable equilibrium point (EP) as soon as possible. There are sixteen stable EPs by solving the Equation (2), however, it is meaningless unless all the elements in the EP are positive, i.e. $x_0^*>0$, $x_1^*>0$, $x_2^*>0$ and $x_3^*>0$.

Then the stable equilibrium point $x^*$, i.e. the desirable transmission rate of each flow, can be calculated as:

$$\left\{\begin{aligned} x_3^* &= \tfrac{a_{32}-\beta}{a_{23}a_{32}-\beta^2}N_3 \\ x_2^* &= \tfrac{a_{23}-\beta}{a_{23}a_{32}-\beta^2}N_2 \\ x_1^* &= \tfrac{N_1}{\beta}\left[1-\tfrac{a_{12}(a_{23}-\beta)+a_{13}(a_{32}-\beta)}{a_{23}a_{32}-\beta^2}\right] \\ x_0^* &= \tfrac{N_0}{\beta}\left\{1-\tfrac{a_{03}(a_{32}-\beta)+a_{02}(a_{23}-\beta)}{a_{23}a_{32}-\beta^2}\right. \\ &\quad \left. -\tfrac{a_{01}}{\beta}\left[1-\tfrac{a_{13}(a_{32}-\beta)+a_{12}(a_{23}-\beta)}{a_{23}a_{32}-\beta^2}\right]\right\} \end{aligned}\right. \tag{3}$$

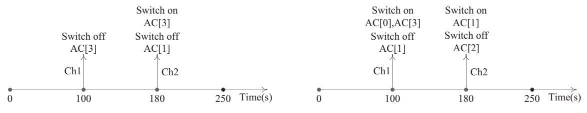
## V. Performance Assessment

### A. Experiment environment and setting

In order to evaluate the performance of the proposed bio-inspired rate control scheme for IEEE 802.11e EDCA WLANs, simulation studies are used to evaluate the performance in terms of graceful performance degradation, self-adaptivity, scalability and service differentiation. In addition, evaluation studies investigate how parameters affect the performance of our approach in terms of stability, convergence and provide effective parameter setting rules. The simulation experiments are conducted in a WLANs, the data flows are transmitted by wireless medium through Access Point (AP) to the Destination Node (DN) as shown in Fig.2. The time interval between successive evaluations of the number of bytes sent by Source Node (SN) is set to 1 second. The parameters (Para) of the proposed rate control scheme shown in Equation (3) are set in Table II.

TABLE II
Parameter setting

| Para | value | Para | value | Para | value | Para | value |
|---|---|---|---|---|---|---|---|
| $a_{01}$ | 1 | $a_{13}$ | 1.3 | $r_1$ | 1 | $N_1$ | 1024 |
| $a_{02}$ | 1.4 | $a_{23}$ | 1.1 | $r_2$ | 1 | $N_2$ | 1024 |
| $a_{03}$ | 1.5 | $a_{32}$ | 1 | $r_3$ | 1 | $N_3$ | 1024 |
| $a_{12}$ | 1.2 | $r_0$ | 1 | $N_0$ | 1024 | $\beta$ | 4 |



(a) Test 1

(b) Test 2

Fig. 3. Removing and/or injecting traffic flow

### B. Stability analysis under different traffic loads

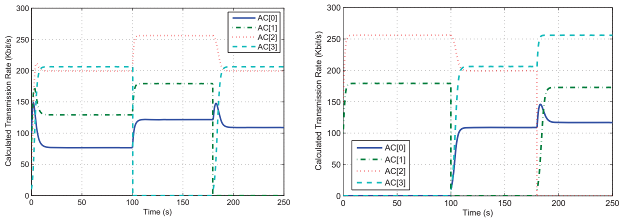
Two random network scenarios (i.e. Test1 and Test2) with two changes, such as removing or injecting traffic flows, are used to evaluate the performance of the proposed rate control scheme, such as the stability, scalability and adaptivity. Each scenario with random initial rate of traffic flows has two changing network states, change1 (Ch1) and change2 (Ch2)), as shown in Fig.3 and three global stable states (i.e. stable, stable1 and stable2) as shown in Table III. In Test1, we switch off the flow $AC[3]$ at $t = 100s$, and then switch $AC[3]$ on and $AC[1]$ off at $t = 180s$. While in Test2, we switch $AC[0]$, $AC[3]$ on and $AC[1]$ off at $t = 100s$, and then switch $AC[1]$ on again and $AC[2]$ off at $t = 180s$. Individual element in Table III is the transmission rate. Moreover, assuming that all traffic flows have the same characteristics of the growth rate $r$, the intra-specific competition coefficient $\beta$ and the maximum capacity $N$ of each data flow.

TABLE III
The EP of the proposed mechanism under different network conditions

| Test | Priority | Initial | Stable | Stable1 | Stable2 |
|---|---|---|---|---|---|
| | | 0s | 100s | 180s | 250s |
| Test1 | $AC[0]$ | 100.0 | 76.6 | 121.6 | 108.9 |
| | $AC[1]$ | 80.0 | 129.2 | 179.2 | 0.0 |
| | $AC[2]$ | 40.0 | 199.3 | 256.0 | 199.3 |
| | $AC[3]$ | 10.0 | 206.1 | 0.0 | 206.1 |
| Test2 | $AC[0]$ | 0.0 | 0.0 | 108.9 | 116.8 |
| | $AC[1]$ | 100.0 | 179.2 | 0.0 | 172.8 |
| | $AC[2]$ | 150.0 | 256.0 | 199.3 | 0.0 |
| | $AC[3]$ | 0.0 | 0.0 | 206.1 | 256.0 |

As observed in Table III, the experiments of Test1 and Test2 have the same values of stable EP, i.e. The point (76.6,129.2,199.3,206.1) is the global stable EP for all flows co-existing in the bandwidth-limited WLANs. When the flow $AC[3]$ becomes extinct at $t = 100s$ for some unknown reasons in Test1, the other priority flows will soon adaptively reach another new EP (121.6,179.2,256.0,0.0). After the instant $t = 180s$, the flow $AC[3]$ is injected into the network again, at the same time $AC[1]$ is switched off, then the system converges to a new EP. Similarly, Test2 also shows the excellent performance of the proposed mechanism in terms of adaptivity, scalability and stability. When $AC[1]$ is switched off and $AC[0]$, $AC[3]$ are switched on at $t = 100s$, the system reaches the stable EP (108.9,0.0,199.3,206.1), and then the WLANs will reach another new stable EP (116.8,172.8,0.0,256.0) after switching off $AC[2]$ and switching on $AC[1]$ at $t = 180s$ in Test2.

Fig. 4 takes a close look at the behavior of all traffic flows with differentiated priority under changing network load. We aim to reveal the process of the system keeping stable after changes in network state. As can be seen, when the data flow is changed for some reasons, the system can re-converge to a new stable EP quickly for its characteristic of self-adaptive. Moreover, the proposed mechanism provides smooth calculated transmission rates for all the traffic flows, which also assists in avoiding the probability of buffer overflow and network congestion. When some high priority emergency data streams are injected into the network, the proposed mechanism can also achieve graceful performance degradation.



(a) Test 1

(b) Test 2

Fig. 4. Calculated transmission rates

### C. Parameter setting and analysis

According to the second assumption of Section IV.A, the effect of intra-specific competition of any traffic flow is a constant, the impact of coefficient $\beta$ on a realistic network environment in investigated in this section. Each scenario, concerning different combinations of $a_{ij}$, $\beta$, $r$ and $N$ values, is executed 10 times and the average values of metrics over all scenarios are presented below.

When $\beta$ increases from 1.5 to 5 as shown in Fig. 5, the difference in transmission rates of all data flows is reduced, the remaining difference in data flow rate is only caused by the different inter-specific competition coefficients. With the increasement of value $\beta$, i.e. the competitive effect of flow $j$ on the sending rate of flow $i$ is much less than that from the inside of flow $i$, the total sending rate of all data flow is decreased. Even though there is no upper bound for $\beta$ value, it is worth pointing out that as $\beta$ increases, the EP value decreases and the quality of the received data at the DN may be reduced.

The phase plane of the scenarios with different values of $\beta$ as shown in Fig. 6 illustrates the stability, rapid convergence and differentiation of the transmission rate of the four different priority flows. As observed in Fig. 6(c) when $\beta = 4$, the transmission rate of each flow can converge to the global stable point without any fluctuations. Moreover, with the increment

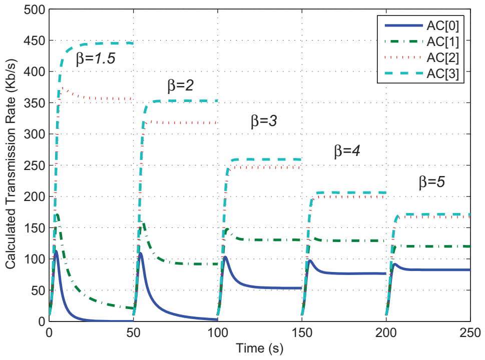
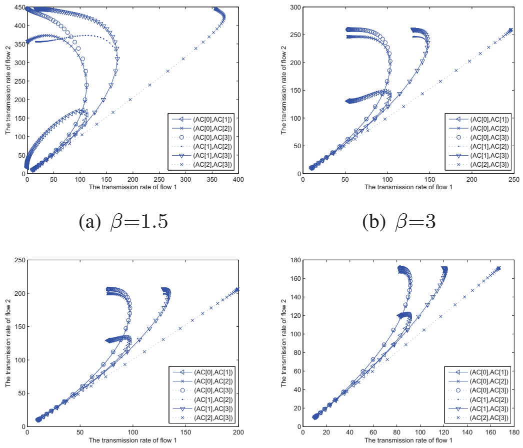


Fig. 5. The equilibrium point with different $\beta$ values

of intra-specific competition coefficient $\beta$, the phase plane inclines to a small region (the EP), which means that there is less difference among four species. However, when the value of intra-specific competition coefficient is equal to 1.5(i.e. the inequality $a_{ij} < \beta$ is not satisfied, for $a_{12} = 1.5$), the network system is not very stable as shown in Fig.6(a), which is identical to the analysis of Section IV.A and can be effectively avoided from proper parameter settings.



(a) $\beta$=1.5

(b) $\beta$=3

(c) $\beta$=4

(d) $\beta$=5

Fig. 6. Phase plane of two species

Fig.7 shows the throughput of different priority flows with $\beta$ varies from 1.5 to 5. The throughput of higher priority AC is much bigger than that of lower priority AC. And according to the curved surface, the value of intra-specific $\beta$ can be set for special purpose.

## VI. Conclusion

In this paper, we proposed a novel bio-inspired rate control scheme to meet the differentiated QoS requirements for various applications in IEEE 802.11e EDCA WLANs. Based on the extended competitive Lotka-Volterra model, the proposed rate control scheme considers traffic flows with differentiated QoS requirements. And the effect of injected bursty traffic

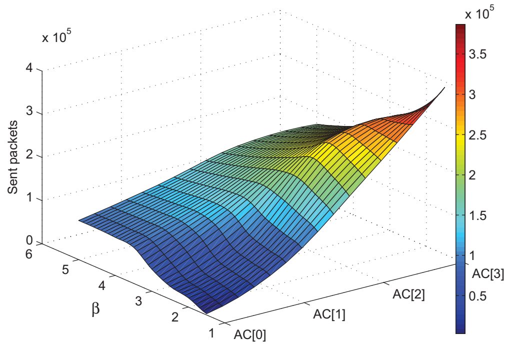


Fig. 7. The throughput of each priority flow with different $\beta$ values

flows on the system stability was also taken into consideration based on the limited network resources and competitions from other traffic flows. It was proven that the scheme has a global stable EP and can fast re-converge to a new EP under changing network conditions, while keeping all traffic flows co-existing and serviced with differentiated QoS. The source traffic rates could be adjusted optimally according to the value of EP. From the analysis of simulation results, we illustrated how the variations of the scheme's parameters influence stability, scalability and distinction of traffic flows. Performance evaluations suggested certain values for parameters $a_{ij}$, $\beta_i$ and $r_i$ that are able to avoid the network congestion and guarantee bandwidth for high priority real time traffic.

## ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (Grant No. 61070043).

## REFERENCES

[1] C.H. Foh, Y. Zhang, Z.F. Ni, J.F. Cai. Scalable Video Transmission over the IEEE 802.11e Networks Using Cross-Layer Rate Control. In Proc. ICC Communications 2007, Glasgow, Scotland, June 2007.

[2] J. Hui, M. Devetsikiotis. A unified model for the performance analysis of IEEE 802.11e EDCA. IEEE Transactions on Communications, 2005, 53(9): 1498-1510.

[3] X.W. Yao, W.L. Wang, S.H. Yang. Video streaming transmission: performance modelling over wireless local area networks under saturation condition. IET Communications, 2012,6(1):13-21.

[4] J.Y. Lee, H.S. Lee. A performance analysis model for IEEE 802.11e EDCA under saturation condition. IEEE Transactions on Communications, 2009, 57(1): 56-63.

[5] P. Serrano, A. Banchs, P. Patras, A. Azcorra. Optimal Configuration of 802.11e EDCA for Real-Time and Data Traffic. IEEE Transactions on Vehicular Technology, 2010, 59(5): 2511-2528.

[6] X. Chen, H.Q. Zhang, X.J. Tian, Y.G. Fang. Supporting QoS in IEEE 802.11e wireless LANs. IEEE Transactions on Wireless Communications, 2006, 5(8): 2217-2227.

[7] J. Hu, G. Min, M.E. Woodward, W. Jia. A Comprehensive Analytical Model for IEEE 802.11e QoS Differentiation Schemes under Unsaturated Traffic Loads. In Pro. ICC Communications, 2008 :241-245.

[8] S.H. Nguyen, H.L. Vu, L.L.H. Andrew. Performance Analysis of IEEE 802.11 WLANs with Saturated and Unsaturated Sources. IEEE Transactions on Vehicular Technology, 2012, 61(1):333-345.

[9] Y. Xiao, Y. Zhang, M. Nolen, J.H. Deng, J. Zhang. A Cross-Layer Approach for Prioritized Frame Transmissions of MPEG-4 Over the IEEE 802.11 and IEEE 802.11e Wireless Local Area Networks. IEEE Systems Journal, 2011, 5(4):474-485.

[10] P. Antoniou, A. Pitsillides. A bio-inspired approach for streaming applications in wireless sensor networks based on the Lotka-Volterra competition model. Computer Communications, 2010, 33(17): 2039-2047.

[11] Yaghmaee M.H., Adjeroh D.A. Priority-based rate control for service differentiation and congestion control in wireless multimedia sensor networks. Computer Networks. 2009,53(11):p. 1798-1811.

[12] Chen Y.L., Lai H.P. Priority-based transmission rate control with a fuzzy logical controller in wireless multimedia sensor networks. Computers and Mathematics with Applications (2011). Doi:10.1016/j.camwa.2011.09.034.

[13] Dressler F., Akan O.B. A survey on bio-inspired networking. Computer Networks, 2010,54(6): 881-900.

[14] Zhu, C. and G. Yin, On competitive Lotka-Volterra model in random environments. Journal of Mathematical Analysis and Applications, 2009,357(1): p. 154-170.

[15] Antoniou, P., Pitsillides, A. Towards a scalable and self-adaptive congestion control approach for autonomous decentralized networks. Third European Symposium on Nature-inspired Smart Information Systems (NiSIS), 2007.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Global Controlled Consensus of Multi-Agent Systems with Different Agent Dynamics and Time-Varying Communication Delay

Wei-Song Zhong
Faculty of Advanced Technology
University of Glamorgan
Cardiff, UK CF37 1DL
Email: wzhong@glam.ac.uk

Guo-Ping Liu
Faculty of Advanced Technology
University of Glamorgan
Cardiff, UK CF37 1DL
Email: gpliu@glam.ac.uk

***Abstract*—This paper investigates the global bounded consensus problem of Networked Multi-Agent Systems exhibiting nonlinear, non-identical agent dynamics with communication time-varying delay. Globally bounded controlled consensus conditions based on pinning control method and adaptive pinning control method are derived. The proposed consensus criteria ensures that all agents eventually move along desired trajectories in terms of boundedness. The proposed controlled consensus criteria generalizes the case of identical agent dynamics to the case of non-identical agent dynamics, and many related results of other researches in this area can be viewed as special cases of the above results. We finally demonstrate the effectiveness of the theoretical results by means of a numerical simulation.**

## I. Introduction

Networked Multi-Agent Systems (NMAS) has attracted many attention due to the broad applications of NMAS in many areas. How to design appropriate protocols and algorithms such that the set of agents can realize common objective, such as consensus, is a critical problem, especially for the case of unreliable information exchange and communication delays, and some relevant important contributions have been made in recent years [1], [2], [3], [4].

The consensus problem requires an agreement to be reached that depends on the state of all agents. The topic has been studied across many fields of science and engineering [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20]. It is noted that the agent dynamics in most existing works are often restricted to linear and identical ones. Obviously, in practice, this is not always the case. The controlled consensus problem of NMAS with nonlinear agent dynamics and communication delay are more complicated and just a few results have been made [21], [22]. In addition, most research in consensus problems usually assume that the final consensus value to be a constant, which may not be the case in the sense that the information state of each agent may be dynamically evolving in time according to some inherent dynamics. It is interesting to study controlled consensus problems where the final consensus value evolves with time or as a function of environmental dynamics.

The behavior of the NMAS with non-identical agent dynamics is much more complicated than the identical case. Usually, no common equilibrium for all agents exists even if each agent has an equilibrium, neither does a consensus manifold exist in the classical sense. The NMAS with non-identical agent dynamics cannot be decoupled into a number of lower dimensional systems exactly like the identical-agent case. Yet, a NMAS with non-identical agents may still exhibit some kinds of consensus behaviors which are far from being fully understood. Certain reasonable and satisfactory boundedness of state motions errors between different agents can be taken as useful consensus properties. The present paper will focus on the global consensus problems of NMAS based on pinning control methods [23], [24], [25], [26], and the proposed controlled consensus property is formulated in terms of certain boundedness of state errors.

The rest of this paper is organized as follows. A controlled continuous-time NMAS model with communication time-delay is presented in Section II. The main results including pinning control and adaptive pinning control bounded consensus criterion are derived in Section III and IV respectively. Section V gives a numerical simulation example to verify the effectiveness of the proposed results, followed by conclusions in Section VI.

## II. Problem description

Let $G = (\mathcal{V}, \mathcal{A})$ be a graph of order $N$ consisting of a set of vertices $\mathcal{V} = \{v_1, v_2, \cdots, v_N\}$ and a set of edges $\mathcal{A} \subseteq \mathcal{V} \times \mathcal{V}$. An edge $(v_j, v_i)$ in graph $G$ means that agent $v_i$ sends some information to agent $v_j$. The set of neighbors of agent $v_i$ is denoted by $\mathcal{N}_i = \{v_j \in \mathcal{V} : (v_j, v_i) \in \mathcal{A}\}$.

We consider a NMAS consisting of $N$ non-identical agents with communication delay:

$$\dot{x}_i = f_i(x_i) + c \sum_{j \in \mathcal{N}_i} a_{ij} \Gamma x_j(t - \tau(t)), i = 1, 2, \cdots, N, \quad (1)$$

where $x_i = (x_{i1}(t), x_{i2}(t), \cdots, x_{in}(t))^T \in R^n$ are the state variables of the agent $v_i$, $f_i(x_i) : R^n \to R^n$ are continuously differentiable mappings with Jacobian $Df_i$, representing the self-dynamics of the agent $v_i$, $c > 0$ denotes the coupling strength, $\Gamma = (\gamma_{ij}) \in R^{n\times n}$ is the inner coupling matrix, and where $\gamma_{ij} \neq 0$ means two connected agents are linked via their $i$th and $j$th state variables, respectively. The adjacency matrix $A = (a_{ij}) \in R^{N\times N}$ (which is symmetric and irreducible) represents the communication topology relation of the NMAS, and is defined by $a_{ij} = a_{ji} = 1(v_j \in \mathcal{N}_i)$, $a_{ij} = 0(v_j \notin \mathcal{N}_i)$ and $a_{ii} = -\sum_{j\neq i} a_{ij}$. $\tau(t)$ is a time-varying coupling delay which reflects the reality that the agent $v_i$ can't obtain information from agent $v_j$ instantaneously.

The average dynamic of all agents is defined by the vector field $\bar{f}(x(t)) = \frac{1}{N}\sum_{k=1}^{N} f_k(x(t))$ with Jacobian $D\bar{f}_i(x(t))$.

The average state trajectory is chosen as the desired moving trajectory

$$s(t) = \frac{1}{N}\sum_{k=1}^{N} x_k(t). \tag{2}$$

We now discuss the problem of global consensus for the system (1). The consensus problem here will be depicted instead via certain boundedness of $x_i(t) - x_j(t)$, $\forall i, j = 1, 2, \cdots, N$ as $t \to \infty$. This better reflects reality as it is impossible for NMAS (1) to achieve exact consensus. To address this case we will focus on making the states of all agents converge to a bounded set.

We denote $x(t)$, $s(t)$, $u(t)$, $e(t)$, $w(t)$, $d_i(t)$ and $V(w(t), t)$ as $x$, $s$, $u$, $e$, $w$, $d_i$ and $V$ respectively.

## III. Linear feedback pinning controller

To achieve the goal, we apply the feedback control strategy on a small fraction $\delta$ $(0 < \delta \leq 1)$ of the agents in system (1). Suppose that nodes $i_1, i_2, \cdots, i_l$ are selected to be under control, where $l = [\delta N]$ stands for the smaller but nearest integer to the real number $\delta N$. This controlled NMAS can be described as

$$\begin{cases} \dot{x}_{i_k} = f_{i_k}(x_{i_k}) + c\sum_{j\in\mathcal{N}_i} a_{i_k j}\Gamma x_j(t-\tau(t)) \\ \qquad +u_{i_k}, \qquad\qquad 1 \leq k \leq l, \\ \dot{x}_{i_k} = f_{i_k}(x_{i_k}) + c\sum_{j\in\mathcal{N}_i} a_{i_k j}\Gamma x_j(t-\tau(t)), \\ \qquad\qquad\qquad\qquad l+1 \leq k \leq N. \end{cases} \tag{3}$$

The local linear negative feedback control law is chosen as follows:

$$\begin{cases} u_{i_k} = -d_{i_k}(x_{i_k} - s), & 1 \leq k \leq l, \\ u_{i_k} = 0, & l+1 \leq k \leq N, \end{cases} \tag{4}$$

where the feedback gain $d_{i_k} > 0$.

Combine (3) and (4) and rearrange the order of the nodes in the network. Let the first $l$ nodes be controlled, and $e_i = x_i - s$, $i = 1, 2, \cdots, N$. It's obvious that $\frac{c}{N}\sum_{k=1}^{N}\sum_{j\in\mathcal{N}_i} a_{kj}\Gamma x_j(t-\tau(t)) = 0$ and $\sum_{i=1}^{N} e_i = 0$. Then by applying the Newton-Leibniz formula, error systems can be written as

$$\begin{cases} \dot{e}_i = D\bar{f}(s)e_i + c\sum_{j\in\mathcal{N}_i} a_{ij}\Gamma e_j(t-\tau(t)) \\ \quad + \int_0^1 (Df_i(s+\tau e_i) - D\bar{f}(s))e_i d\tau \\ \quad - \frac{1}{N}\sum_{k=1}^{N}\int_0^1 Df_k(s+\tau e_k)e_k d\tau \\ \quad + f_i(s) - \bar{f}(s) - d_i e_i, \quad 1 \leq i \leq l, \\ \dot{e}_i = D\bar{f}(s)e_i + c\sum_{j\in\mathcal{N}_i} a_{ij}\Gamma e_j(t-\tau(t)) \\ \quad + \int_0^1 (Df_i + \tau e_i) - D\bar{f}(s))e_i d\tau \\ \quad - \frac{1}{N}\sum_{k=1}^{N}\int_0^1 Df_k(s+\tau e_k)e_k d\tau \\ \quad + f_i(s) - \bar{f}(s), \quad l+1 \leq i \leq N. \end{cases} \tag{5}$$

The following work will focus on simplifying the error systems (5) by means of a series of transformations using a procedure similar to [22].

Define the following matrix

$$D = diag(D_1, D_2, \cdots, D_N) \in R^{nN\times nN},$$

where $D_i = diag\{-d_i, -d_i, \cdots, -d_i\} \in R^{n\times n}$.

Let $e = (e_1^T, e_2^T, \cdots, e_N^T)^T$, then (5) becomes

$$\dot{e} = \bar{\Sigma}(t)e + cA\otimes\Gamma e(t-\tau(t)) + I(t)e - \frac{1}{N}H(t)e + F(t), \tag{6}$$

where $I(t) = diag\{\int_0^1 (Df_1(s+\tau e_1) - D\bar{f}(s))d\tau \cdots \int_0^1 (Df_N(s+\tau e_N) - D\bar{f}(s))d\tau\}$, $\bar{\Sigma}(t) = I_N \otimes D\bar{f}(s) + D$, $H^T(t) = (H_1^T(t), \cdots, H_N^T(t))$, $H_i(t) = (\int_0^1 Df_1(s + \tau e_1)d\tau, \cdots, \int_0^1 Df_N(s + \tau e_N)d\tau)$, $F_i^T(t) = (f_1^T(s) - \bar{f}^T(s), \cdots, f_N^T(s) - \bar{f}^T(s))$.

Since $A$ is symmetric and irreducible, according to [22], there exists a unitary matrix $\Phi = (\varphi_{ij})_{N\times N} = (\Phi_1, \Phi_2, \cdots, \Phi_N)$. This together with $w(t) = (\Phi^T \otimes I_n)e$ gives

$$\begin{aligned} \dot{w} &= (\Phi^T\otimes I_n)\bar{\Sigma}(t)(\Phi\otimes I_n)w \\ &\quad + (\Phi^T\otimes I_n)(cA\otimes\Gamma)(\Phi\otimes I_n)w(t-\tau(t)) \\ &\quad + (\Phi^T\otimes I_n)I(t)(\Phi\otimes I_n)w + (\Phi^T\otimes I_n)F(t) \\ &\quad - \frac{1}{N}(\Phi^T\otimes I_n)H(t)(\Phi\otimes I_n)w. \end{aligned} \tag{7}$$

Note that $H(t) = \sqrt{N}\sum_{k=1}^{N}(\mathbf{0}\ \cdots\ \mathbf{0}\ \bar{\Phi}_k\ \mathbf{0}\ \cdots\ \mathbf{0}) \otimes \int_0^1 Df_k(s + \tau e_k)d\tau$, where $\bar{\Phi}_k$ stands for the matrix with its $k$-th column equal to $\Phi_1$ and the remaining elements are zero. Then we have $\frac{1}{N}(\Phi^T \otimes I_n)H(t)(\Phi \otimes I_n) = \frac{1}{\sqrt{N}}\sum_{k=1}^{N}(\mathbf{0}\ \cdots\ \mathbf{0}\ I_k\ \mathbf{0}\ \cdots\ \mathbf{0})\otimes\int_0^1 Df_k(s+\tau e_k)d\tau(\Phi\otimes I_n)$, where $I_k$ stands for the matrix with its $k$-th column equals $(1\ 0\ \cdots\ 0)^T$ and the remaining of its elements are zero.

Thus, a simple calculation gives $\frac{1}{N}(\Phi^T \otimes I_n)H(t)(\Phi \otimes I_n) = \frac{1}{\sqrt{N}}\sum_{k=1}^{N}\begin{pmatrix} \Upsilon_k & 0 \end{pmatrix}^T \otimes \int_0^1 Df_k(s(t) + \tau e_k(t))d\tau$, where $\Upsilon_k \in R^{1\times N}$ and $0 \in R^{(N-1)\times N}$. Therefore, $\dot{w} = \bar{\Sigma}(t)w + c\Lambda\otimes\Gamma w(t-\tau(t)) + (\Phi^T\otimes I_n)I(t)(\Phi\otimes I_n)w - \begin{pmatrix} * & 0 \end{pmatrix}^T w + (\Phi^T \otimes I_n)F(t)$. Since $w_1 \equiv 0$, we only need to consider $w_2, w_3, \cdots, w_N$. Rewriting in the component form we have

$$\begin{aligned} \dot{w}_i &= \Sigma_i(t)w_i + c\lambda_i\Gamma w_i(t-\tau(t)) + (\Phi_i^T\otimes I_n)F(t) \\ &\quad + (\Phi_i^T\otimes I_n)I(t)(\Phi\otimes I_n)w,\ i = 2, 3, \cdots, N, \end{aligned} \tag{8}$$

where $\Sigma_i = \bar{D}f(s) + D_i$.

So far, we have transferred the consensus problem of system (1) to the stability problem of the $N-1$ of $n-$dimensional systems.

**Theorem 1** Suppose there exist positive definite matrices $P_i(t) \in \mathcal{PC}^1_{n\times n}$, $Q_i$ and constants $\zeta > 0$, $\gamma \geq 0$, $a > 0$ and $b > 0$ such that

$$a\|x\|^2 \leq x^T P_i(t) x + \int_{t-\tau(t)}^{t} w_i^T(\alpha) Q_i w_i(\alpha) d\alpha \leq b\|x\|^2,$$
$$\forall t \in R^+,\ x \in R^n, i = 2, 3, \cdots, N, \tag{9}$$

$$\dot{P}_i(t) + P_i(t)\Sigma_i(t) + \Sigma_i^T(t) P_i(t) + Q_i$$
$$+ c^2\lambda_i^2 P_i(t)\Gamma Q_i^{-1}\Gamma^T P_i(t) + \zeta I \leq 0,\ i = 1, 2, \cdots, N, \tag{10}$$

$$\|I(t)\| \leq \gamma,\ i = 1, 2, \cdots, N. \tag{11}$$

Let

$$\mu(t) = \|F(t)\| \tag{12}$$

be bounded and

$$\beta = (\sum_{i=2}^{N} \|P_i(t)\|^2)^{\frac{1}{2}}, \tag{13}$$

if $\zeta > 2\gamma\beta$, then system (6) converges to the set

$$M = \{e | \|e\| \leq \frac{2b}{a}\frac{\beta \overline{lim}_{t\to\infty}\mu(t)}{\zeta - 2\gamma\beta - \delta}\}, \tag{14}$$

for any time-varying delay $\tau(t) > 0$, namely, $e(t) = x_i(t) - \frac{1}{N}\sum_{k=1}^{N} x_k(t) \to \Omega$ as $t \to \infty$, where $\delta > 0$ is any constant satisfying $\delta < \zeta - 2\gamma\beta$, Furthermore, the NMAS (1) achieves bounded consensus for any fixed time delay $\tau(t) > 0$, $0 \leq \dot{\tau}(t) \leq 1$.

**Proof.** Choose the following Lyapunov-Krasovskii functional as

$$V = \sum_{i=2}^{N} V_i, \tag{15}$$

$$V_i = w_i^T P_i(t) w_i + \int_{t-\tau(t)}^{t} w_i^T(\alpha) Q_i w_i(\alpha) d\alpha. \tag{16}$$

Differentiating (16) along the trajectory of (8) gives

$$\begin{aligned}\dot{V}_i &= w_i^T(\dot{P}_i(t) + P_i(t)\Sigma_i(t) + \Sigma_i^T(t)P_i(t) + Q_i)w_i \\ &+ 2w_i^T P_i(t)(\Phi_i^T \otimes I_n) I(t)(\Phi_i \otimes I_n) w \\ &+ 2w_i^T P_i(t)(\Phi_i^T \otimes I_n) F(t) + 2w_i^T(c\lambda_i P_i(t)\Gamma) w_i(t-\tau(t)) \\ &- w_i^T(t-\tau(t)) Q_i w_i(t-\tau(t)).\end{aligned} \tag{17}$$

Applying the Young Inequality to the equality (17) results in

$$\begin{aligned}\dot{V}_i &\leq w_i^T(\dot{P}_i(t) + P_i(t)\Sigma_i(t) + \Sigma_i^T(t)P_i(t) + Q_i \\ &+ c^2\lambda_i^2 P_i(t)\Gamma Q_i^{-1}\Gamma^T P_i(t))w_i + 2w_i^T P_i(t)(\Phi_i^T \otimes I_n)F(t) \\ &+ 2w_i^T P_i(t)(\Phi_i^T \otimes I_n) I(t)(\Phi \otimes I_n) w.\end{aligned} \tag{18}$$

Condition (10) implies that the first term on the right hand side of (18) satisfies

$$\begin{aligned}&w_i^T(\dot{P}_i(t) + P_i(t)\Sigma_i(t) + \Sigma_i^T(t)P_i(t) + Q_i \\ &\quad + c^2\lambda_i^2 P_i(t)\Gamma Q_i^{-1}\Gamma^T P_i(t))w_i \leq -\zeta\|w_i\|^2.\end{aligned} \tag{19}$$

The second term on the right hand side of (18) satisfies

$$2w_i^T P_i(t)(\Phi_i^T \otimes I_n) F(t) \leq 2\mu(t)\|P_i(t)\|\|w_i\|. \tag{20}$$

Applying condition (11) we know the third term on the right hand side of (18) satisfies

$$2w_i^T P_i(t)(\Phi_i^T \otimes I_n) I(t)(\Phi_i \otimes I_n) w \leq 2\gamma\|P_i(t)\|\|w_i\|\|w\|. \tag{21}$$

Since $V = \sum_{i=2}^{N} V_i$, we have

$$\begin{aligned}\dot{V} &= \sum_{i=2}^{N} \dot{V}_i \\ &= -\zeta\|w\|^2 + 2(\gamma\|w\| + \mu(t))\sum_{i=2}^{N}\|w_i\|\|P_i(t)\| \\ &\leq -\zeta\|w\|^2 + 2(\gamma\|w\| + \mu(t))\|w\|(\sum_{i=2}^{N}\|P_i(t)\|^2)^{\frac{1}{2}} \\ &= \|w\|((2\gamma\beta - \zeta)\|w\| + 2\beta\mu(t)).\end{aligned} \tag{22}$$

Thus, when

$$\|w\| \geq \frac{2\beta\mu(t)}{\zeta - 2\gamma\beta - \delta}, \tag{23}$$

we have

$$\dot{V} \leq -\delta\|w\|^2. \tag{24}$$

Applying the result in [22] completes the proof.

## IV. Adaptive pinning controller

In this section, we will derive globally consensus criteria via direct adaptive pinning control method. Without loss of generality, we still assume that the first $l$ agents are selected as pinned agents with the adaptive controllers:

$$\begin{cases} u_i = -d_i(x_i - s), & 1 \leq i \leq l, \\ \dot{d}_i = h_i e_i^T P_i(t) e_i, & \\ u_i = 0, & l+1 \leq i \leq N, \end{cases} \tag{25}$$

where constant $h_i > 0$ and positive definite matrix $P_i(t) \in R^{n\times n}$. Applying Newton-Leibniz formula, then the error N-

MAS can be rewritten as

$$\begin{cases} \dot{e}_i = D\bar{f}(s)e_i + c\sum_{j\in\mathcal{N}_i} a_{ij}\Gamma e_j(t-\tau(t)) \\ \qquad + \int_0^1 (Df_i(s+\tau e_i) - D\bar{f}(s))e_i d\tau \\ \qquad -\frac{1}{N}\sum_{k=1}^N \int_0^1 Df_k(s+\tau e_k)e_k d\tau \\ \qquad +f_i(s) - \bar{f}(s) - d_i e_i, \qquad 1 \le i \le l, \\ \dot{d}_i = h_i e_i^T P_i(t) e_i, \\ \dot{e}_i = D\bar{f}(s)e_i + c\sum_{j\in\mathcal{N}_i} a_{ij}\Gamma e_j(t-\tau(t)) \\ \qquad + \int_0^1 (Df_i(s+\tau e_i) - D\bar{f}(s))e_i d\tau \\ \qquad -\frac{1}{N}\sum_{k=1}^N \int_0^1 Df_k(s+\tau e_k)e_k d\tau \\ \qquad +f_i(s) - \bar{f}(s), \qquad\qquad l+1 \le i \le N. \end{cases} \tag{26}$$

Repeating a similar procedure to the previous subsection, the controlled consensus problem of system (1) is equivalent to the stability problem of the following $N-1$ of $n$-dimensional systems.

$$\begin{cases} \dot{w}_i = D\bar{f}(s(t))w_i - d_i w_i + c\lambda_i\Gamma w_i(t-\tau(t)) \\ \qquad +(\Phi_i^T \otimes I_n)I(t)(\Phi \otimes I_n)w \\ \qquad +(\Phi_i^T \otimes I_n)F(t), \qquad\qquad 2 \le i \le l, \\ \dot{d}_i = h_i w_i^T P_i(t) w_i, \\ \dot{w}_i = D\bar{f}(s)w_i + c\lambda_i\Gamma w_i(t-\tau(t)) \\ \qquad +(\Phi_i^T \otimes I_n)I(t)(\Phi \otimes I_n)w \\ \qquad +(\Phi_i^T \otimes I_n)F(t), \qquad\qquad l+1 \le i \le N, \end{cases} \tag{27}$$

where $w_i$, $w$, $\Phi$, $\Phi_i$, $I(t)$ and $F(t)$ are the same as the previous subsection.

**Theorem 2** Suppose there exist positive definite matrices $P_i(t) \in \mathcal{PC}^1_{n\times n}$, $Q_i$ and constants $\bar{\zeta} > 0$, $\gamma \ge 0$, $a > 0$ and $b > 0$ such that

$$\begin{aligned} a\|x\|^2 \le x_i^T P_i(t)x_i + \int_{t-\tau(t)}^t x_i^T(\alpha)Q_i x_i(\alpha)d\alpha \\ + \frac{(d_i-d)^2}{h_i} \le b\|x\|^2, \forall t \in R^+,\ x \in R^n, i = 2, 3, \cdots, N, \end{aligned} \tag{28}$$

$$\begin{aligned} \dot{P}_i(t) + P_i(t)D\bar{f}(s) + (D\bar{f}(s))^T P_i(t) + Q_i - 2dP_i(t) \\ + c^2\lambda_i^2 P_i(t)\Gamma Q_i^{-1}\Gamma^T P_i(t) + \bar{\zeta} I \le 0,\ i = 1, 2, \cdots, N, \end{aligned} \tag{29}$$

(11) and $\bar{\zeta} > 2\gamma\beta$ are satisfied, then the system (6) converges to the set (14) for any time-varying delay $\tau(t) > 0$, where $\mu(t)$ and $\beta$ are the same as in (12) and (13) respectively, $\bar{\delta} > 0$ is any constant satisfying $\bar{\delta} < \bar{\zeta} - 2\gamma\beta$, and then the NMAS (1) achieves bounded consensus for any fixed time delay $\tau(t) > 0$, $0 \le \dot{\tau}(t) \le 1$.

**Proof.** Construct the following Lyapunov-Krasovskii functional as

$$V = \sum_{i=2}^N V_i + \sum_{i=2}^l \frac{(d_i-d)^2}{h_i}, \tag{30}$$

where

$$\begin{cases} V_i = w_i^T P_i(t) w_i + \int_{t-\tau(t)}^t w_i^T(\alpha)Q_i w_i(\alpha)d\alpha \\ \qquad + \frac{(d_i-d)^2}{h_i}, \qquad\qquad 2 \le i \le l, \\ V_i = w_i^T P_i(t) w_i + \int_{t-\tau(t)}^t w_i^T(\alpha)Q_i w_i(\alpha)d\alpha, \\ \qquad\qquad\qquad\qquad\qquad l+1 \le i \le N, \end{cases} \tag{31}$$

where $d$ is a positive constant to be determined.

Differentiating (31) along the trajectory of (27) gives

$$\begin{aligned} \dot{V}_i = {} & w_i^T(\dot{P}_i(t) + P_i(t)D\bar{f}(s) + (D\bar{f}(s))^T P_i(t) + Q_i \\ & - 2dP_i(t))w_i + 2w_i^T P_i(t)(\Phi_i^T \otimes I_n)I(t)(\Phi_i \otimes I_n)w \\ & + 2w_i^T P_i(t)(\Phi_i^T \otimes I_n)F(t) + 2w_i^T(c\lambda_i P_i(t)\Gamma)w_i(t-\tau(t)) \\ & - w_i^T(t-\tau(t))Q_i w_i(t-\tau(t)). \end{aligned} \tag{32}$$

The remaining part of the proof is similar to that of Theorem 1, so is therefore omitted here. This completes the proof.

## V. Examples

To demonstrate the theoretical results obtained above, we construct a NMAS consisting of 12 agents described as follows

$$\dot{x}_i(t) = f_i(x_i(t)) + c\sum_{j\in\mathcal{N}_i} a_{ij}\Gamma x_j(t-\tau(t)), \tag{33}$$

where $f_i(x_i(t)) = B_i x_i(t) + g(x_i(t))$, $B_i(i = 1, 2, \cdots, 6)$ and $B_i(i = 7, 8, \cdots, 12)$ are chosen as follows:

$$\begin{pmatrix} -10 + 0.1 \times (i-1) & 10 - 0.1 \times (i-1) & 0 \\ 1 & -1 & 1 \\ 0 & -15 - 0.1 \times (i-1) & 0 \end{pmatrix},$$

$$\begin{pmatrix} -10 - 0.1 \times (i-6) & 10 + 0.1 \times (i-6) & 0 \\ 1 & -1 & 1 \\ 0 & -15 + 0.1 \times (i-6) & 0 \end{pmatrix},$$

and

$$g(x_i(t)) = (-9.5sin(\frac{\pi x_{i1}(t)}{3.2} + \pi)\ 0\ 0)^T, \quad i = 1, 2, \cdots, 12.$$

The communication coupling matrix $C = (C_1^T C_2^T \cdots C_{12}^T)$, $C_1 = (-8\ 1\ 1\ 0\ 1\ 1\ 1\ 0\ 1\ 1\ 1\ 0)$, $C_2 = (1\ -8\ 1\ 1\ 1\ 0\ 1\ 0\ 1\ 1\ 1\ 0)$, $C_3 = (1\ 1\ -7\ 1\ 0\ 0\ 0\ 1\ 0\ 1\ 1\ 1)$, $C_4 = (0\ 1\ 1\ -6\ 0\ 1\ 1\ 0\ 0\ 1\ 0\ 1)$, $C_5 = (1\ 1\ 0\ 0\ -6\ 0\ 1\ 1\ 1\ 1\ 0\ 0)$, $C_6 = (1\ 0\ 0\ 1\ 0\ -5\ 1\ 0\ 1\ 1\ 0)$, $C_7 = (1\ 1\ 0\ 1\ 1\ 1\ -7\ 1\ 0\ 1\ 0)$, $C_8 = (0\ 0\ 1\ 0\ 1\ 0\ 1\ -6\ 0\ 1\ 1\ 1)$, $C_9 = (1\ 1\ 0\ 0\ 1\ 1\ 0\ 0\ -7\ 1\ 1\ 1)$, $C_{10} = (1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ -10\ 1)$, $C_{11} = (1\ 1\ 1\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ -7\ 1)$, $C_{12} = (0\ 0\ 1\ 1\ 0\ 0\ 0\ 1\ 1\ 0\ 1\ -5)$. $\Gamma = diag\{2, 2, 2\}$, respectively, where the matrix $A$ is produced by means of the Scale-Free network program.

Design the following controllers

$$\begin{cases} u_{i_k} = -d_{i_k}(x_{i_k}(t) - s(t)), & i_k = 1, 2 \texttt{ and } 10, \\ u_{i_k} = 0, & \texttt{else}, \end{cases}$$

with $d_1 = 0.5$, $d_2 = 0.5$, $d_{10} = 0.5$ and

$$\begin{cases} u_{i_k} = -d_{i_k}(t)(x_{i_k}(t) - s(t)), & i_k = 1, 2 \texttt{ and } 10, \\ \dot{d}_{i_k}(t) = h_{i_k} e_{i_k}^T P_{i_k}(t) e_{i_k}, \\ u_{i_k} = 0, & \texttt{else}, \end{cases}$$

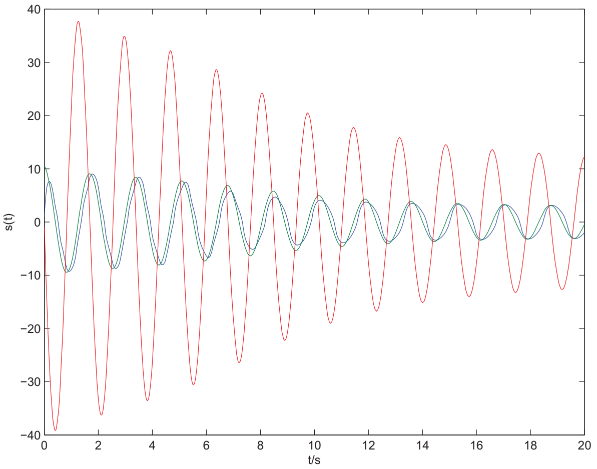


Fig.1. Desired agent dynamics under pinning control.

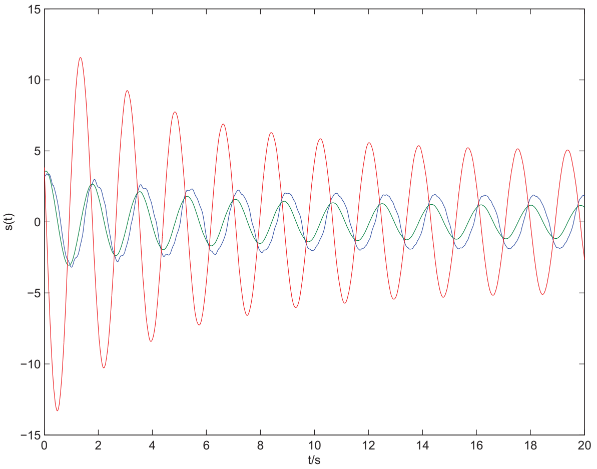


Fig.2. Desired agent dynamics under adaptive pinning control.

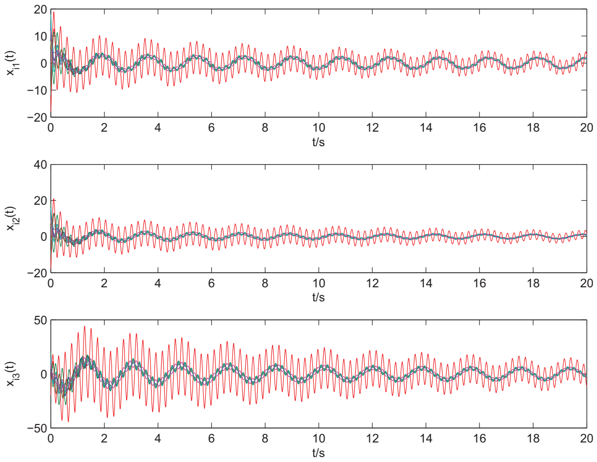


Fig.3. All agent dynamics under pinning control.

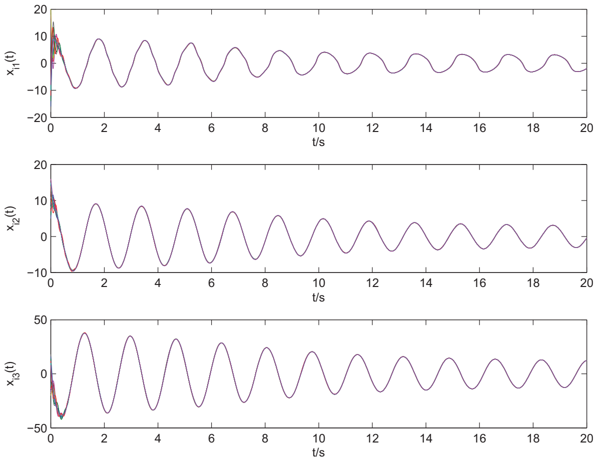


Fig.4. All agent dynamics under adaptive pinning control.

with $h_1 = 0.1$, $h_2 = 0.2$, $h_{10} = 0.3$, $s(t)$ can then be evaluated by simulation.

Given the initial values of 12 agents as $(10\ 5\ \ -10)^T$, $(12\ 6\ \ -12)^T$, $(14\ 7\ \ -14)^T$, $(16\ 8\ \ -16)^T$, $(18\ 9\ \ -18)^T$, $(20\ 10\ -20)^T$, $(-18\ 11\ 18)^T$, $(-16\ 12\ 16)^T$, $(-14\ 13\ 14)^T$, $(-12\ 14\ 12)^T$, $(-10\ 15\ 10)^T$, $(-8\ 16\ 8)^T$ respectively and $P_{i_k}(t) = I_3$, $d_1(0) = 1$, $d_2(0) = 1$, $d_{10}(0) = 1$ and $\tau(t) = \frac{\pi}{2} + arctan(t)$. The conditions of Theorem 1 and Theorem 2 are satisfied readily. Bounded consensus of the NMAS is achieved for any time varying delay satisfying $0 < \tau \leq \frac{\pi}{2} + arctan(t)$. Simulation results are depicted in Fig.1 to Fig.8 for $\tau(t) = \frac{\pi}{2} + arctan(t)$ and $c = 1$.

## VI. Conclusion

In this paper, we've investigated the controlled consensus problems of NMAS with different agent dynamics. The derived criteria are verified via theoretical analysis and numerical simulation. The consensus for the NMAS is achieved based on pinning control and adaptive pinning control methods. Many related results for the case of identical agent dynamics have been viewed as the special cases of the proposed results. However, it should be noted that the conditions are still restrictive and the time-varying delay is chosen as fixed case. Further investigations will focus on relaxing these limitations and more generalized cases.

## References

[1] Sandro Zampieri. Trends in networked control systems. *Proceedings of 17th World Congress of IFAC*, Seoul, Korea, July 6-11, 2008, pp. 2886 – 2894.

[2] J.P. Desai, J.P. Ostrowski and V. Kumar. Modeling and control of formation of nonholonomic mobile robust. *IEEE Int. Trans. on Robotics and Automation*, 2001, 17(6): 905 – 908.

[3] M. Porfiri, D.G. Roberson and D.J. Stilwell. Tracking and formation control of multiple autonomous agents: a two-level consensus approach. *Automatica*, 2007, 43: 1318 – 1328.

[4] J. Cortés. Global formation-shape stabilization of relative sensing networks. *2009 American Control Systems*, Hyatt Regency Riverfront, St. Louis, USA, June 10-12, 2009, pp. 1460 – 1465.

[5] F. Xiao and L. Wang. Asynchronous consensus in continuous-time multi-agent systems with switching topology and time-varying delays. *IEEE Transactions on Automatic Control*, 2008, 53(8): 1804 – 1816.

[6] T. Li and J.F. Zhang. Decentralized tracking-type games for multi-agent systems with coupled ARX models: asymptotic Nash equilibria. *Automatica*, 2008, 44: 713 – 725.

[7] E.S. Kazerooni and K. Khorasani. Optimal consensus algorithms for cooperative team of agents subject to partial information. *Automatica*, 2008, 44: 2766 – 2777.

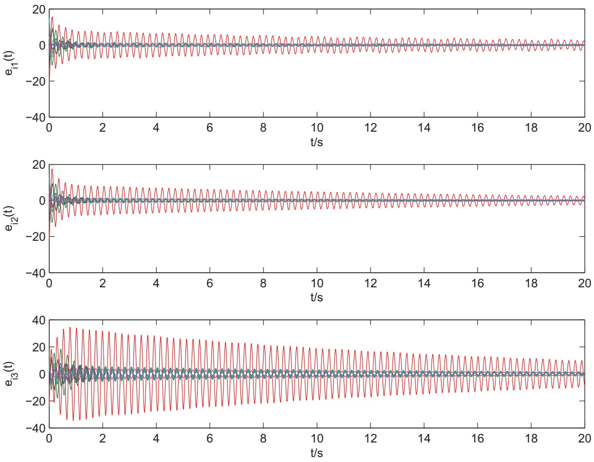


Fig.5. All agent error dynamics under pinning control.

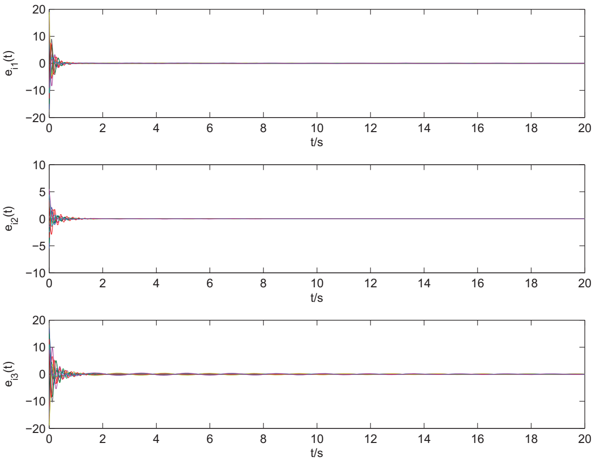


Fig.6. All agent error dynamics under adaptive pinning control.

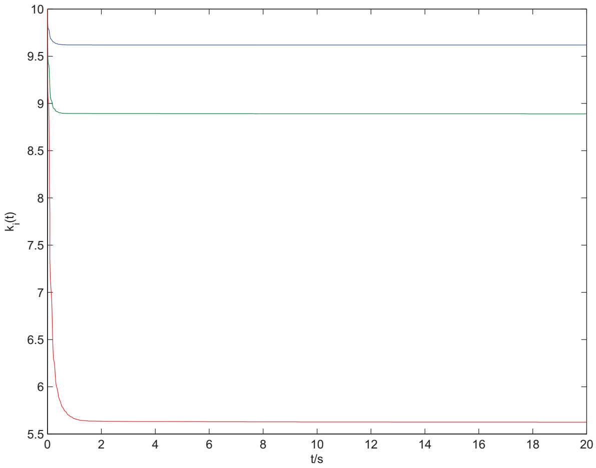


Fig.7. Adaptive gain curves.

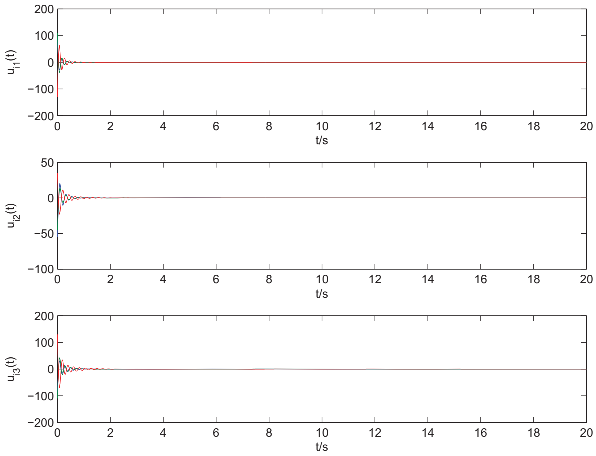


Fig.8. Adaptive pinning controllers curves.

[8] Z.K. Li, Z.S. Duan, G.R. Chen and L. Huang. Consensus of multiagent systems and synchronization of complex networks: a unified viewpoint. *IEEE Transactions on Circuits and Systems-I*, 2010, 57(1): 213 – 224.

[9] R. Olfati-Saber and R.M. Murray. Consensus problmes in networks with switching topology and time-delays. *IEEE Transactions on Automatic Control*, 2004, 49(9): 1520 – 1533.

[10] M. Arack. Passivity as a design tool for group coordination. *IEEE Transactions on Automatic Control*, 2007, 52(8): 1380 – 1390.

[11] F. Chen, Z.Q. Chen, L.Y. Xiang, Z.X. Liu and Z.Z. Yuan. Reaching a consensus via pinning control. *Automatica*, 2009, 45: 1215 – 1220.

[12] Y.P. Tian and C.L Liu. Robust consensus of multi-agent systems with diverse input delays and asymmetric interconnection perturbations. *Automatica*, 2009, 45: 1347 – 1353.

[13] P. Lin, Y.M. Jia and L. Li. Distributed robust $H_\infty$ Consensus control in directed networks of agents with time-delay. *Systems and Control Letters*, 2008, 57(8): 643 – 653.

[14] W. Ren, R.W. Beard. Formation feedback control for multiple spacecraft via virtual structures. *IEE Proc. Control Theory and Applications*, 2004, 151(3): 357 – 368.

[15] H. Yamaguchi, T. Arai and G. Beni. A distributed control scheme for multiple robotic vehicles to make group formations. *Robitics and Autonomous Systems*, 2001, 36: 125-147.

[16] Z.P. Wu, Z.H. Guan, T. Li. Group motion control of multi-agent systems based on complex network, *Proceedings of the 26th Chinese Control Conference*, Hunan, China, July 26-31, 2007, pp: 63 – 67.

[17] J.Q. Lu, D.W.C. Ho and J. Kurths. Consensus over directed static networks with arbitrary finite communication delays. *Physical Review E*, 2009, 80(6): 066121.

[18] J. Cortes. Distributed algorithms for reaching consensus on general functions. *Automatica*, 2008, 44(3): 726 – 737.

[19] J. Zhao, D. J. Hill and T. Liu. Synchronization of Dynamical Networks With Nonidentical Nodes: Criteria and Control. *IEEE Tran.Circuits and Syst.I*, 2011, 58(3): 584 – 594.

[20] P.A. Bliman and G.F. Trecate. Average consensus problems in networks of agents with delayed communications. *Automatica*, 2008, 44: 1985 – 1995.

[21] Z.S Duan and G.R. Chen. Global Robust Stability and Synchronization of Networks With Lorenz-Type Nodes. *IEEE Transactions on Circuits and Systems-II: Express Briefs*, 2009, 56(8): 679 – 683.

[22] D. J.Hill and J. Zhao. Global synchronization of complex dynamical networks with non-identical nodes. *Proceedings of the 47th IEEE Conference on Decision and Control*, Cancun, Mexico, Dec. 9-11, 2008, 817 – 822.

[23] J.C. Zhao, J.A. Lu and Q.J. Zhang. Pinning a complex dynamical network to a homogenous trajectory. *IEEE Transactions on Circuits and Systems-II: Express Briefs*, 2009, 56(6): 514 – 518.

[24] X. Li, X.F. Wang and G.R. Chen. Pinning a complex dynamical network to its equilibrium. *IEEE Tran.Circuits and Syst.I*,2004, 51: 2074 – 2087.

[25] W. Wu, W.J. Zhou and T.P. Chen. Cluster Synchronization of Linearly Coupled Complex Networks Under Pinning Control. *IEEE Transactions on Circuits and Systems-I*, 2009, 56(4): 829 – 839.

[26] Y.Y. Wu, W. Wei, G.Y. Li and J. Xiang. Pinning Control of Uncertain Complex Networks to a Homogeneous Orbit. *IEEE Transactions on Circuits and Systems-II: Express Briefs*, 2009, 56(3): 235 – 239.

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# Global Bounded Controlled Consensus of Multi-Agents Systems with Non-Identical Nodes and Communication Time-Delay Topology

Wei-Song Zhong
Faculty of Advanced Technology
University of Glamorgan
Cardiff, UK CF37 1DL
Email: wzhong@glam.ac.uk

Guo-Ping Liu
Faculty of Advanced Technology
University of Glamorgan
Cardiff, UK CF37 1DL
Email: gpliu@glam.ac.uk

***Abstract*—This paper investigates the global bounded consensus problem of Networked Multi-Agent Systems exhibiting nonlinear, non-identical agent dynamics with communication time-varying delay. Globally bounded controlled consensus conditions based on pinning control method and adaptive pinning control method are derived. The proposed consensus criteria ensures that all agents eventually move along desired trajectories in terms of boundedness. The proposed controlled consensus criteria generalizes the case of identical agent dynamics to the case of non-identical agent dynamics, and many related results of other researches in this area can be viewed as special cases of the above results. We finally demonstrate the effectiveness of the theoretical results by means of a numerical simulation.**

## I. Introduction

Networked Multi-Agent Systems (NMAS) analysis involves the study of how the network architectures and interactions between network components influence global control goals and some important contributions have been made in recent years [1], [2], [3], [4].

The consensus problem has been studied across many fields of science and engineering [5], [6], [7], [8], [9], [10], [11], [12], [13]. The controlled consensus problem of NMAS with non-identical agent dynamics is much more complicated than the identical case and few results have been reported to date [14].

The present paper will focus on the global consensus problems of NMAS based on pinning control methods [15], [16], [17], and the proposed controlled consensus property is formulated in terms of certain boundedness of state errors. In this paper, we'll generalize many existing results for the case of identical agent dynamics to the case of non-identical agent dynamics based on the pinning control method.

The rest of this paper is organized as follows. A controlled continuous-time NMAS model with communication time-delay is presented in Section II. The main results including pinning control and adaptive pinning control bounded consensus criterion are derived in Section III and V respectively. Section IV gives a numerical simulation example to verify the effectiveness of the proposed results, followed by conclusions in Section VI.

## II. Problem description

Let $G = (\mathcal{V}, \mathcal{A})$ be a graph of order $N$ consisting of a set of vertices $\mathcal{V} = \{v_1, v_2, \cdots, v_N\}$ and a set of edges $\mathcal{A} \subseteq \mathcal{V} \times \mathcal{V}$. An edge $(v_j, v_i)$ in graph $G$ means that agent $v_i$ sends some information to agent $v_j$. The set of neighbors of agent $v_i$ is denoted by $\mathcal{N}_i = \{v_j \in \mathcal{V} : (v_j, v_i) \in \mathcal{A}\}$.

We consider a MAS consisting of $N$ non-identical agents with communication delay:

$$\dot{x}_i = f_i(x_i) + c\sum_{j\in\mathcal{N}_i}^{N} a_{ij}\Gamma x_j(t-\tau), i = 1, 2, \cdots, N, \qquad (1)$$

where $x_i = (x_{i1}(t), x_{i2}(t), \cdots, x_{in}(t))^T \in R^n$ are the state variables of the agent $v_i$, $f_i(x_i) : R^n \to R^n$ are continuously differentiable mappings with Jacobian $Df_i$, representing the self-dynamics of the agent $v_i$, $c > 0$ denotes the coupling strength, $\Gamma = (\gamma_{ij}) \in R^{n\times n}$ is the inner coupling matrix, and where $\gamma_{ij} \neq 0$ means two connected agents are linked via their $i$th and $j$th state variables, respectively. The adjacency matrix $A = (a_{ij}) \in R^{N\times N}$ (which is symmetric and irreducible) represents the communication topology relation of the MAS, and is defined by $a_{ij} = a_{ji} = 1(v_j \in \mathcal{N}_i)$, $a_{ij} = 0(v_j \notin \mathcal{N}_i)$ and $a_{ii} = -\sum_{j\neq i} a_{ij}$. $\tau$ is a constant coupling delay which reflects the reality that the agent $v_i$ can't obtain information from agent $v_j$ instantaneously.

The average dynamic of all agents is defined by the vector field $\bar{f}(x(t)) = \frac{1}{N}\sum_{k=1}^{N} f_k(x(t))$ with Jacobian $D\bar{f}_i(x(t))$.

The average state trajectory is chosen as the desired moving trajectory

$$s(t) = \frac{1}{N}\sum_{k=1}^{N} x_k(t). \qquad (2)$$

We now discuss the problem of global consensus for the system (1). The consensus problem formulation in the present paper is quite different from many others, where the consensus problem is solvable if the states of all agents satisfy $x_i(t) \to x_j(t)$, $\forall i, j = 1, 2, \cdots, N$ as $t \to \infty$. The consensus problem here will be depicted instead via certain boundedness

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

of $x_i(t) - x_j(t)$, $\forall i, j = 1, 2, \cdots, N$ as $t \to \infty$. This better reflects reality as it is impossible for MAS (1) to achieve exact consensus. To address this case we will focus on making the states of all agents converge to a bounded set.

We denote $x(t)$, $s(t)$, $u(t)$, $e(t)$, $w(t)$ and $V(w(t), t)$ as $x$, $s$, $u$, $e$, $w$ and $V$ respectively.

## III. Linear feedback pinning controller

To achieve the goal, we apply the feedback control strategy on a small fraction $\delta$ $(0 < \delta \leq 1)$ of the agents in system (1). Suppose that nodes $i_1, i_2, \cdots, i_l$ are selected to be under control, where $l = [\delta N]$ stands for the smaller but nearest integer to the real number $\delta N$. This controlled MAS can be described as

$$\begin{cases} \dot{x}_{i_k} = f_{i_k}(x_{i_k}) + c\sum_{j=1}^{N} a_{i_k j}\Gamma x_j(t-\tau) + u_{i_k}, 1 \leq k \leq l, \\ \dot{x}_{i_k} = f_{i_k}(x_{i_k}) + c\sum_{j=1}^{N} a_{i_k j}\Gamma x_j(t-\tau), l+1 \leq k \leq N. \end{cases} \tag{3}$$

The local linear negative feedback control law is chosen as follows:

$$\begin{cases} u_{i_k} = -d_{i_k}(x_{i_k} - s), & 1 \leq k \leq l, \\ u_{i_k} = 0, & l+1 \leq k \leq N, \end{cases} \tag{4}$$

where the feedback gain $d_{i_k} > 0$.

Combine (3) and (4) and rearrange the order of the n-odes in the network. Let the first $l$ nodes be controlled, and $e_i = x_i - s$, $i = 1, 2, \cdots, N$. It's obvious that $\frac{c}{N}\sum_{k=1}^{N}\sum_{j=1}^{N} a_{kj}\Gamma x_j(t-\tau) = 0$ and $\sum_{i=1}^{N} e_i = 0$. Then by applying the Newton-Leibniz formula, error systems can be written as

$$\begin{cases} \dot{e}_i = D\bar{f}(s)e_i + c\sum_{j=1}^{N} a_{ij}\Gamma e_j(t-\tau) \\ \quad + \int_0^1 (Df_i(s+\tau e_i) - D\bar{f}(s))e_i d\tau \\ \quad - \frac{1}{N}\sum_{k=1}^{N}\int_0^1 Df_k(s+\tau e_k)e_k d\tau \\ \quad + f_i(s) - \bar{f}(s) - d_i e_i, & 1 \leq i \leq l, \\ \dot{e}_i = D\bar{f}(s)e_i + c\sum_{j=1}^{N} a_{ij}\Gamma e_j(t-\tau) \\ \quad + \int_0^1 (Df_i + \tau e_i) - D\bar{f}(s))e_i d\tau \\ \quad - \frac{1}{N}\sum_{k=1}^{N}\int_0^1 Df_k(s+\tau e_k)e_k d\tau \\ \quad + f_i(s) - \bar{f}(s), & l+1 \leq i \leq N. \end{cases} \tag{5}$$

The following work will focus on simplifying the error systems (5) by means of a series of transformations using a procedure similar to [14].

Define the following matrix

$$D = diag(D_1, D_2, \cdots, D_N) \in R^{nN \times nN},$$

where $D_i = diag\{-d_i, -d_i, \cdots, -d_i\} \in R^{n \times n}$.

Let $e = (e_1^T, e_2^T, \cdots, e_N^T)^T$, then (5) becomes

$$\dot{e} = \bar{\Sigma}(t)e + cA \otimes \Gamma e(t-\tau) + I(t)e - \frac{1}{N}H(t)e + F(t), \tag{6}$$

where $I(t) = diag\{\int_0^1 (Df_1(s+\tau e_1) - D\bar{f}(s))d\tau \cdots \int_0^1 (Df_N(s+\tau e_N) - D\bar{f}(s))d\tau\}$, $\bar{\Sigma}(t) = I_N \otimes D\bar{f}(s) + D$, $H^T(t) = (H_1^T(t), \cdots, H_N^T(t))$, $H_i(t) = (\int_0^1 Df_1(s+\tau e_1)d\tau, \cdots, \int_0^1 Df_N(s+\tau e_N)d\tau)$, $F_i^T(t) = (f_1^T(s) - \bar{f}^T(s), \cdots, f_N^T(s) - \bar{f}^T(s))$.

Since $A$ is symmetric and irreducible, according to [14], there exists a unitary matrix $\Phi = (\varphi_{ij})_{N\times N} = (\Phi_1, \Phi_2, \cdots, \Phi_N)$. This together with $w(t) = (\Phi^T \otimes I_n)e$ gives

$$\begin{aligned} \dot{w} = {} & (\Phi^T \otimes I_n)\bar{\Sigma}(t)(\Phi \otimes I_n)w \\ & + (\Phi^T \otimes I_n)(cA \otimes \Gamma)(\Phi \otimes I_n)w(t-\tau) \\ & + (\Phi^T \otimes I_n)I(t)(\Phi \otimes I_n)w \\ & - \frac{1}{N}(\Phi^T \otimes I_n)H(t)(\Phi \otimes I_n)w + (\Phi^T \otimes I_n)F(t). \end{aligned} \tag{7}$$

Note that $H(t) = \sqrt{N}\sum_{k=1}^{N}(\mathbf{0}\ \cdots\ \mathbf{0}\ \bar{\Phi}_k\ \mathbf{0}\ \cdots\ \mathbf{0}) \otimes \int_0^1 Df_k(s+\tau e_k)d\tau$, where $\bar{\Phi}_k$ stands for the matrix with its $k$-th column equal to $\Phi_1$ and the remaining elements are zero. Then we have $\frac{1}{N}(\Phi^T \otimes I_n)H(t)(\Phi \otimes I_n) = \frac{1}{\sqrt{N}}\sum_{k=1}^{N}(\mathbf{0}\ \cdots\ \mathbf{0}\ I_k\ \mathbf{0}\ \cdots\ \mathbf{0}) \otimes \int_0^1 Df_k(s+\tau e_k)d\tau(\Phi \otimes I_n)$, where $I_k$ stands for the matrix with its $k$-th column equals $(1\ 0\ \cdots\ 0)^T$ and the remaining of its elements are zero.

Thus, a simple calculation gives $\frac{1}{N}(\Phi^T \otimes I_n)H(t)(\Phi \otimes I_n) = \frac{1}{\sqrt{N}}\sum_{k=1}^{N}\begin{pmatrix} \Upsilon_k \\ 0 \end{pmatrix} \otimes \int_0^1 Df_k(s(t)+\tau e_k(t))d\tau$, where $\Upsilon_k \in R^{1\times N}$ and $0 \in R^{(N-1)\times N}$. Therefore, $\dot{w} = \bar{\Sigma}(t)w + c\Lambda \otimes \Gamma w(t-\tau) + (\Phi^T \otimes I_n)I(t)(\Phi \otimes I_n)w - \begin{pmatrix} * \\ 0 \end{pmatrix} w + (\Phi^T \otimes I_n)F(t)$. Since $w_1 \equiv 0$, we only need to consider $w_2, w_3, \cdots, w_N$. Rewriting in the component form we have

$$\begin{aligned} \dot{w}_i = {} & \Sigma_i(t)w_i + c\lambda_i\Gamma w_i(t-\tau) + (\Phi_i^T \otimes I_n)F(t) \\ & + (\Phi_i^T \otimes I_n)I(t)(\Phi \otimes I_n)w,\ i = 2, 3, \cdots, N, \end{aligned} \tag{8}$$

where $\Sigma_i = \bar{D}f(s) + D_i$.

So far, we have transferred the consensus problem of system (1) to the stability problem of the $N-1$ of $n-$dimensional systems.

**Theorem 1** Suppose that $\|I(t)\| \leq \gamma$ is satisfied. If there exist matrices $P_i(t) \in \mathcal{PC}_{n\times n}^1$, $Q_i > 0$, $\Theta_i > 0$, $\Pi_i > 0$, $X_i$, $Y_i$ and $Z_i$ of appropriate dimensions such that

$$B = \begin{pmatrix} B_1 & B_2 \\ B_2^T & B_3 \end{pmatrix} < 0, \quad \begin{pmatrix} X_i & Y_i \\ Y_i^T & Z_i \end{pmatrix} \geq 0, \tag{9}$$

for $i = 2, 3, \cdots, N$, where $B_1 = \dot{P}_i(t) + P_i(t)\Sigma_i(t) + \Sigma_i^T(t)P_i(t) + hX_i + Y_i^T + Y_i + Q_i + h\Sigma_i^T(t)Z_i\Sigma_i(t)$, $B_2 = c\lambda_i P_i(t)\Gamma - Y_i + hc\lambda_i\Sigma_i^T(t)Z_i\Gamma$ and $B_3 = \Pi_i^{-1} + \Theta_i^{-1} - Q_i + hc^2\lambda_i^2\Gamma^T Z_i\Gamma$, then the MAS (1) will achieve bounded consensus for the time-invariant delay $\tau \in [0, h]$ for some $h < \infty$.

**Proof.** Construct the following Lyapunov-Krasovskii functional as

$$V = \sum_{i=2}^{N}\sum_{k=1}^{3} V_k, \tag{10}$$

where

$$\begin{aligned} V_1 &= w_i^T P_i(t) w_i, \\ V_2 &= \int_{-\tau}^{0} \int_{t+\beta}^{t} \dot{w}_i^T(\alpha) Z_i \dot{w}_i(\alpha) d\alpha d\beta, \\ V_3 &= \int_{t-\tau}^{t} w_i^T(\alpha) Q_i w_i(\alpha) d\alpha. \end{aligned}$$

The $i$-th ($i = 2, 3, \cdots, N$) equation in system (8) can be written as

$$\begin{aligned} \dot{w}_i = (\Sigma_i(t) + c\lambda_i \Gamma) w_i - c\lambda_i \Gamma \int_{t-\tau}^{t} \dot{w}_i(\alpha) d\alpha \\ + (\Phi_i^T \otimes I_n) I(t) (\Phi \otimes I_n) w + (\Phi_i^T \otimes I_n) F(t). \quad (11) \end{aligned}$$

Defining $a(.)$, $b(.)$ and $M$ in [18] as $a(\alpha) = w_i(t)$, $b(\alpha) = \dot{w}_i(\alpha)$ and $M = c\lambda_i P_i(t)\Gamma$ for all $\alpha \in [t-\tau, t]$ then we have

$$\begin{aligned} \dot{V}_1 \le w_i^T [\dot{P}_i(t) + P_i(t)\Sigma_i(t) + \Sigma_i^T(t) P_i(t) \\ + hX_i + Y_i^T + Y_i] w_i + \int_{t-\tau}^{t} \dot{w}_i^T(\alpha) Z_i \dot{w}_i(\alpha) d\alpha \\ + 2w_i^T (c\lambda_i P_i(t)\Gamma - Y_i) w_i(t-\tau) \\ + 2w_i^T P_i(t) (\Phi_i^T \otimes I_n) I(t) (\Phi_i \otimes I_n) w \\ + 2w_i^T P_i(t) (\Phi_i^T \otimes I_n) F(t). \quad (12) \end{aligned}$$

Moreover, $\dot{V}_2$ can be enlarged as

$$\begin{aligned} \dot{V}_2 \le h[\Sigma_i(t) w_i + c\lambda_i \Gamma w_i(t-\tau)]^T Z_i [\Sigma_i(t) w_i \\ + c\lambda_i \Gamma w_i(t-\tau)] + 2h(\Sigma_i(t) w_i)^T Z_i (\Phi_i^T \otimes I_n) I(t) \\ (\Phi \otimes I_n) w + 2h(\Sigma_i(t) w_i)^T Z_i (\Phi_i^T \otimes I_n) F(t) \\ + 2h(c\lambda_i \Gamma w_i(t-\tau))^T Z_i (\Phi_i^T \otimes I_n) I(t) (\Phi \otimes I_n) w \\ + 2h(c\lambda_i \Gamma w_i(t-\tau))^T Z_i (\Phi_i^T \otimes I_n) F(t) \\ + 2h((\Phi_i^T \otimes I_n) I(t) (\Phi \otimes I_n) w)^T Z_i (\Phi_i^T \otimes I_n) F(t) \\ + h((\Phi_i^T \otimes I_n) F(t))^T Z_i ((\Phi_i^T \otimes I_n) F(t)) \\ + h((\Phi_i^T \otimes I_n) I(t) (\Phi \otimes I_n) w)^T Z_i ((\Phi_i^T \otimes I_n) I(t) \\ (\Phi \otimes I_n) w) - \int_{t-\tau}^{t} \dot{w}_i^T(\alpha) Z_i \dot{w}_i(\alpha) d\alpha. \quad (13) \end{aligned}$$

and

$$\dot{V}_3 = w_i^T Q_i w_i - w_i^T(t-\tau) Q_i w_i(t-\tau). \quad (14)$$

Applying the Young Inequality, then we have $2h(c\lambda_i \Gamma w_i(t - \tau))^T Z_i (\Phi_i^T \otimes I_n) I(t) (\Phi \otimes I_n) w \le w_i^T(t-\tau) \Pi_i^{-1} w_i(t-\tau) + h^2 c^2 \lambda_i^2 w^T ((\Phi \otimes I_n)^T I(t) (\Phi_i^T \otimes I_n)^T Z_i \Gamma \Pi_i \Gamma^T Z_i (\Phi_i^T \otimes I_n) I(t) (\Phi \otimes I_n)) w(t)$, and $2h(c\lambda_i \Gamma w_i(t-\tau))^T Z_i (\Phi_i^T \otimes I_n) F(t) \le w_i^T(t-\tau) \Theta_i^{-1} w_i(t-\tau) + h^2 c^2 \lambda_i^2 F^T(t) (\Phi_i^T \otimes I_n)^T Z_i \Gamma \Theta_i \Gamma^T Z_i (\Phi_i^T \otimes I_n) F(t)$. Applying these two inequalities and the conditions of the theorem results

$$\begin{aligned} \dot{V} \le \sum_{i=2}^{N} \begin{pmatrix} w_i \\ w_i(t-\tau) \end{pmatrix}^T B \begin{pmatrix} w_i \\ w_i(t-\tau) \end{pmatrix} \\ + 2\mu(t)\beta + (\|w\|(2\gamma\beta + 2h\gamma\|\Sigma_i(t)\| \sum_{i=2}^{N} \lambda_{max}(Z_i) \\ + 2h\mu(t)\|\Sigma_i(t)\| \sum_{i=2}^{N} \lambda_{max}(Z_i) + h\gamma^2 \sum_{i=2}^{N} \lambda_{max}(Z_i) \\ + h^2 c^2 \gamma^2 \lambda_{max}^{\frac{1}{2}}(\Gamma\Gamma^T) \sum_{i=2}^{N} \lambda_{max}(\Pi_i) \lambda_i^2 \lambda_{max}^2(Z_i) \\ + h^2 c^2 \mu^2(t) \lambda_{max}^{\frac{1}{2}}(\Gamma\Gamma^T) \sum_{i=2}^{N} \lambda_{max}(\Theta_i) \lambda_i^2 \lambda_{max}^2(Z_i)) \|w\| \\ + 2h\gamma \sum_{i=2}^{N} \lambda_{max}(Z_i) \mu(t)) + h\mu^2(t) \sum_{i=2}^{N} \lambda_i^2 \lambda_{max}(Z_i). \quad (15) \end{aligned}$$

Thus when

$$\|w\| \ge \frac{2\mu(t)\beta + 2h\gamma \sum_{i=2}^{N} \lambda_{max}(Z_i)\mu(t)}{\varpi(t)},$$

we have

$$\dot{V} \le -\delta\|w\|^2 + h\mu^2(t) \sum_{i=2}^{N} \lambda_{max}(Z_i) \lambda_i^2, \quad (16)$$

where $\varpi(t) = -(2\gamma\beta + 2h\gamma\|\Sigma_i(t)\| \sum_{i=2}^{N} \lambda_{max}(Z_i) + 2h\mu(t)\|\Sigma_i(t)\| \sum_{i=2}^{N} \lambda_{max}(Z_i) + h\gamma^2 \sum_{i=2}^{N} \lambda_{max}(Z_i) + h^2 c^2 \gamma^2 \lambda_{max}^{\frac{1}{2}}(\Gamma\Gamma^T) \sum_{i=2}^{N} \lambda_{max}(\Pi_i) \lambda_i^2 \lambda_{max}^2(Z_i) + h^2 c^2 \mu^2(t) \lambda_{max}^{\frac{1}{2}}(\Gamma\Gamma^T) \sum_{i=2}^{N} \lambda_{max}(\Theta_i) \lambda_i^2 \lambda_{max}^2(Z_i)) - \delta$. Thus, according to [19] and Lyapunov stability theory, bounded consensus is ultimately achieved. This completes the proof.

## IV. Adaptive pinning controller

In this section, we will derive globally consensus criteria via direct adaptive pinning control method. Without loss of generality, we still assume that the first $l$ agents are selected as pinned agents with the adaptive controllers:

$$\begin{cases} u_i = -d_i(t)(x_i - s), & 1 \le i \le l, \\ \dot{d}_i(t) = h_i e_i^T P_i(t) e_i, & \\ u_i = 0, & l+1 \le i \le N, \end{cases} \quad (17)$$

where constant $h_i > 0$ and positive definite matrix $P_i(t) \in R^{n \times n}$. Applying Newton-Leibniz formula, then the error MAS can be rewritten as

$$\begin{cases} \dot{e}_i = D\bar{f}(s)e_i + c\sum_{j=1}^{N} a_{ij}\Gamma e_j(t-\tau) \\ \qquad + \int_0^1 (Df_i(s+\tau e_i) - D\bar{f}(s))e_i d\tau \\ \qquad -\frac{1}{N}\sum_{k=1}^{N}\int_0^1 Df_k(s+\tau e_k)e_k d\tau \\ \qquad +f_i(s) - \bar{f}(s) - d_i(t)e_i, \qquad 1 \le i \le l, \\ \dot{d}_i(t) = h_i e_i^T P_i(t) e_i, \\ \dot{e}_i = D\bar{f}(s)e_i + c\sum_{j=1}^{N} a_{ij}\Gamma e_j(t-\tau) \\ \qquad + \int_0^1 (Df_i(s+\tau e_i) - D\bar{f}(s))e_i d\tau \\ \qquad -\frac{1}{N}\sum_{k=1}^{N}\int_0^1 Df_k(s+\tau e_k)e_k d\tau \\ \qquad +f_i(s) - \bar{f}(s), \qquad l+1 \le i \le N. \end{cases} \tag{18}$$

Repeating a similar procedure to the previous subsection, the controlled consensus problem of system (1) is equivalent to the stability problem of the following $N-1$ of $n$-dimensional systems.

$$\begin{cases} \dot{w}_i = D\bar{f}(s(t))w_i - d_i(t)w_i + c\lambda_i\Gamma w_i(t-\tau) \\ \qquad +(\Phi_i^T \otimes I_n)I(t)(\Phi \otimes I_n)w \\ \qquad +(\Phi_i^T \otimes I_n)F(t), \qquad 2 \le i \le l, \\ \dot{d}_i(t) = h_i w_i^T P_i(t) w_i, \\ \dot{w}_i = D\bar{f}(s)w_i + c\lambda_i\Gamma w_i(t-\tau) \\ \qquad +(\Phi_i^T \otimes I_n)I(t)(\Phi \otimes I_n)w \\ \qquad +(\Phi_i^T \otimes I_n)F(t), \qquad l+1 \le i \le N, \end{cases} \tag{19}$$

where $w_i$, $w$, $\Phi$, $\Phi_i$, $I(t)$ and $F(t)$ are the same as the previous subsection.

**Theorem 2** Suppose that $\|I(t)\| \le \gamma$ is satisfied. If there exist matrices $P_i(t) \in \mathcal{PC}^1_{n\times n}$, $Q_i > 0$, $\Theta_i > 0$, $\Pi_i > 0$, $X_i$, $Y_i$ and $Z_i$ of appropriate dimensions and constant $d > 0$ such that

$$B = \begin{pmatrix} B_1 & B_2 \\ B_2^T & B_3 \end{pmatrix} < 0, \quad \begin{pmatrix} X_i & Y_i \\ Y_i^T & Z_i \end{pmatrix} \ge 0, \tag{20}$$

for $i = 2, 3, \cdots, N$, where $B_1 = \dot{P}_i(t) + P_i(t)(Df(s)) + (Df(s))^T P_i(t) - 2dP_i(t) + hX_i + Y_i^T + Y_i + Q_i + h\Sigma_i^T(t)Z_i\Sigma_i(t)$, $B_2 = c\lambda_i P_i(t)\Gamma - Y_i + hc\lambda_i\Sigma_i^T(t)Z_i\Gamma$ and $B_3 = \Pi_i^{-1} + \Theta_i^{-1} - Q_i + hc^2\lambda_i^2\Gamma^T Z_i\Gamma$, then the system (1) will achieve bounded consensus for the time-invariant delay $\tau \in [0, h]$ for some $h < \infty$.

**Proof.** Construct the following Lyapunov-Krasovskii functional as

$$V = \sum_{i=2}^{N}\sum_{k=1}^{3} V_k + \sum_{i=2}^{l}\frac{(d_i(t) - d)^2}{h_i}, \tag{21}$$

where

$$V_1 = w_i^T P_i(t) w_i,$$

$$V_2 = \int_{-\tau}^{0}\int_{t+\beta}^{t} \dot{w}_i^T(\alpha) Z_i \dot{w}_i(\alpha) d\alpha d\beta,$$

$$V_3 = \int_{t-\tau}^{t} w_i^T(\alpha) Q_i w_i(\alpha) d\alpha.$$

The remaining part of the proof is similar to that of Theorem 1 and is therefore omitted here. This completes the proof.

## V. EXAMPLE

To demonstrate the theoretical results obtained above, we construct a MAS consisting of 11 agents described as follows

$$\dot{x}_i(t) = f_i(x_i(t)) + c\sum_{j\in\mathcal{N}_i}^{N} a_{ij}\Gamma x_j(t-\tau), \tag{22}$$

where $f_i(x_i(t)) = B_i x_i(t) + g(x_i(t))$, $B_i (i = 1, 2, \cdots, 6)$ and $B_i (i = 7, 8, \cdots, 11)$ are chosen as follows:

$$\begin{pmatrix} -10 + 0.1 \times (i-1) & 10 - 0.1 \times (i-1) & 0 \\ 1 & -1 & 1 \\ 0 & -15 - 0.1 \times (i-1) & 0 \end{pmatrix},$$

$$\begin{pmatrix} -10 - 0.1 \times (i-6) & 10 + 0.1 \times (i-6) & 0 \\ 1 & -1 & 1 \\ 0 & -15 + 0.1 \times (i-6) & 0 \end{pmatrix},$$

and

$$g(x_i(t)) = (-9.5sin(\frac{\pi x_{i1}(t)}{3.2} + \pi)\ 0\ 0)^T, \quad i = 1, 2, \cdots, 11.$$

The communication coupling matrix $C = (C_1^T C_2^T \cdots C_{11}^T)$, $C_1 = (-8\ 1\ 1\ 0\ 1\ 1\ 1\ 0\ 1\ 1\ 1)$, $C_2 = (1\ -8\ 1\ 1\ 1\ 0\ 1\ 0\ 1\ 1\ 1)$, $C_3 = (1\ 1\ -6\ 1\ 0\ 0\ 0\ 1\ 0\ 1\ 1)$, $C_4 = (0\ 1\ 1\ -5\ 0\ 1\ 1\ 0\ 0\ 1\ 0)$, $C_5 = (1\ 1\ 0\ 0\ -6\ 0\ 1\ 1\ 1\ 1\ 0)$, $C_6 = (1\ 0\ 0\ 1\ 0\ -5\ 1\ 0\ 1\ 1\ 0)$, $C_7 = (1\ 1\ 0\ 1\ 1\ 1\ -7\ 1\ 0\ 1\ 0)$, $C_8 = (0\ 0\ 1\ 0\ 1\ 0\ 1\ -5\ 0\ 1\ 1)$, $C_9 = (1\ 1\ 0\ 0\ 1\ 1\ 0\ 0\ -6\ 1\ 1)$, $C_{10} = (1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ -10\ 1)$, $C_{11} = (1\ 1\ 1\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ -6)$. $\Gamma = diag\{2, 2, 2\}$, respectively, where the matrix $A$ is produced by means of the Scale-Free network program.

Design the following controllers

$$\begin{cases} u_{i_k} = -d_{i_k}(x_{i_k}(t) - s(t)), & i_k = 1, 2 \texttt{ and } 10, \\ u_{i_k} = 0, & \texttt{else}, \end{cases}$$

with $d_1 = 0.5$, $d_2 = 0.5$, $d_{10} = 0.5$ and

$$\begin{cases} u_{i_k} = -d_{i_k}(t)(x_{i_k}(t) - s(t)), & i_k = 1, 2 \texttt{ and } 10, \\ \dot{d}_{i_k}(t) = h_{i_k} e_{i_k}^T P_{i_k}(t) e_{i_k}, \\ u_{i_k} = 0, & \texttt{else}, \end{cases}$$

with $h_1 = 0.1$, $h_2 = 0.2$, $h_{10} = 0.3$, $s(t)$ can then be evaluated by simulation.

Given the initial values of 11 agents as $(10\ 5\ -10)^T$, $(12\ 6\ -12)^T$, $(14\ 7\ -14)^T$, $(16\ 8\ -16)^T$, $(18\ 9\ -18)^T$, $(20\ 10\ -20)^T$, $(-18\ 11\ 18)^T$, $(-16\ 12\ 16)^T$, $(-14\ 13\ 14)^T$, $(-12\ 14\ 12)^T$, $(-10\ 15\ 10)^T$ respectively and $P_{i_k}(t) = I_3$. We may verify the conditions of Theorem 1 and Theorem 2 readily. This demonstrates the bounded consensus of the MAS is achieved for any time delay $0 < \tau \le 0.061$. Simulation results are depicted in Fig.1 to Fig.4 for $\tau = 0.061$ and $c = 1$.

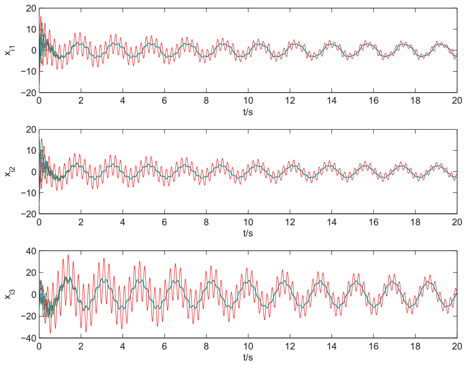


Fig.1. All agent dynamics under pinning control.

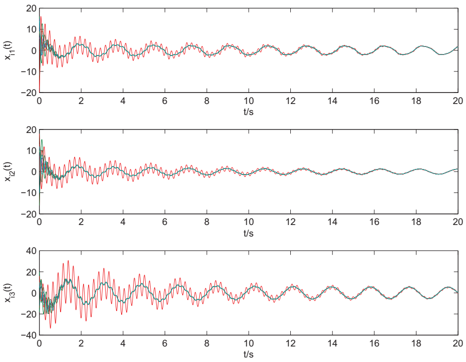


Fig.2. All agent dynamics under adaptive pinning control.

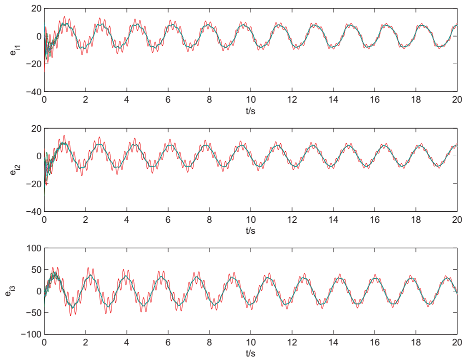


Fig.3. All agent error dynamics under pinning control.

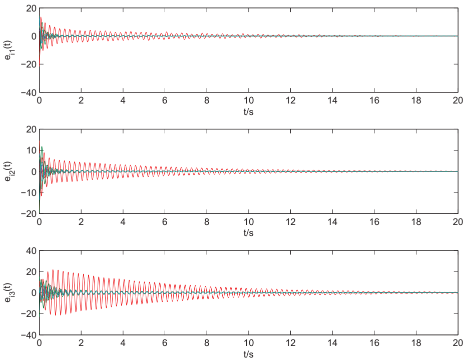


Fig.4. All agent error dynamics under adaptive pinning control.

## VI. Conclusion

In this paper, we've investigated the controlled consensus problems of NMAS with different agent dynamics. The derived criteria are verified via theoretical analysis and numerical simulation. The consensus for the NMAS is achieved based on pinning control and adaptive pinning control methods. It should be noted that the conditions are still restrictive and all the delays are the same. Further investigations will focus on relaxing these limitations.

## References

[1] Sandro Zampieri. Trends in networked control systems, *Proceedings of 17th World Congress of IFAC*, Seoul, Korea, July 6-11, 2008, pp. 2886 – 2894.

[2] J.P. Desai, J.P. Ostrowski and V. Kumar. Modeling and control of formation of nonholonomic mobile robust. *IEEE Int. Trans. on Robotics and Automation*, 2001, 17(6): 905 – 908.

[3] M. Porfiri, D.G. Roberson and D.J. Stilwell. Tracking and formation control of multiple autonomous agents: a two-level consensus approach. *Automatica*, 2007, 43: 1318 – 1328.

[4] J. Cortés. Global formation-shape stabilization of relative sensing networks. *2009 American Control Systems*, Hyatt Regency Riverfront, St. Louis, USA, June 10-12, 2009, pp. 1460 – 1465.

[5] Y.G. Hong, J.P. Hu and L.X. Gao. Tracking control for multi-agent consensus with an active leader and variable topology. *Automatica*, 2006, 42: 1177 – 1182.

[6] F. Xiao and L. Wang. Asynchronous consensus in continuous-time multi-agent systems with switching topology and time-varying delays. *IEEE Transactions on Automatic Control*, 2008, 53(8): 1804 – 1816.

[7] E.S. Kazerooni and K. Khorasani. Optimal consensus algorithms for cooperative team of agents subject to partial information. *Automatica*, 2008, 44: 2766 – 2777.

[8] R. Olfati-Saber and R.M. Murray. Consensus problmes in networks with switching topology and time-delays. *IEEE Transactions on Automatic Control*, 2004, 49(9): 1520 – 1533.

[9] M. Arack. Passivity as a design tool for group coordination. *IEEE Transactions on Automatic Control*, 2007, 52(8): 1380 – 1390.

[10] F. Chen, Z.Q. Chen, L.Y. Xiang, Z.X. Liu and Z.Z. Yuan. Reaching a consensus via pinning control. *Automatica*, 2009, 45: 1215 – 1220.

[11] Y.P. Tian and C.L Liu. Robust consensus of multi-agent systems with diverse input delays and asymmetric interconnection perturbations. *Automatica*, 2009, 45: 1347 – 1353.

[12] P. Lin, Y.M. Jia and L. Li. Distributed robust $H_\infty$ Consensus control in directed networks of agents with time-delay. *Systems and Control Letters*, 2008, 57(8): 643 – 653.

[13] P.A. Bliman and G.F. Trecate. Average consensus problems in networks of agents with delayed communications. *Automatica*, 2008, 44: 1985 – 1995.

[14] D. J.Hill and J. Zhao. Global synchronization of complex dynamical networks with non-identical nodes, *Proceedings of the 47th IEEE Conference on Decision and Control*, Cancun, Mexico, Dec. 9-11, 2008, 817 – 822.

[15] J.C. Zhao, J.A. Lu and Q.J. Zhang. Pinning a complex dynamical network to a homogenous trajectory. *IEEE Transactions on Circuits and Systems-II: Express Briefs*, 2009, 56(6): 514 – 518.

[16] W. Wu, W.J. Zhou and T.P. Chen. Cluster Synchronization of Linearly Coupled Complex Networks Under Pinning Control. *IEEE Transactions on Circuits and Systems-I*, 2009, 56(4): 829 – 839.

[17] Y.Y. Wu, W. Wei, G.Y. Li and J. Xiang. Pinning Control of Uncertain Complex Networks to a Homogeneous Orbit. *IEEE Transactions on Circuits and Systems-II: Express Briefs*, 2009, 56(3): 235 – 239.

[18] Y.M. Moon, P. Park, W.H. Kwon, Y.S. Lee. Delay-dependent robust stabilization of uncertain state-delayed systems. *Int J. Control*, 2001, 74: 1447 – 1455.

[19] C.C. Hua, X.P. Guan and P. Shi. Decentralized robust model reference adaptive cotnrol for interconnected time-delay systems. *J.Comput.Appl.Math.*, 2006, 193: 383 – 396.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Real Time Estimation of the Wheel-Rail Contact Condtions using Multi-Kalman Filtering and Fuzzy Logic

I Hussain
Mehran University, Jamshoro, Pakistan
imtiaz.hussain@faculty.muet.edu.pk

T. X. Mei
School of Computing, Science and Engineering, University of Salford
T.X.Mei@salford.ac.uk

Mohammad Mirzapour
School of Computing, Science and Engineering, University of Salford
M.Mirzapour@edu.salford.ac.uk

*Abstract*— **This paper presents a novel technique for the real time estimation of the contact conditions by using a combination of multi-Kalman filtering and fuzzy logic approach. The proposed solution exploits the variations in the dynamic behaviour of a railway wheelset with the changes in wheel-rail contact condition. The proposed system involves the use of multiple model based estimation of the wheelset dynamics in response to different track conditions. Each of the estimators is tuned to match one particular track condition to give the best results at the specific design point. Residuals of each filter are calculated and the level of matches/mismatches is reflected in the residual values of the models concerned. The residuals from all the models are then be assessed by a fuzzy inference system to determine the present operating condition and hence to give real time information about the track conditions.**

***Keywords- Wheel rail contact; State Estimation; Kalman filters; Fuzzy Logic***

## I. Introduction

Adhesion is a very important factor in the operation of the railway vehicles. The delivery of traction and braking is achieved through the available adhesion at the wheel-rail interface. Insufficient level of adhesion can lead to severe safety and operational problems resulting in huge financial losses to railway industry around the world. Although in last few decades the railway industry is able to manage low adhesion to some extent but currently available measures are not sufficient to eliminate the safety incidents and train delays. This is because the adhesion is affected by a large number of parameters such as weather, season changes and contaminations and therefore cannot be predicted with certainty. Changes in the adhesion conditions can be rapid and also short-lived, and the adhesion coefficient can differ from position to position along a route depending upon the type and degree of contamination which presents a great scientific challenge to effectively design a suitable technique to tackle this problem.

Current wheel slip/slide protection (WSP) technologies for traction and braking systems are incorporated in the rail vehicles to maximize the use of available adhesion [1]. WSPs control the slip ratio (relative speed between a wheel and the train) below a pre-defined threshold to avoid slip/slide during traction or braking [1]. In general, WSPs are effectively reactive systems, i.e. only 'activated' to stop wheel slip/slide when detected by the sensors. There is still a need for a system which is proactive and can prevent slip/slide from its occurrence, such that real time information about the track condition can be provided to the traction and braking control systems to maximize the use of available adhesion. On the other hand, the wheel-rail contact mechanics is extremely complex and vary with time which presents a great scientific challenge to effectively design a suitable technique to tackle this problem.

A number of ideas have been proposed that is related to the monitoring of the running condition of the wheel-rail interface that use low cost inertial sensing mounted on the vehicle and advanced processing, e.g. an inverse modelling approach for the estimation of creep forces [2, 23-25]; and a model based estimation [3-4].

A multiple model approach has been proposed previously by the authors for the real time estimation of the wheel-rail contact conditions [5-8], this paper extends the study to take into account contact conditions that are directly used in design of the Kalman filters. This is of particular practice importance, as real track conditions can be affected by uncertain external factors and hence unpredictable. Furthermore, this paper also covers the complete design of fuzzy inference system and presents a formula to convert the fuzzy logic output into percentage adhesion.

## II. Modelling of Contact Mechanics

Wheelsets are a key component of railway vehicles that interacts directly with the track and consequently the dynamics of the wheelset are directly influenced by changing contact conditions - therefore this study focusses on a single solid axle wheelset and the outcome of the study may be readily extended to the full vehicles [9-19]. The dynamic behaviour of the railway wheelset is governed by the creep forces generated at the wheel rail contact patches. These creep forces are the result of creepages which are the relative speed of the wheels to rail and can be characterized as lateral ($\lambda_y$) and longitudinal creep ($\lambda_x$) in accordance with the direction of motion as given in equations 1-3 [5-8].

$$\lambda_{xL} = \frac{r_o \omega_L - v}{v} + [\frac{L_g \dot{\psi}}{v} + \frac{\gamma(y - y_t)}{r_o}] \tag{1}$$

$$\lambda_{xR} = \frac{r_o \omega_R - v}{v} - [\frac{L_g \dot{\psi}}{v} + \frac{\gamma(y - y_t)}{r_o}] \tag{2}$$

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

$$\lambda_y = \lambda_{yR} = \lambda_{yL} = \frac{\dot{y}}{v} - \psi \qquad (3)$$

where the subscripts $L$ and $R$ represent left and right wheels, $r_o$ is the nominal radius of the wheels, $v$ is the vehicle forward speed, $\gamma$ is the conicity of the wheels, $\Psi$ is the yaw angle, $L_g$ is the track half gauge, $\omega_L$ and $\omega_R$ are the angular velocities of the left and right wheels respectively, $y$ is the lateral motion, and $y_t$ represents the track irregularity in lateral direction. The total creepage $\lambda_j$ is the combination of the lateral and longitudinal creepages.

$$\lambda_j = \sqrt{\lambda_{ij}^2 + \lambda_{ij}^2} \qquad i = x, y \; and \; j = L, R \qquad (4)$$

The total creep force $F_j$ is a nonlinear function of the total creepage and can be represented using equation 5.

$$F_j = \mu_j N_j \qquad j = L, R \qquad (5)$$

The distribution of the contact forces in the longitudinal and lateral directions is thoroughly studied by Polach [20] and can be represented using (6).

$$F_{ij} = F_j . \frac{\lambda_{ij}}{\lambda_j}, \qquad i = x, y \; and \; j = L, R \qquad (6)$$

The equations of motion of the wheelset are given in equations (7-12).

$$\begin{aligned} M_v \ddot{x} = {} & \frac{\mu_R N_R}{\sqrt{\lambda_{xR}^2 + \lambda_{yL}^2}} \left[ \frac{r_o \omega_R - v}{v} - [\frac{L_g \dot{\psi}}{v} + \frac{\gamma(y - y_t)}{r_o}] \right] \\ & + \frac{\mu_L N_L}{\sqrt{\lambda_{xL}^2 + \lambda_{yL}^2}} \left[ \frac{r_o \omega_L - v}{v} + [\frac{L_g \dot{\psi}}{v} + \frac{\gamma(y - y_t)}{r_o}] \right] \end{aligned} \qquad (7)$$

$$\begin{aligned} I_w \ddot{\psi} = {} & \frac{\mu_R N_R}{\sqrt{\lambda_{xR}^2 + \lambda_{yL}^2}} \left[ \frac{r_o \omega_R - v}{v} - [\frac{L_g \dot{\psi}}{v} + \frac{\gamma(y - y_t)}{r_o}] \right] L_g \\ & - \frac{\mu_L N_L}{\sqrt{\lambda_{xL}^2 + \lambda_{yL}^2}} \left[ \frac{r_o \omega_L - v}{v} + [\frac{L_g \dot{\psi}}{v} + \frac{\gamma(y - y_t)}{r_o}] \right] L_g \text{ - } k_w \psi \end{aligned} \qquad (8)$$

$$m_w \ddot{y} = - \frac{\mu_L N_L}{\sqrt{\lambda_{xL}^2 + \lambda_{yL}^2}} \left[ \frac{\dot{y}}{v} - \psi \right] - \frac{\mu_R N_R}{\sqrt{\lambda_{xR}^2 + \lambda_{yL}^2}} \left[ \frac{\dot{y}}{v} - \psi \right] + F_c + F_g \qquad (9)$$

$$\begin{aligned} I_R \dot{\omega}_R = {} & T_t - K_s \theta_s - r_o \frac{\mu_R N_R}{\sqrt{\lambda_{xR}^2 + \lambda_{yL}^2}} [\frac{r_o \omega_R - v}{v} - (\frac{L_g \dot{\psi}}{v} + \\ & \frac{\gamma(y - y_t)}{r_o}) \end{aligned} \qquad (10)$$

$$\begin{aligned} I_L \dot{\omega}_L = {} & K_s \theta_s - r_o \frac{\mu_L N_L}{\sqrt{\lambda_{xL}^2 + \lambda_{yL}^2}} [\frac{r_o \omega_L - v}{v} + (\frac{L_g \dot{\psi}}{v} \\ & + \frac{\gamma(y - y_t)}{r_o})] \end{aligned} \qquad (11)$$

$$T_s = k_s \int (\omega_R - \omega_L) dt + C_s (\omega_R - \omega_L) \qquad (12)$$

where $M_v$ is the mass of the vehicle $\ddot{x}$ is the vehicle forward acceleration, $I_w$ is the yaw moment of inertia, $k_w$ is a yaw stiffness necessary to stabilise the wheelset, $m_w$ is the mass of the wheelset, $\ddot{y}$ is the lateral acceleration, $F_c$ is a centrifugal force which is taken into consideration when the wheelset runs on a curved track, $F_g$ is the gravitational stiffness force related to the lateral displacement and roll angle of the wheelset. The tractive torque $T_t$ is applied to one side of the wheelset (right side in this case) and the other wheel is driven by the torsional torque $T_s$. $\theta_s = \int(\omega_R - \omega_L)dt$. $k_s$ is the torsional stiffness of the shaft connecting the two wheels and $C_s$ is material damping of the shaft, which is usually very small.

## III. Design of Multiple Kalman Filters

The main objective of this study is to detect the changes in the wheel-rail contact condition with practical sensors. The design of the estimator is simplified by considering the wheelset modes that are directly related to contact conditions. Previous studies have suggested that the lateral and yaw dynamics are sufficient for the study of plan-view dynamics of a wheelset [4, 5, 6, 8, 15, 21-22]. The use of a simplified model has several advantages in the estimator design without having a significant effect on the results [5-8]. The major advantage is the simple design of the estimator with minimum number of states which will allow the estimator to converge quickly. The yaw and lateral dynamics are excited by lateral track irregularities. The contact forces given in (5) and (6) are nonlinear in nature and are linearized at specific points on the creep curves in order to enable the design of the Kalman filters. The small signal model of linearized creep forces is given in following equation [5-7].

$$\begin{bmatrix} \Delta\dot{y} \\ \Delta\dot{\psi} \\ \Delta\ddot{y} \\ \Delta\ddot{\psi} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & \frac{2g_{22}}{m_w} & -\frac{2g_{22}}{vm_w} & 0 \\ -\frac{2L_g\gamma g_{11}}{r_o I_w} & -\frac{k_w}{I_w} & 0 & -\frac{2L_g^2 g_{11}}{vI_w} \end{bmatrix} \begin{bmatrix} \Delta y \\ \Delta\psi \\ \Delta\dot{y} \\ \Delta\dot{\psi} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ \frac{2L_g\gamma g_{11}}{r_o I_w} \end{bmatrix} \Delta y_{\mathrm{t}} \qquad (13)$$

The track disturbances ($y_t$) are very difficult/expensive to measure in practice and therefore highly undesirable to be used as an input to the Kalman filters, therefore (13) is reformulated to include the track input as an additional state, as

$$\frac{d}{dt} \begin{bmatrix} \Delta\psi \\ \Delta\dot{y} \\ \Delta\dot{\psi} \\ \Delta y_t \\ \Delta y - \Delta y_t \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ \frac{2g_{22}}{m_w} & -\frac{2g_{22}}{vm_w} & 0 & 0 & 0 \\ -\frac{k_w}{I_w} & 0 & -\frac{2L_g^2 g_{11}}{vI_w} & 0 & -\frac{2L_g\gamma g_{11}}{r_o I_w} \\ 0 & 0 & 0 & N & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix} \qquad (14)$$

$$\begin{bmatrix} \Delta\psi \\ \Delta\dot{y} \\ \Delta\dot{\psi} \\ \Delta y_t \\ \Delta y - \Delta y_t \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ -1 \end{bmatrix} \dot{y}_{\mathrm{t}}$$

A gyro sensor to measure yaw rate and accelerometer to measure lateral acceleration are found to be sufficient to produce satisfactory results. The output equation is given in (15).

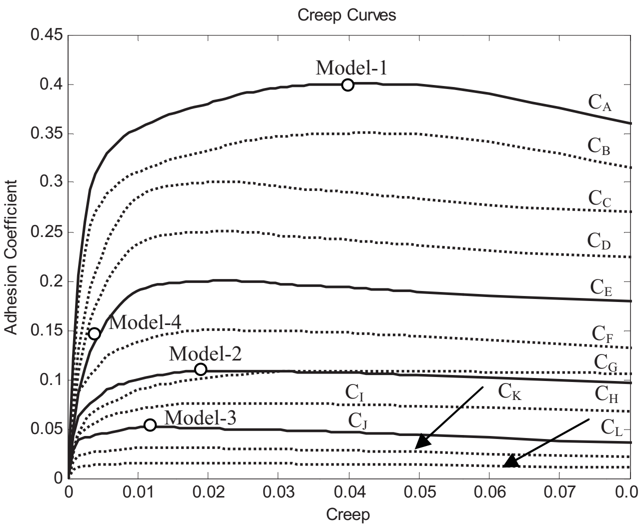


Figure 1. Creep Curves.

$$z(t)=\begin{bmatrix}0 & 0 & 1 & 0 & 0\\ \frac{2g_{22}}{m_w} & -\frac{2g_{22}}{vm_w} & 0 & 0 & 0\end{bmatrix}\begin{bmatrix}\Delta\psi\\ \Delta\dot{y}\\ \Delta\dot{\psi}\\ \Delta y_t\\ \Delta y-\Delta y_t\end{bmatrix}+v \qquad (15)$$

## IV. Estimation oF Contact Conditions

Several creep curves representing high to very low adhesion conditions are used for the design and the evaluation of the detection system. The creep curves used for the Kalman filter design are shown in solid lines and the creep curves shown in dotted lines are used as additional cases for the assessment of detection system. The proposed scheme is shown in Fig.2. This scheme identifies the contact condition based on the residuals of the Kalman filters, which are the difference generated by the observations and the system's mathematical model. The design of each Kalman filter is based on linerised creep coefficients $g_{11}$ and $g_{22}$ values. At the saturation point of creep curves, $g_{11}$, which depends upon the slope of the creep curve $d\mu/d\lambda$, is zero and $g_{22}$, which depends upon traction ratio $\mu/\lambda$, is different for different adhesion levels. Therefore the residual signal of a Kalman filter designed and tuned to operate in specific contact condition is expected to be at lowest when the vehicle is operated in similar contact condition, in comparison with those from other Kalman filters. In this study, four Kalman filters are found to be sufficient to detect the changes in the contact conditions. Model-1 is tuned to operate at the saturation region of the creep curve $C_A$, model-2 is designed to operate on the saturation region of the creep curve $C_C$, model-3 is designed to operate on the saturation region of creep curve $C_D$ and model-4 is designed to operate in the linear region of creep curve $C_B$. The purpose of the model-4 is to identify whether the wheelset is operating at the saturation region of the creep curve or not.

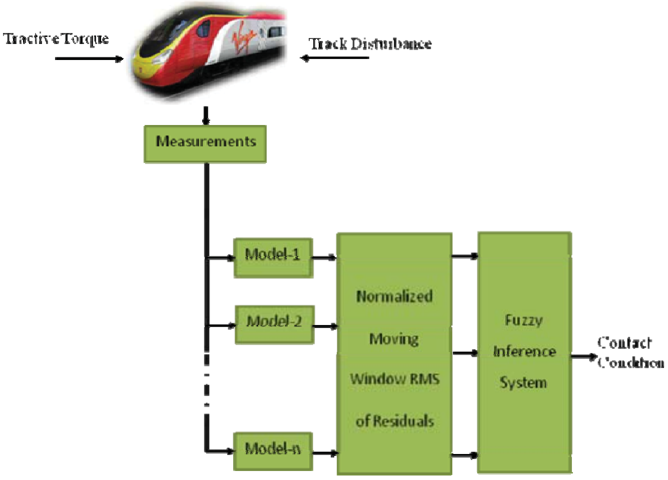


Figure 2. Contact Condition Estimation Scheme

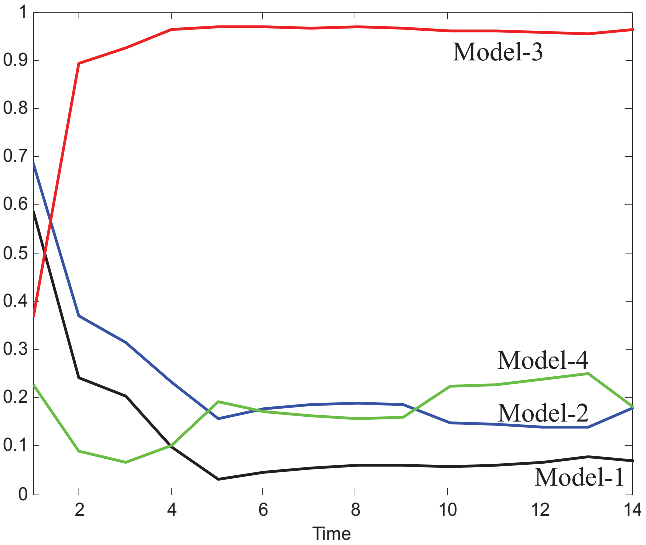


Figure 3. Residuals of filters at $P_1$

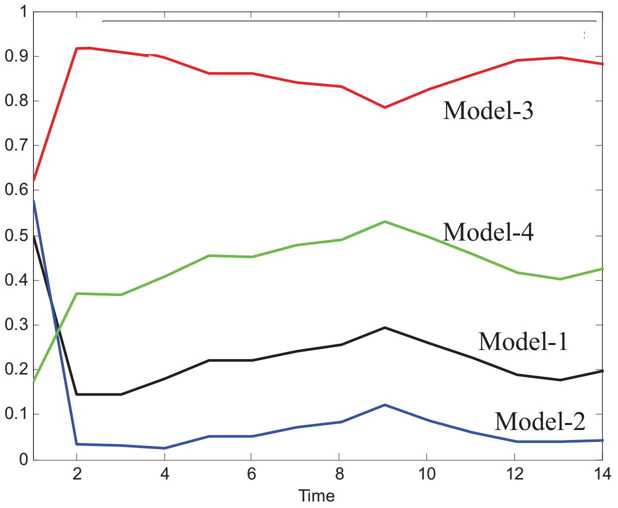


Figure 4. Residuals of filters at $P_3$

Fig.3 shows the residuals of the filters when the wheelset is operated at the saturation region of the creep curve-1. The residual of model-1 is the lowest as expected confirming that the wheelset is operating at $P_a$. When the wheelset is operated at $P_c$ the residual of model-2 is at the lowest and the residual of model-1 is increased as shown in Fig.4. If the wheelset is not operated on the saturation points of the creep curves the residual of model-4 designed at $P_e$ is at the lowest. For instance when the wheelset is operated in the linear region of the creep curve-1 the residual of model-4 is at the lowest as shown in Fig.5.

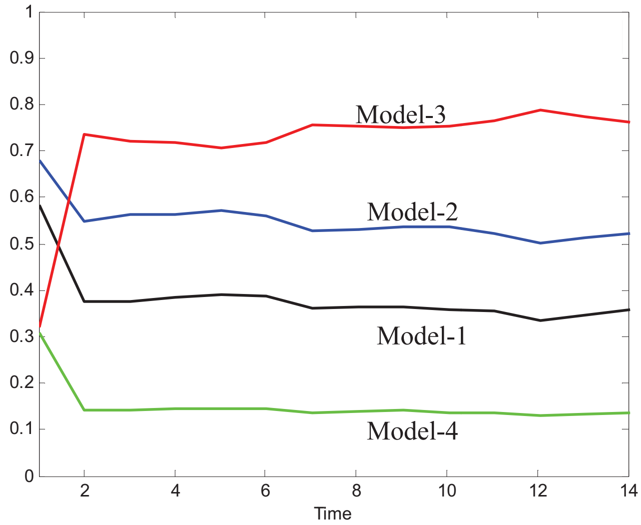


Figure 5. Residuals the when the wheelset is operated in linear region of $C_A$

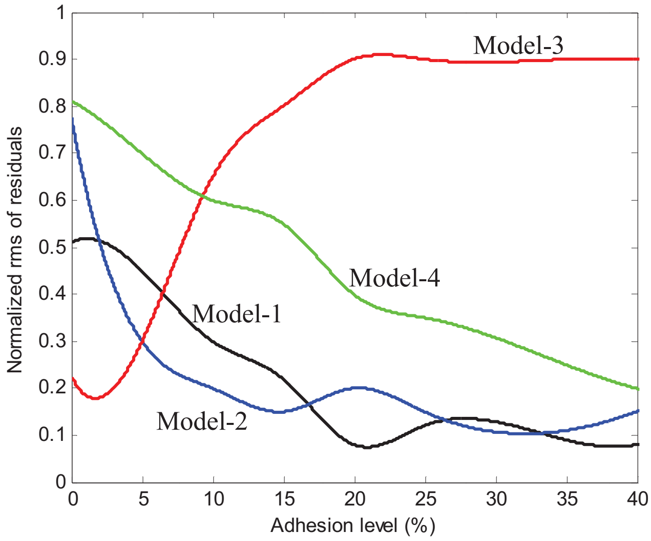


Figure 6. Residuals of filters after interpolation

As only four models are used in the proposed detection scheme, it is necessary to evaluate how the models would respond to other contact conditions that are not directly included in the design. This is carried out by simulating the filters in different contact conditions shown in figure-1 and the residual data is interpolated (Fig.6) to develop a fuzzy logic based detection system to detect other possible contact conditions.

## V. Design Of Fuzzy Inference System

The basic idea of the fuzzy inference system is that if wheelset operating point is at the saturation region of the creep curve then the residual information together with the tractive torque can easily be used to determine the adhesion level. The fuzzy inference system (FIS) that analyzes the residuals and the tractive torque is shown in Fig.7. As any other fuzzy logic system it has three main divisions. First part is the input division which scales and weighs the inputs and determines the magnitude of participation in producing output. The inputs are then processed according to set rules and the final output is determined by the averaging the output of individual rules.

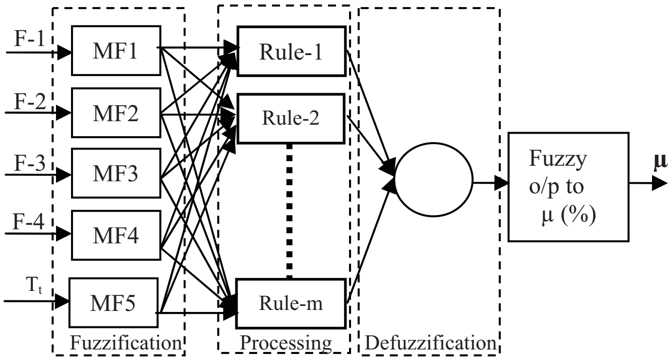


Figure 7. Fuzzy Inference System

The residual of each model in Fig.6 is divided as 'Low', 'Moderate' and 'High' to develop input membership function. Input membership function for residual of model and modle-2 are shown in Fig.8 and Fig.9 respectively.

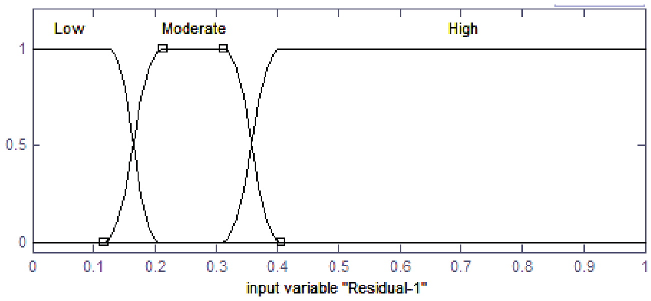


Figure 8. Input membership function of Residual-1

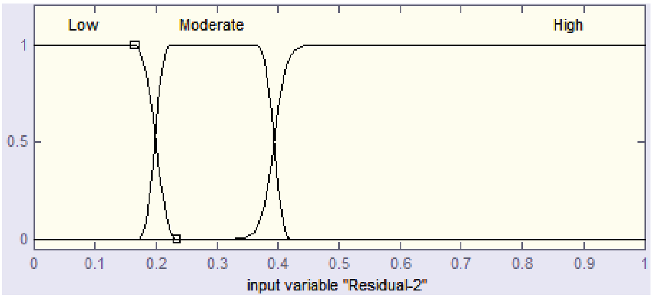


Figure 9. Input membership function of Residual-2

After the residual values are scaled and weighted they are processed according to set rules. The fuzzy logic rules are developed by examining the residual values in different contact conditions e.g. in 40% adhesion level the value of residual of model-1 is 'Low', the value of residual of Model-2 is 'Low', the value of residual of model-3 is 'High' and the value of residual of model-4 is 'Moderate'. Similarly rules for other possible contact conditions are developed and some of the rules are given below.

If *Residual-1* is '*Low*' and *Residual-2* is '*Low*' and *Residual-3* is '*High*' and *Residual-4* is '*Moderate*' and $T_t$ is '$T_8$' then '$\mu \geq 40\%$'.
If *Residual-1* is '*Low*' and *Residual-2* is '*Low*' and *Residual-3* is '*Moderate*' and *Residual-4* is '*Low*' and $T_t$ is '$T_3$' then '$20\% \geq \mu \geq 10\%$'.

The output is determined by averaging the outcome of all the rules and final numeric output ranging from 0 to 100 is produced. The output fuzzy set is shown in Fig-10.

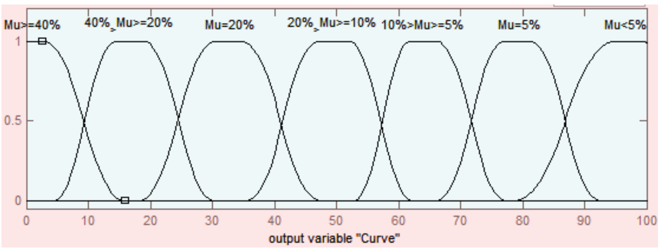


Figure 10. Input membership function of Residual-2

The final stage is to convert the fuzzy logic output into percentage adhesion, which is done by fitting a sixth order polynomial on fuzzy output data set and following equation is obtained.

$$\mu = -2.213\times10^{-6} n_1^{\;4} + 0.00042 n_1^{\;3} - 0.0205 n_1^{\;2} - 0.38 n_1 + 42 \quad (16)$$

where $n_1$ is fuzzy logic output and $\mu$ is percentage adhesion level. The simulations are run different contact conditions and the results are given below.

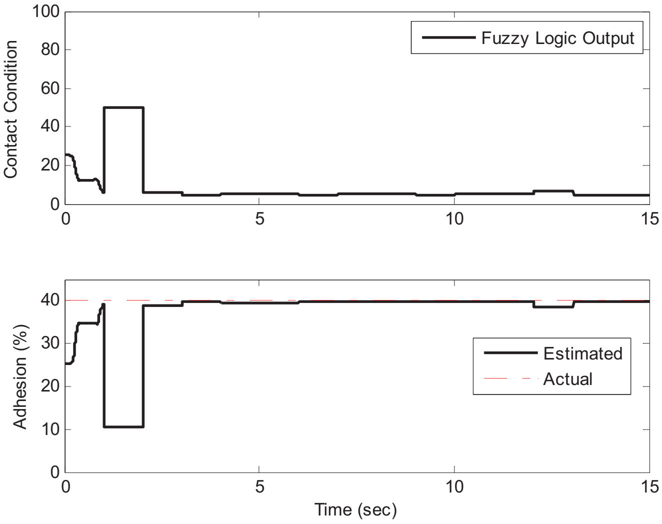


Figure 11. Simulation carried out using creep Curve $C_A$

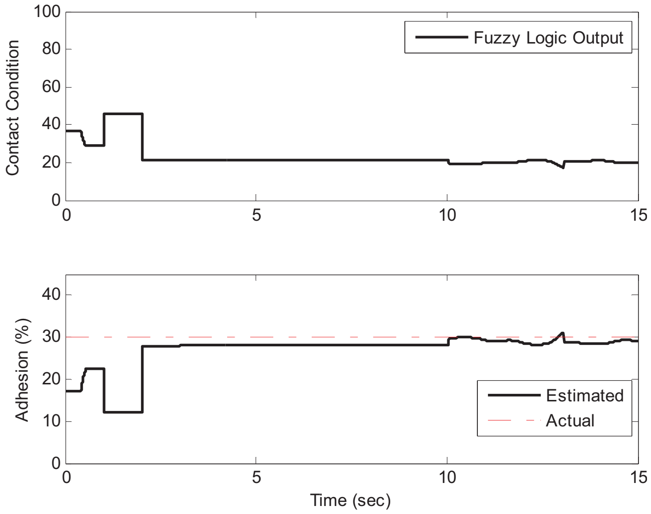


Figure 12. Simulation carried out using creep Curve $C_C$

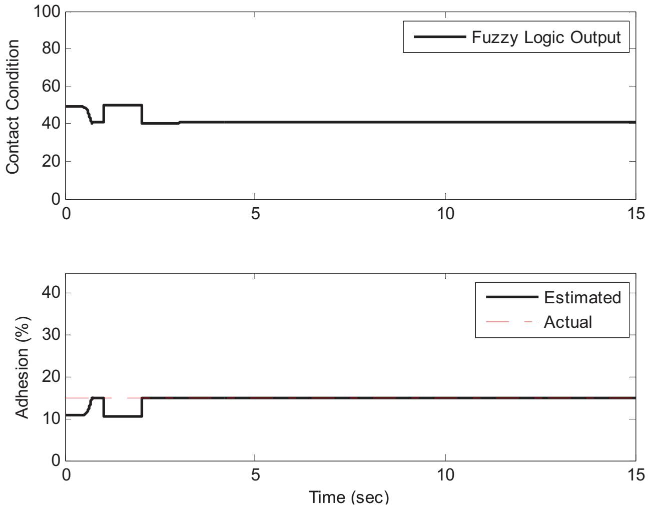


Figure 13. Simulation carried out using creep Curve $C_F$

Fig.11 shows the output of the fuzzy inference system when the system is operated on creep curve $C_A$. the fuzzy logic output is converted to adhesion information using (16). The system takes approximately two seconds to react and produce correct output. The delay in this case is due to the time needed in computer simulation to reach steadily contact conditions and also due to the time required to calculate the windowed rms of residuals. After 2 seconds the output fuzzy logic system is steady and the estimated adhesion level is almost equal to the simulated adhesion level. Fig.12 shows the result obtained by simulating the creep curve $C_C$. Again the steady output is produced after a delay of approximately 2 seconds. In practice while the wheelset already be in motion the amount of delay would not be as much. After the expected delay, the estimated adhesion level is approximately equal to the actual adhesion level. The difference in the actual output and the estimated output is caused by several reasons that include inaccuracy in fuzzy interpretation and the error due to curve fitting formula. Similarly Fig.13 shows the result of the simulation when the wheelset is operated on creep curve $C_F$.

## VI. Conclusion and Further Work

The problem of low adhesion and its adverse impact on train control systems and rail network operations present a significant technological challenge to the railway industry. Measures taken by railway industries around the world, such as sanding, water jetting, WSPs e.t.c, have solved the problems caused by low adhesion to some extent. But these measures are not sufficient to eliminate the problems completely. The adhesion detection method presented in this paper lays a scientific foundation for a new way forward. The simulation results presented in this paper affirm the potential of the idea presented. Further work, e.g. track testing and experimental validation, will be needed before it can be put into practice.

## References

[1] Watanabe, T. and A. Yamanaka. "Optimisation of readhesion control of Shinkansen trains with wheel–rail adhesion". Proceedings of the power conversion conferecne, Nagaoka Japan, 1997, p. 47-50.

[2] Xia, F., C. Cole, and P. Wolfs, *An inverse railway wagon model and its applications.* Vehicle System Dynamics, 2007. 45(6): p. 583-605

[3] C. Ward, P.Weston, E. Stewart, H. Li, R. Goodall, C. Roberts, T.X. Mei, G. Charles and R. Dixon, "Condition monitoring opportunities using vehicle based sensors". *IMechE proceedings, Part F: Rail and Rapid Transit*, Vol 225, No.2/2011, pp.202-218.

[4] Charles. G, R. Goodall and R. Dixon. "Model-based condition monitoring at the wheel-rail interface". *Journal of Vehicle System Dynamics*, Volume 46, Supplement 1, pp. 415-430(16), September 2008

[5] Hussain, I. and M.T. X, *Identification of the Wheel Rail Contact Condition for the Traction and Braking Control* Proceedings of the 22nd International Symposium on Dynamics of Vehicles on Roads and Tracks, Manchester Metropolitan University, 14-19 August 2011.

[6] Hussain, M.T.X., *Multi Kalman Filtering Approach for Estimation of Wheel-Rail Contact Conditions* Proceedings of the United kingdom Automatic Control Conferecne 2010, 2010: p. 459-464.

[7] Hussain, M.T.X.a.A.H.J., *Modeling and Estimation of Nonlinear Wheel-rail Contact Mechanics.* Proceedings of the twentieth Intenational conference on System Engineering, 2009: p. 219-223.

[8] Mei, T.X. and I. Hussain. *Detection of wheel-rail conditions for improved traction control*. in *Railway Traction Systems (RTS 2010), IET Conference on*. 2010.

[9] Park. K. T, Lee. H. W, Park. C. H, Kim. D. H and Lee. M. H, "The characteristics of driving control of crane", Proceedings of IEEE International Symposium on Industrial Electronics, Pusan, South Korea, 2001, p. 734 – 739.

[10] Mei, T.X., J.H. Yu, and D.A. Wilson, "A Mechatronic Approach for Anti-slip Control in Railway Traction". Proceedings of the 17th world congress The international federation of automatic control (IFAC), Seoul, Korea, July 2008, p. 8275-8280.

[11] Mei, T.X, J. Yu, and D. Wilson, "A mechatronic approach for effective wheel slip control in railway traction". Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, 2009. Vol. 223(3): p. 295-304.

[12] Li. H and R. Goodal, "Modelling and analysis of a railway wheelset for active control", proceedings of the United Kingdom Automatic Control Conferecne, Swansea UK, 1998, p. 1289 – 1293.

[13] Mei, T.X. and R.M. Goodall, "Practical Strategies for Controlling Railway Wheelsets Independently Rotating Wheels". Journal of Dynamic Systems, Measurement, and Control, 2003. Vol. 125(3): p. 354-360.

[14] Yu, J. H, T.X. Mei and D.A. Wilson, "Re-Adhesion control based on wheelset dynamics in railway traction system". Proceedings of the United Kingdom automatic control conferecne, 2006, Sheffield UK.

[15] Yu. J.H, "Re-adhesion control for AC traction system in railway application", PhD thesis, The University of Leeds, 2007.

[16] Iwnicki. S, "Simulation of wheel-rail contact forces". *Journal of Fatigue & Fracture of Engineering Materials & Structures*, vol. 26 No.10, p. 887-900, 2003.

[17] Goodall, R. and H. Li, "Solid Axle and Independently-Rotating Railway Wheelsets-A Control Engineering Assessment of Stability". *International Journal of Vehicle System Dynamics*, Vol.33(1): p. 57-67, 2000.

[18] Goodall, R., "Tilting trains and beyond. The future for active railwaysuspensions. 2. Improving stability and guidance". *Computing & Control Engineering Journal*, 1999. 10(5): p. 221-230.

[19] Beagley, T., I. McEwen, and C. Pritchard, "Wheel/rail adhesion--Boundary lubrication by oily fluids". *Wear*, Vol. 31(1): p. 77-88, 1975.

[20] Polach, O., *Creep forces in simulations of traction vehicles running on adhesion limit.* Wear, 2005. **258**(7-8): p. 992-1000

[21] Charles, G. and R. Goodall. "Low adhesion estimation". IET International Conference on Railway Condition Monitoring, Birmingham, UK, 29-30 Nov. 2006, ISBN: 0 86341 732 9.

[22] Mei, T.X and R. Goodall. "LQG and GA solutions for active steering of railway vehicles". IEEE proceedings on control theory and applications, January 2000, vol. 147(1), p. 111-117.

[23] F. Xia and P. J. Wolfs, "Estimation of wheel rail interaction forces". U.S Patent 7853412 B2, December 2010.

[24] *F.* Xia, S. Bleakley and P. Wolfs, "The estimation of wheel-rail interaction forces from wagon accelerations", Proceedings of the fourth Australasian Congress on Applied Mechanics, Melbourne, Australia, 16-18 February 2005, pp. 333-338

[25] F. Xia and P. J. Wolfs, "Estimation des forces D'Interaction entre des roues et un rail". WO Patent WO/2006/130,908, 2006.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Fault Diagnosis of Diesel Engine Based on Energy Spectrum Analysis

Hongxia Pan
School of Mechanical Engineering & Automation
North University of China
Taiyuan, China
panhx1015@163.com

Jifang Men
School of Mechanical Engineering & Automation
North University of China
Taiyuan, China
men_jfang@163.com

*Abstract*—**This paper has studied the application of wavelet package energy spectrum and frequency energy spectrum analysis in the diesel engine fault diagnosis. Extracting the fault features by wavelet package energy spectrum and frequency energy spectrum analysis of the fault angle of fuel supply decreased 2.5° and plug of air filter, then making those as the input character of neural networks and implementing the fault diagnosis. It is concluded that frequency energy spectrum analysis is more strongly of the practicability than wavelet package energy spectrum analysis by comparing the test results.**

***Keywords-wavelet packet; energy spectrum; neural networks; fault diagnosis***

## I. INTRODUCTION

With today's society development of the industrialization level, the diesel engine as a common power mechanical equipment, fault diagnosis has become an important research object in the fault diagnosis research field. Main signals used for the analysis of diesel engine's state are vibration signal, noise signal, pressure signal and temperature signal, etc. In diesel engine fault diagnosis, mainly adopts the vibration signal analysis method, and the methods of vibration signal processing include amplitude spectrum, power spectrum, zoom spectral analysis, energy spectrum analysis, cepstrum analysis and other many kinds analysis methods. Diesel engine in the actual operation, the frequency range is not the same in the energy spectrum of the response from the different components. If the energy spatial distribution of some output signals compared with normal system output changes, can distinguish the frequency band corresponding to the component should be in the abnormal working state. Usually some frequency components significantly inhibited, the band signal energy is reduced, and some components are enhanced, the frequency band energy increase. On the energy spectrum density or power spectral density for the analysis, which can be concluded that all kinds of fault characteristic features, so the energy spectrum analysis on fault diagnosis is practical and feasible[1]. This paper based on the energy spectrum analysis on fault diagnosis, is mainly to the wavelet packet energy spectrum analysis and energy spectrum analysis are compared.

## II. WAVELET PACKET ENERGY SPECTRUM ANALYSIS

If the use of wavelet analysis to fault diagnosis of diesel engine, although able to signal the effective time-frequency decomposition, but only on the low frequency band signals decomposition effect is good, the high frequency frequency resolution is poor. Wavelet packet analysis wavelet transform is an improvement of wavelet analysis, it can do further decomposition to the high frequency part for no analysis, is a more accurate analysis method, effectively improves the time-frequency resolution[2]. So here only discuss the wavelet packet energy spectrum analysis method[3].

Steps of Extracting wavelet packet energy spectrum are:

(1) Do wavelet packet decomposition to the sampling signal, this paper adopts four layer wavelet packet decomposition.

(2) on each layer of wavelet packet decomposition coefficients were reconstructed, extracting each frequency band signal on fourth layers, using $S_{4i}$ signify the reconstructed signal of nodes *(4, i)* .

(3) calculation of the frequency band energy

Set $S_{4i}$ corresponding energy is $E_{4i}$ , to seek the total energy of each frequency band signal. It is $E_{4i} = \int \left|S_{4i}(t)\right|^2 dt = \sum_{k=1}^{n} \left|x_{i,k}\right|^2$ . where, $x_{i,k}$ is the amplitude of discrete points of reconstruction signal $S_{4i}$ , and n is sampling points the signal. With these energy values to structure feature vector T, $T = \left[E_{41}, E_{42}, \cdots, E_{415}\ \right]$.

(4) To construct the feature vector. If the each frequency band energy values are larger, then the amount of computation required is large, therefore needs will feature vector is normalized, normalized formula is $S_i = E_{4i}/E$ , where, the total energy E expression is $E = sqrt\left(\sum_{i=0}^{15} \left|E_{4i}\right|^2\right)$.

## III. FREQUENCY DOMAIN ENERGY SPECTRUM ANALYSIS

When the diesel engine system is the normal operation, the system's vibration signal with certain frequency spectrum, if the system in a component failure occurs, with the vibration changes will make the original vibration composition changes, corresponding to its spectrum will have change[4]. Energy spectrum described signal energy along the frequency axis

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

distribution, frequency domain energy spectrum expression is $E = \left| fft\left(x(t)\right)\right|^2$.

The specific method is, this paper adopted the data after reduced frequency, the sample frequency is 10KHz, after FFT transform analysis frequency is 5000Hz. So is will 0 Hz ~ 5000 Hz frequency domain is divided into N bands, separately calculate the each frequency band energy. When the energy is larger, corresponding to the energy of each frequency band value is relatively large, so in order to calculate conveniently, need to put all the energy values for the normalized. Normalized formula is $a_i = \dfrac{a_{\max} - a_i}{a_{\max} - a_{\min}}$

Where, $a_i$ is the each frequency band energy value, $a_{\max}$ expressed the maximum value in $a_i$, $a_{\min}$ is the minimum value in $a_i$. Normalized values can be used as feature value required diesel engine fault diagnosis.

## IV. Application of Neural Networks in Fault Diagnosis

The nature of the fault diagnosis is through the fault symptom conclude fault reason, and the artificial neural network is through to simulate the infer function of human brain to fault diagnosis, and can through the adaptive learning and adjusting the network size, to achieve pattern recognition and feature extraction. Currently main neural network used in fault diagnosis be the BP network, Elman network and RBF network. Which BP network is a multilayer feed-forward neural network, and because of its simple structure, training algorithm is much, so it has been widely used in practice[5].

The BP network learning process includes forward computation and the error back propagation calculation within network, after all training samples input to neural network to calculate the output of the network. If the data from the output layer is not the desired output, then the network automatically enter the error back propagation, by modifying the weights value and the threshold of each layer neurons, and the error is reduced gradually, in this way, until meet the requirements, then train stop. This paper will be mentioned in the wavelet energy spectrum analysis method and frequency domain energy spectrum analysis of the feature value as the input of neural network, select the number of training samples and test samples, through the training to review output, which testified in front of the feasibility of the proposed method[6].

## V. Diesel Engine Fault Diagnosis Example Analysis

In this paper, the experimental data are in the normal operation of diesel engines, artificial set of fuel supply advance angle decreased 2.5 degree and the air filter is clogged two fault, through the acquisition of diesel engine vibration signal to do fault diagnosis. To collecting data the sampling frequency is 40KHz, diesel engine speed is 1500r / min.

First the collected vibration signals preprocessing, then using wavelet packet transform to extract signals feature vector. the vector signal. Using the previously described extracting wavelet packet energy spectrum method, using the db1 of in Daubechies wavelet series do 4 layer wavelet decomposition of signal, a total of 16 frequency bands. Extracting the fourth layer wavelet packet each frequency band energy spectrum scales as the neural network's input vector P, namely *P = [E40 / E, E41 / E, E42 / E, E43 / E, E44 / E, E45 / E, E46 / E, E47 / E, E48 / E, E49 / E, E410 / E, E411 / E, E412 / E, E413 / E, E414 / E, E415 / E]*.

Each condition calculation six samples, set the number of training sample is 12, in which the normal sample number, fault sample number of fuel supply advance angle decreased 2.5 degree and fault sample number of air filter clog all are 4. The test samples were set to 6, normal and two kinds of fault sample number all are 2. Because there are two kinds of failure mode and a normal state, so the network output is expressed as follow, fuel supply advance angle decreased 2.5 degree fault is (1,0,0), the air filter is clogged fault is (0,1,0), the normal state is (0,0,1). The training samples and test samples are shown as Table 7 and Table 1.

TABLE I. Test Samples

| Serial Number | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| E40/E | 0.026 | 0.045 | 0.037 | 0.024 | 0.007 | 0.011 |
| E41/E | 0.044 | 0.066 | 0.059 | 0.062 | 0.022 | 0.020 |
| E42/E | 0.133 | 0.151 | 0.152 | 0.138 | 0.046 | 0.068 |
| E43/E | 0.079 | 0.109 | 0.131 | 0.115 | 0.043 | 0.059 |
| E44/E | 0.129 | 0.140 | 0.190 | 0.150 | 0.074 | 0.108 |
| E45/E | 0.130 | 0.114 | 0.137 | 0.117 | 0.058 | 0.057 |
| E46/E | 0.180 | 0.242 | 0.243 | 0.244 | 0.103 | 0.111 |
| E47/E | 0.231 | 0.172 | 0.251 | 0.218 | 0.153 | 0.170 |
| E48/E | 0.130 | 0.114 | 0.137 | 0.117 | 0.058 | 0.057 |
| E49/E | 0.180 | 0.242 | 0.243 | 0.244 | 0.103 | 0.111 |
| E410/E | 0.566 | 0.334 | 0.343 | 0.412 | 0.256 | 0.261 |
| E411/E | 0.200 | 0.173 | 0.212 | 0.211 | 0.106 | 0.072 |
| E412/E | 0.136 | 0.159 | 0.215 | 0.193 | 0.084 | 0.093 |
| E413/E | 0.173 | 0.142 | 0.156 | 0.165 | 0.126 | 0.082 |
| E414/E | 0.556 | 0.678 | 0.526 | 0.519 | 0.859 | 0.829 |
| E415/E | 0.293 | 0.346 | 0.437 | 0.446 | 0.318 | 0.380 |
| Ideal Output | 100 | 100 | 010 | 010 | 001 | 001 |

Set the input layer neuron number is 16, the output layer neuron number is 3, number of hidden layer neurons is approximately 33, the number of training is 1000, target of the training error is 0.00001, the learning rate is 0.1. After setting up parameters, establishes the BP neural network, and got the training results are shown as figure 1.

From the diagram that the curve of training error can be seen, after 70time after training, the network can achieve our learning goals. Then will the test sample as a neural network's input, to get the test results are shown as Table 2.

From the test results, and we expected output consistent with, can accurate analysis the failures, this method is desirable, but consume time is longer.

Repeat the above, using the wavelet db5 of in Daubechies wavelet series do 4 layer signal decomposition, the same extracting the fourth layer wavelet packet each frequency band energy spectrum scales as the neural network's input vector P, output mode and the above ware same, training samples and test samples used the same calculation method, to train of neural network set up the same parameters, get training results are shown as figure 2.

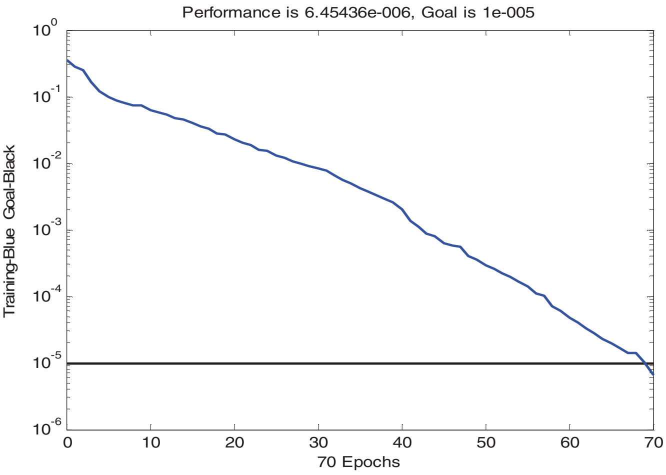


Figure 1. BP neural network error training curve of wavelet packet transform db1

TABLE II. TESTING RESULTS

| Sample Output | 1 | 2 | 3 | Testing Results |
|---|---|---|---|---|
| 1 | 0.9998 | 0.0000 | 0.0001 | Fuel supply advance angle decreased 2.5° |
| 2 | 0.9855 | 0.0001 | 0.0003 | Fuel supply advance angle decreased 2.5° |
| 3 | 0.0000 | 1.0000 | 0.0091 | Air filter clog |
| 4 | 0.0000 | 1.0000 | 0.0019 | Air filter clog |
| 5 | 0.0141 | 0.0000 | 0.9978 | Normal state |
| 6 | 0.0000 | 0.0000 | 1.0000 | Normal state |

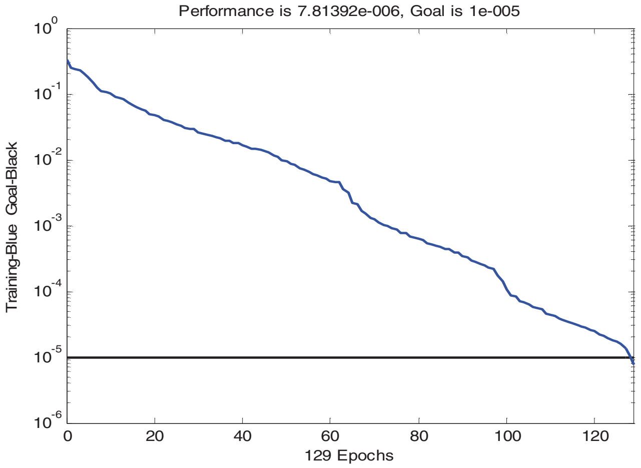


Figure 2. BP neural network error training curve of wavelet packet transform db5

After 129 time training, the error of the network meet the requirements. Then will the test sample as the input of neural network, to get the test results are shown in Table 3.

TABLE III. TESTING RESULTS

| Sample Output | 1 | 2 | 3 | Testing Results |
|---|---|---|---|---|
| 1 | 0.9997 | 0.0001 | 0.0004 | Fuel supply advance angle decreased 2.5° |
| 2 | 1.0000 | 0.0000 | 0.0000 | Fuel supply advance angle decreased 2.5° |
| 3 | 0.0009 | 0.5604 | 0.0002 | Air filter clog |
| 4 | 0.9888 | 0.9954 | 0.0000 | Can't judge |
| 5 | 0.0000 | 0.0228 | 1.0000 | Normal state |
| 6 | 0.0000 | 1.0000 | 0.3913 | Air filter clog |

From the test results, some value and our ideal output is not consistent, emerged condition of failure error separation, so this method is not desirable.

It is introduced with the frequency domain energy spectrum method to extract the feature vector. This paper uses the data sampling frequency is 40KHz. Because the frequency is too high, so the need to reduce analysis frequency, the reduced frequency is 10 KHz, after done power spectrum the frequency is 5 KHz, will this frequency are divided into 10 band, each band energy be calculated. Choose after normalized energy value S as fault diagnosis feature value, namely the neural network's input vector P, P = [S1, S2, S3, S4, S5, 'S6', S7, S8, S9, S10].

Similarly, each condition selection six samples, set the number of training sample is 12, in which the normal sample number and fault sample number all are 4. The test samples were set to 6, normal and two kinds of fault sample number all are 2. The network output is expressed as follow, fuel supply advance angle decreased 2.5 degree fault is (1,0,0), the air filter is clogged fault is (0,1,0), the normal state is (0,0,1). The training samples and test samples are shown as Table 8 and Table 4.

TABLE IV. TEST SAMPLES

| Serial Number | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| S0 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| S1 | 0.934 | 0.934 | 0.935 | 0.935 | 0.929 | 0.929 |
| S2 | 0.867 | 0.867 | 0.900 | 0.900 | 0.854 | 0.854 |
| S3 | 0.474 | 0.474 | 0.697 | 0.697 | 0.556 | 0.556 |
| S4 | 0.790 | 0.790 | 0.870 | 0.870 | 0.838 | 0.838 |
| S5 | 0.867 | 0.867 | 0.852 | 0.852 | 0.862 | 0.862 |
| S6 | 0.154 | 0.154 | 0.231 | 0.231 | 0.017 | 0.017 |
| S7 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| S8 | 0.810 | 0.810 | 0.842 | 0.842 | 0.902 | 0.902 |
| S9 | 0.764 | 0.764 | 0.774 | 0.774 | 0.802 | 0.802 |
| Ideal Output | 100 | 100 | 010 | 010 | 001 | 001 |

Set the input layer neuron number is 10, the output layer neuron number is 3, number of hidden layer neurons is approximately 21, the number of training is 1000, training target is 0.00001, the learning rate is 0.1. After setting up parameters, establishes the BP neural network, and got the training results are shown as figure 3.

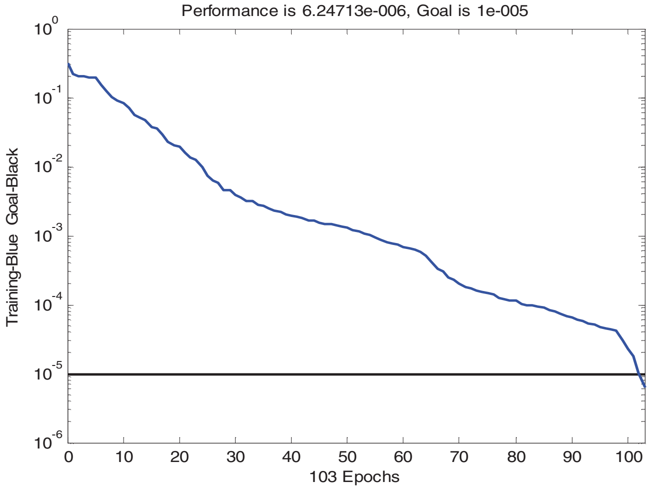


Figure 3. BP neural network error training curve of frequency domain energy spectrum

After 103 time training, the error of the network meet the requirements. Then will the test sample as the input of neural network, to get the test results are shown in Table 5.

TABLE V. TESTING RESULTS

| Sample Output | 1 | 2 | 3 | Test Results |
|---|---|---|---|---|
| 1 | 0.9983 | 0.0011 | 0.0012 | Fuel supply advance angle decreased 2.5° |
| 2 | 0.9983 | 0.0011 | 0.0012 | Fuel supply advance angle decreased 2.5° |
| 3 | 0.0036 | 0.9956 | 0.0003 | Air filter clog |
| 4 | 0.0036 | 0.9956 | 0.0003 | Air filter clog |
| 5 | 0.0006 | 0.0000 | 0.9998 | Normal state |
| 6 | 0.0006 | 0.0000 | 0.9998 | Normal state |

From the curve of training error, network can rapidly astringe. From the testing results, the fault separation accuracy is relatively high, and the time is short.

Using the three analysis methods to extract feature vector which input to BP neural network, the training time is respectively are shown in Table 6.

TABLE VI. TRAINING TIMES

| Analysys methode | wavelet packet transform db1 | wavelet packet transform db5 | Frequency domain energy spectrum |
|---|---|---|---|
| Elapsed_time | 39.5000 | 73.4540 | 3.3440 |

From the above two methods of comparison can be seen, the wavelet packet energy spectrum analysis although can to signal fine time-frequency decomposition, but the wavelet function selection is a difficult problem of the wavelet packet transform. Some can accurately diagnose the fault, but some can not, like the db5 used above. Because the kinds of fault could not only one, so want to choose for each fault are suitable wavelet packet transform is more difficult. Due to the signal decomposition and reconstruction, so the calculation amount is larger, longer time consuming, not suitable for online diagnosis. And the frequency domain energy spectrum analysis method needs less computation amount, shorter time, and the correct rate is higher, thus suitable for online diagnosis.

## VI. CONCLUSION

This paper uses wavelet packet energy spectrum analysis method and frequency domain energy spectrum analysis method to the diesel engine vibration signal fault diagnosis, after verification, if the problem of selecting wavelet function can be improved, then the wavelet packet energy spectrum analysis method is a good method of fault diagnosis on the above analysis, the results show the frequency domain energy spectrum method in the practical application is feasible, less time-consuming, accuracy is relatively high.

## ACKNOWLEDGMENT

The work described in this paper has been supported by National Natural Science Foundation of China under the grant No. 51175480, National Doctoral Fund of Ministry of Education under the grant No.20091420110002, Province Natural Science Foundation of Shanxi under the grant No.2011011019 and Shanxi Scholarship Council of China. The authors would like to express their gratitude for the support of this study.

## REFERENCES

[1] Tang Youhuai, Zhang Haitao, Luo Shan Jiang Zhe. “Based on wavelet packed energy spectrum analysis of motor fault diagnosis”, Foreign Electronic Measurement Technology, 2008, Vol. 27 (2), pp.54-55.

[2] Ge, Zhexue. and Sha, Wei. 2007. “Wavelet Analysis Theory MATLAB R2007 Realizaton”, Beijing, Publishing House of Electronics Industry.

[3] Pan Hongxia, Yao Zhuting, Huang Jinying. “Techniques on Monitoring and Fault Diagnosis for Vehicle Power Transmission Systems”, Journal of North China Institute of Technology, 2002, Vol.23(2): pp.109-113

[4] Zhai Guangrui, Pan Hongxia. “Diesel Engine Vibration Signal Analysis Based on Improved Local Wave Method”, Vehicle Engine, 2009, (1), pp.77-81

[5] Cheng Baojia, Li Li, Zhang Yuan. “Application of Scale-Wave Power Spectrum to Fault Diagnosis of Internal Combustion Engine”, Transactions of CSICE, 2006 (3), pp.284-287.

[6] Cao, Longhan. and Cao, Changxiu. The Research of Diesel Engine Fault Diagnosis with ANN Based on Rough Sets Theory. *Transactions of Csice.* 2002, Vol. 20 (4), pp. 357-361.

TABLE VII. TRAIN SAMPLES

| **Serial Number** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| E40/E | 0.028 | 0.038 | 0.025 | 0.035 | 0.031 | 0.029 | 0.035 | 0.043 | 0.014 | 0.020 | 0.019 | 0.012 |
| E41/E | 0.055 | 0.069 | 0.034 | 0.056 | 0.060 | 0.055 | 0.075 | 0.065 | 0.024 | 0.025 | 0.032 | 0.025 |
| E42/E | 0.148 | 0.154 | 0.156 | 0.120 | 0.165 | 0.138 | 0.127 | 0.164 | 0.074 | 0.076 | 0.076 | 0.100 |
| E43/E | 0.097 | 0.125 | 0.113 | 0.129 | 0.133 | 0.128 | 0.136 | 0.121 | 0.061 | 0.078 | 0.074 | 0.093 |
| E44/E | 0.134 | 0.181 | 0.117 | 0.157 | 0.170 | 0.124 | 0.178 | 0.152 | 0.105 | 0.110 | 0.134 | 0.108 |
| E45/E | 0.099 | 0.155 | 0.112 | 0.170 | 0.152 | 0.136 | 0.143 | 0.150 | 0.092 | 0.105 | 0.118 | 0.092 |
| E46/E | 0.176 | 0.226 | 0.240 | 0.248 | 0.224 | 0.221 | 0.279 | 0.239 | 0.121 | 0.124 | 0.176 | 0.173 |
| E47/E | 0.180 | 0.283 | 0.163 | 0.195 | 0.254 | 0.256 | 0.242 | 0.220 | 0.229 | 0.213 | 0.270 | 0.265 |
| E48/E | 0.099 | 0.155 | 0.112 | 0.170 | 0.152 | 0.136 | 0.143 | 0.150 | 0.092 | 0.105 | 0.118 | 0.092 |
| E49/E | 0.176 | 0.226 | 0.240 | 0.248 | 0.224 | 0.221 | 0.279 | 0.239 | 0.121 | 0.124 | 0.176 | 0.173 |
| E410/E | 0.388 | 0.453 | 0.634 | 0.539 | 0.503 | 0.459 | 0.446 | 0.478 | 0.264 | 0.305 | 0.280 | 0.463 |
| E411/E | 0.155 | 0.223 | 0.196 | 0.245 | 0.216 | 0.198 | 0.200 | 0.261 | 0.126 | 0.112 | 0.122 | 0.120 |
| E412/E | 0.120 | 0.161 | 0.112 | 0.164 | 0.212 | 0.199 | 0.202 | 0.160 | 0.114 | 0.100 | 0.128 | 0.104 |
| E413/E | 0.172 | 0.187 | 0.147 | 0.132 | 0.162 | 0.194 | 0.175 | 0.157 | 0.114 | 0.126 | 0.133 | 0.104 |
| E414/E | 0.683 | 0.527 | 0.468 | 0.508 | 0.505 | 0.546 | 0.508 | 0.462 | 0.807 | 0.777 | 0.700 | 0.615 |
| E415/E | 0.385 | 0.311 | 0.294 | 0.262 | 0.290 | 0.346 | 0.321 | 0.401 | 0.343 | 0.377 | 0.436 | 0.439 |
| Ideal Output | 100 | 100 | 100 | 100 | 010 | 010 | 010 | 010 | 001 | 001 | 001 | 001 |

TABLE VIII. TRAIN SAMPLES

| **Serial Number** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| S0 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| S1 | 0.957 | 0.934 | 0.934 | 0.934 | 0.932 | 0.935 | 0.935 | 0.935 | 0.929 | 0.929 | 0.929 | 0.929 |
| S2 | 0.919 | 0.867 | 0.867 | 0.867 | 0.891 | 0.900 | 0.900 | 0.900 | 0.818 | 0.854 | 0.854 | 0.854 |
| S3 | 0.731 | 0.474 | 0.474 | 0.474 | 0.687 | 0.697 | 0.697 | 0.697 | 0.582 | 0.556 | 0.556 | 0.556 |
| S4 | 0.895 | 0.790 | 0.790 | 0.790 | 0.870 | 0.870 | 0.870 | 0.870 | 0.806 | 0.838 | 0.838 | 0.838 |
| S5 | 0.910 | 0.867 | 0.867 | 0.867 | 0.858 | 0.852 | 0.852 | 0.852 | 0.810 | 0.862 | 0.862 | 0.862 |
| S6 | 0.168 | 0.154 | 0.154 | 0.154 | 0.432 | 0.231 | 0.231 | 0.231 | 0.071 | 0.017 | 0.017 | 0.017 |
| S7 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| S8 | 0.825 | 0.810 | 0.810 | 0.810 | 0.832 | 0.842 | 0.842 | 0.842 | 0.824 | 0.902 | 0.902 | 0.902 |
| S9 | 0.829 | 0.764 | 0.764 | 0.764 | 0.863 | 0.774 | 0.774 | 0.774 | 0.793 | 0.802 | 0.802 | 0.802 |
| Ideal Output | 100 | 100 | 100 | 100 | 010 | 010 | 010 | 010 | 001 | 001 | 001 | 001 |

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Fault Diagnosis using Magnetic Image of PCB

Zhuting Yao
College of Mechanical Engineering and Automatization
North University of China
030051 Taiyuan, Shanxi, China
ztyao713@163.com

Hongxia Pan
College of Mechanical Engineering and Automatization
North University of China
030051 Taiyuan, Shanxi, China
Panhx1015@163.com

*Abstract*—**With the digital technology and very large scale integrated circuit technology widely used, the structure and function of electronic devices are becoming more and more complex, the defects of contact diagnosis has become increasingly prominent for using the probe or needle bed, it demands people began to in-depth study the diagnostic techniques of PCB non-contact. Based on magnetic image of the PCB fault diagnosis technology is a new non-destructive testing technology of PCB, developed in recent years, it can achieve rapid detection and location of the circuit board failure, improve the reliability of the circuit board. Aimed at the specific PCB, its different failure modes of the magnetic image of the PCB are attained by Ansoft software, are carried out through the magnetic image filtering, image enhancement technology; they are completed on multi-level wavelet decomposition, the construction and extraction of wavelet energy features of the magnetic image. The PCB fault diagnosis based on the magnetic image is completed by using the improved momentum-adaptive rate neural network algorithm. The results show that the magnetic image method is effectively diagnosis method for PCB, and it's a new fault diagnosis approach of PCB.**

***Keywords- fault diagnosis；magnetic image；printed circuit boards；wavelet neural network***

## I. INTRODUCTION

As we all know, the circuit board is in the state of power, the current line and the circuit board components are magnetized, such as resistors, integrated chips will produce electromagnetic radiation, and a particular magnetic field is formed in space [1]. It has one to one relationship between the magnetic field distribution characteristics of the circuit and current distribution characteristics of the circuit. When a component or circuit of the circuit board appears failure, it will cause the change of the current size and direction, and the magnetic field strength at any point in the near-field region of board depends on the size of current and its distribution, it makes nearly magnetic field distribution also undergo corresponding changes. The state of the circuit board is determined by detecting the magnetic field distribution in the circuit or system work, and it will be able to achieve non-contact detection and diagnosis, the fault diagnosis of the magnetic image is based on [2].

During the circuit board fault diagnosis, the first it needs to add excitation source, it makes the circuit board works in the largest working condition, then the magnetic field distribution of the circuit board is detected, and a test data is formed. Comparing the measured magnetic field distribution data of the circuit board with the trouble-free data, if the circuit board presences the fault, both of the magnetic field distribution data must be some difference. According to the difference, the faulty components on the circuit board can be determined and located. If it can not locate to the fault components, it can compress the range of failure, and add power in possible malfunction of the various components, it makes them on the maximum working state, measures the magnetic field distribution, extracts the characteristics vector of test data, based on the correspondence relation between the failure and characteristics vector of the magnetic image, reasoning judge, to determine the fault source, and show the location and types of faults.

## II. SYSTEM COMPONENTS AND FUNCTIONS

Based on the magnetic image of PCB fault diagnosis system, the first, it need process the input magnetic image, extract the characteristics information of the magnetic image of board in the work state, and the characteristics information are as neural network inputs, the corresponding fault cause are as network output and network is trained. Until the network training is completed, the magnetic image of the pre-test circuit board is inputted, and the relevant features are extracted, and the board working states can be diagnosed. The fault diagnosis system block diagram is shown in Fig.1.

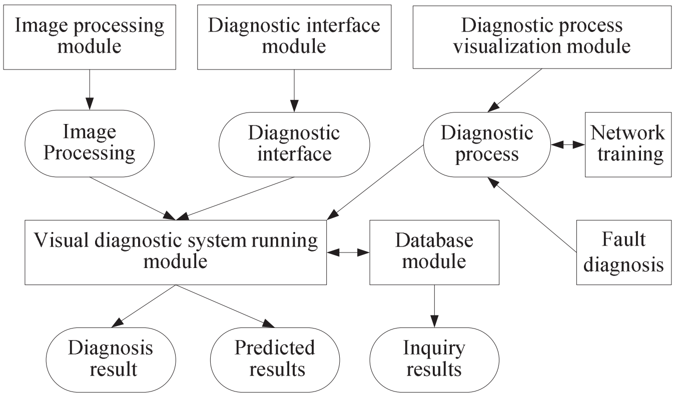


Figure 1. The design frame of system

## III. To Obtain The Magnetic Images

Sometimes the magnetic images can be obtained by the EMSCAN devices and software simulation. In this paper, the electromagnetic image in near-field experimental data is indirectly obtained by Ansoft Designer simulation software, and it provides a new diagnostic method for circuit board.

The electromagnetic near-field image of a circuit board by simulating is shown in Fig. 2. As the software simulation from the distribution of near-field magnetic image is color, making image analysis must be conducted before the gray-scale image processing.

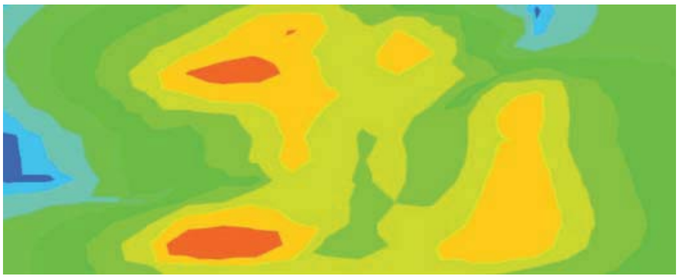


Figure 2. Simulation image of a circuit board magnetic image

## IV. The Magnetic Image Processing Techniques

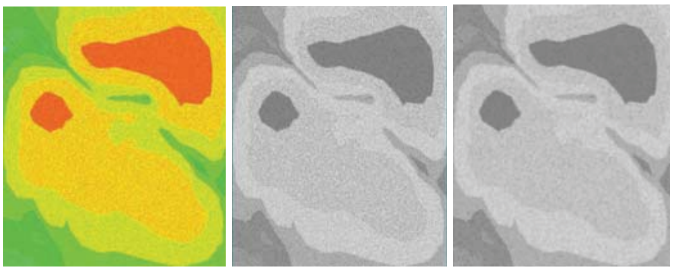
In practice, the magnetic image in the detection process is affected by noise, electromagnetic interference and other external factors, always makes the original image acquires a certain degree of distortion, the images must be digital processing before the magnetic image is analyzed, the images processing in this system includes image filtering, enhancement, and edge sharpening and so on.

### A. Gray-scale and filter of the magnetic image

Gray-scale image is the process transforms a color image into grayscale image. The magnetic image is a color image, shown in Fig. 3(a). Color images contain a large number of R, G, B color information, it not only would take a lot of storage space, but also it will require the system to assign a lot of resources in the image application process. As the grayscale images can be the overall reaction and the local color, brightness distribution and characteristics of the whole image. It can save a variety of hardware and software resources when color images convert to grayscale images.

By detecting the magnetic field data to generate the magnetic field grayscale images, it reflects the magnetic field distribution of the circuit board. Image grayscale of the magnetic images generation process includes the magnetic field data acquisition, extracting the maximum and minimum of array data classification and the data is conversed into the magnetic image gray-scale values.

The gray-scale magnetic image is shown in Fig. 3(b). It can be seen from Fig. 3(b), there are some noise points in the image, therefore, median filtering is used in the grayscale image, the point in digital image is replaced by the median value of brightness of each point in a little neighborhood, so that it can protect the edge of the image, and it has good results to the filter pulse interference and scanning image noise [3].



(a)Source image (b)Gray-scale images (c)Median filtered image

Figure 3. The contrast image from source image with gray-scale image and median filtering image

Set the number of groups: $X_1, X_2, \cdots X_n$,If the order is as follows: $X_{i1} \le X_{i2} \le X_{i3} \le \cdots \le X_{in}$, then:

$$Y = Med\{X_1, X_2, \cdots X_n\} = \begin{cases} X_{i[(n+1)/2]} & n \text{ is odd number} \\ \frac{1}{2}\left[X_{i[(n/2]} + X_{i[(n/2+1]}\right] & n \text{ is even number} \end{cases} \qquad (1)$$

Y is called the median of $X_1, X_2, \cdots X_n$, and it is extended to two-dimensional, then the median filtering of the output is:

$$Y_i = med\{X_{ij}\} = med\{X_{(i+r)(j+s)}(r,s) \in A, (i,j) \in I^2\} \qquad (2)$$

According to the characteristics of the magnetic image, selecting the 5×5 window median filtering method to obtain the magnetic image, it not only can improve the edge characteristics of image signal, and also get faster processing speed and higher accuracy. After filtering the image is shown in Fig.3(c), it can be seen that the noise points in the image are significantly reduced, the smoothness of magnetic image is better, the overall effect is satisfactory.

### B. The magnetic image enhancement

Gray value of pixels within the image is an important data of the image, the result of magnetic image gray value is often less than ideal due to testing equipment or filtering effect. Image enhancement can highlight some interesting information or attenuate no interesting information in image, and it is more effective than the original image in the specific application.

Common methods of image enhancement are gray-scale transformation and Histogram Equalization. The histogram equalization method is adopted in this paper. The basic idea of Histogram equalization is to increase the image's pixel gray value dynamic range. The actual processing of histogram correction achieves on the cumulative distribution function transformation. For digital images, the transformation function is set as follow:

$$s_k = T(r_k) = \sum_{j=0}^{k} \frac{n_j}{n} = \sum_{j=0}^{k} p(r_j) \qquad (3)$$

Where, $0 \le r_j \le 1, k = 0,1,\cdots,L-1, p(r_j)$ is the gray probability that gray level appears, and $T(r_k)$ is the cumulative distribution function of *r*. In practice, according to meet the actual needs in digital image processing, the conversion result is a real number, also need to re-quantitative, quantitative formula is as follow:

$$\hat{s}_k = INT\left[\frac{s_k - s_{\min}}{1 - s_{\min}}(L-1) + 0.5\right] \qquad (4)$$

Fig.4 shows the magnetic image before and after histogram equalization. It can be seen, with the histogram correction, the gray-scale range and gray level intervals of the image histogram are pulled larger, the image become clearer.

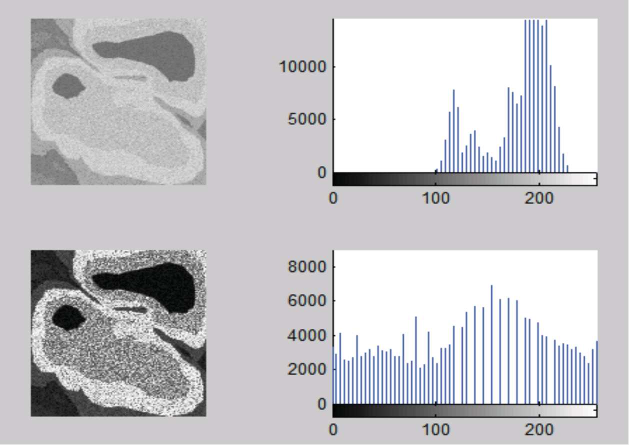


Figure 4. The magnetic image after the Histogram equalization

### *C. Sharpen edge of the magnetic image*

In the image smoothing process, it tends to the image boundaries, contours and details become blurred due to image conversion system transfer function attenuation effects on high-frequency components. In order to reduce these impacts, the image enhancement techniques are adopted in various parts of the contour lines and details of more clarity in the image [4].

Laplacian is selected because it has better effect when it is used in edge detection, and it has rotation invariance. Laplacian is a linear quadratic differential operators, it can meet the different sharpening image requirements.

For images $F(x,y)$ , the expression for the Laplacian operator is as follow:

$$\nabla^2 F(x,y) = \frac{\partial^2 F(x,y)}{\partial^2 x} + \frac{\partial^2 F(x,y)}{\partial^2 y} \qquad (5)$$

Figure 5 shows the magnetic image of PCB after edge sharpening by Laplacian.

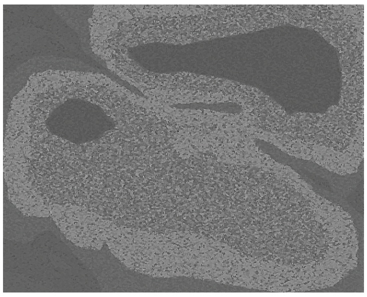


Figure5. Magnetic image of edge sharpening

## V. WAVELET ENERGY FEATURE EXTRACTION OF MAGNETIC IMAGE

In practice fault diagnosis, it is always hope to collect a number of samples to train the neural network, and the fault diagnosis is more accurate and reliable. But too much input sample can cause not only the training time is too long, or even obstruct the convergence effects of network, mainly the training accuracy can not arrive, and ultimately affect the circuit board fault diagnosis. Therefore, it requires to extract useful information is a greater contribution to the faults diagnosis from the sample, image feature extraction is to use existing image samples by mapping (or function transformation) to construct a lower dimensional feature space, the original image contains useful information is mapped in a few features.

The two-dimensional discrete wavelet transform (DWT) is adopted because the wavelet transformation has a good effect in the signal weakened or even removed the correlation between the various characteristics by selecting the appropriate filter function, and it has a multi-scale, multi-resolution features. It can be used correlation analysis in the low frequency band by high frequency resolution and low time resolution; at high frequencies, the available low-frequency resolution and high time resolution achieve correlation analysis, and the magnetic image data processing is a two-dimensional gray-scale image data [5].

### *A. Two-dimensional discrete wavelet transform*

First, define the scale and translation functions as follow:

$$\varphi_{j,m,n}(x,y) = 2^{j/2}\varphi\left(2^j x - m, 2^j y - n\right) \qquad (6)$$

$$\psi^i_{j,m,n}(x,y) = 2^{j/2}\psi^i\left(2^i x - m, 2^j y - n\right), i = \{H,V,D\} \qquad (7)$$

Thereinto, $\varphi(x,y)$ represents a two-dimensional scaling function, $\psi^H(x,y)$、$\psi^V(x,y)$、$\psi^D(x,y)$ represent two-dimensional wavelet measurement function, respectively. The properties of $\varphi(x,y)$、$\psi^H(x,y)$、$\psi^V(x,y)$、$\psi^D(x,y)$ are as follow:

$$\varphi(x,y) = \varphi(x)\varphi(y) \qquad (8)$$

$$\psi^H(x,y) = \psi(x)\varphi(y) \qquad (9)$$

$$\psi^V(x,y) = \varphi(x)\psi(y) \qquad (10)$$

$$\psi^D(x,y) = \psi(x)\psi(y) \qquad (11)$$

These wavelet measured functions are changed along different directions of the image gray value: $\psi^H$ changes along the image column, $\psi^V$ changes along the line of the image, $\psi^D$ changes along the diagonal direction of the image.

Matrix of size $M \times N$ image $f(x,y)$ of the discrete wavelet transformation is as follow:

$$W_\varphi(j_0,m,n) = \frac{1}{\sqrt{MN}}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)\varphi_{j_0,m,n}(x,y) \qquad (12)$$

$$W^i_\psi(j_0,m,n) = \frac{1}{\sqrt{MN}}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)\varphi^i_{j,m,n}(x,y), i = \{H,V,D\} \qquad (13)$$

$j_0$ is beginning scale, it can be taken any value, $W_\varphi(j_0,m,n)$ coefficient defined in the scale $j_0$ factor approximation of $f(x,y)$ · $W_\psi^i(j,m,n)$ coefficient will be additional the horizontal, vertical and diagonal details if $j \geq j_0$. In normal circumstances, let $j_0 = 0$ , $M = N = 2^j$ , $j = 0,1,2,\cdots,J-1, m,n = 0,1,2,\cdots 2^j - 1$ , equation (12) and (13) give the value of $W_\phi$ and $W_\psi^i$ , the original image $f(x,y)$ can be got by inverse discrete wavelet transformation, it expresses as follow:

$$f(x,y) = \frac{1}{\sqrt{MN}} \sum_m \sum_n W_\varphi(j_0,m,n) \varphi_{j_0,m,n}(x,y) \tag{14}$$
$$+ \frac{1}{\sqrt{MN}} \sum_{i=H,V,D} \sum_{j=j_0}^{\infty} \sum_m \sum_n W_\varphi(j_0,m,n) \psi_{j,m,n}^{i}(x,y)$$

*B. The structure and extraction of wavelet energy feature*

Set, $H_i$ 、 $V_i$ and $D_i$ are the details of the image in the $i^{\text{th}}$ wavelet decomposition in the horizontal direction, vertical direction and diagonal direction, the wavelet energy of image in different directions of the $i^{\text{th}}$ level is defined as[6]:

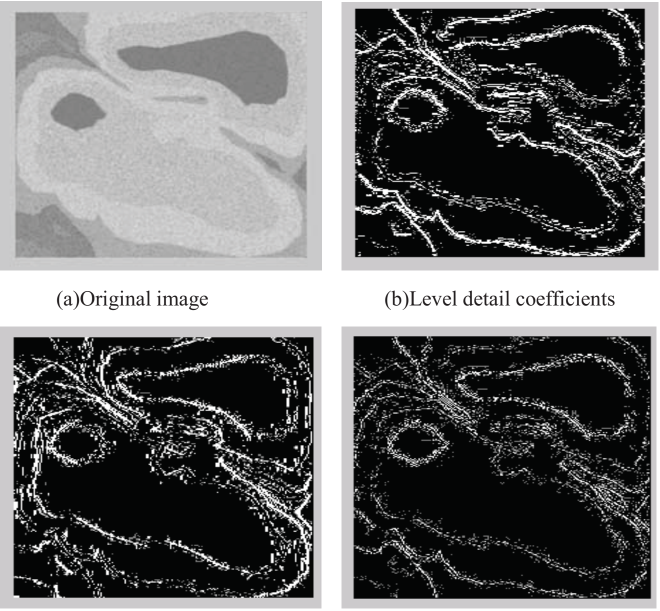
$$E_i^h = \sum_{x=1}^{M} \sum_{y=1}^{N} [H_i(x,y)]^2 \tag{15}$$

$$E_i^v = \sum_{x=1}^{M} \sum_{y=1}^{N} [V_i(x,y)]^2 \tag{16}$$

$$E_i^d = \sum_{x=1}^{M} \sum_{y=1}^{N} [D_i(x,y)]^2 \tag{17}$$

From (15) to (17), they stand for the intensity information of the horizontal, vertical and diagonal directions in the $i^{\text{th}}$ level wavelet decomposition of images, respectively.

The wavelet decomposition details image in each direction of the circuit board's magnetic image are shown in Figure 6.


(a)Original image

(b)Level detail coefficients

(c) Vertical detail coefficients

(d) Diagonal detail coefficients

Figure 6 Image single decomposition

During image wavelet decomposition, the non-oscillating signal wavelet coefficient will increase with the wavelet decomposition series increases, the coefficients of wavelet decomposition level become the maximum when the oscillation signal is corresponding with the oscillation frequency. Therefore, the energy of the non-oscillation of the magnetic image part is concentrated in the large-scale wavelet decomposition detail image, and the energy of the oscillation of the magnetic image part is mainly concentrated in the smaller-scale wavelet decomposition detail image.

The vector of the energy component by wavelet decomposition level is as follow:

$$\left(E_i^h, E_i^v, E_i^d\right)_{1,2,\cdots M} \tag{18}$$

It can reflect the energy distribution of the magnetic image. Where, $M$ is the maximum series number of wavelet decomposition.

Obviously, equation (18) describes the overall information of the magnetic image, and it can not describe some local information of the magnetic image distribution. To solve this problem, the each images details are divided into $S$×$S$ disjoint blocks, and the each wavelet energy is calculated. Finally, the local feature vector $\widetilde{V}$ contains many details of the resolution information are composed by the energy of these blocks:

$$\widetilde{V} = \left(\widetilde{V}_{(1)}^1, \widetilde{V}_{(2)}^1, \cdots \widetilde{V}_{(3\times S\times S)}^1, \cdots \widetilde{V}_{(1)}^M, \widetilde{V}_{(2)}^M, \cdots, \widetilde{V}_{(3\times S\times S)}^M\right) \tag{19}$$

Where, M is the total series number of wavelet decomposition, $\widetilde{V}_{(j)}^1 (j = 1,\cdots,3 \times S \times S)$ are each images block energy that detail coefficients image $H_i$、 $V_i$、 $D_i$ of the $i^{\text{th}}$ level wavelet decomposition is divided into $S \times S$ blocks. Since the wavelet energy feature is a combination of the each level wavelet energy feature, the local features can be reflected the image at different resolutions, different locations, different directions, it can be used as basis to system identification or diagnosis.

Typically, the various samples eigenvalues are required normalized before the image feature value is as the basis for fault diagnosis, it is better identification or diagnosis when the normalized eigenvalues replace the original features.

$$V = \left(V_{(1)}^1, V_{(2)}^1, \cdots V_{(3\times S\times S)}^1, \cdots V_{(1)}^M, V_{(2)}^M, \cdots, V_{(3\times S\times S)}^M\right) \tag{20}$$

$$V_{(j)}^i = \frac{\widetilde{V}_{(j)}^i}{\sum_{k=1}^{M} \sum_{l=1}^{3\times S\times S} \widetilde{V}_{(l)}^k} \quad (i = 1,\cdots,M; j = 1,\cdots,3 \times S \times S \tag{21}$$

The feature vector $V$ is the wavelet energy feature.

$i^{\text{th}}$ level wavelet energy feature is as follow:

$$V^i = \left(V_{(1)}^i, V_{(2)}^i, \cdots V_{(3\times S\times S)}^i\right) \tag{22}$$

Wavelet energy feature is combined by the wavelet energy levels character, as the magnetic image, the wavelet energy feature of image is combined all levels coefficients of the wavelet energy feature by the wavelet decomposition, and multi-scale wavelet analysis of texture features is obtained, it can reflect the characteristics of the magnetic image distribution.

## VI. The Diagnosis Algorithm Of The Magnetic Images

### A. Improved BP Algorithm

To speed up the learning speed of the network, the improved BP network learning algorithms- Momentum - adaptive learning rate BP algorithm is used when the weights and thresholds are adjusted. The algorithm synthesis the advantage of momentum BP algorithm can avoid training process causes local minima during and adaptive variable learning rate back-propagation (VLBP) algorithm has character of weight adjustment smooth and short training time.

Momentum BP algorithm: when the momentum BP algorithm amend its weight, it not only to consider the error role in the gradient, but also it consider the change trends in the error surface. It adds a proportional the previous weight change value in the each weights change, and according to the back-propagation method to generate a new weight change. Its essence is through a momentum factor to transfer the impacts of weight change in the last time. The weight adjustment formula with additional momentum factor is as follow:

$$mc = \begin{cases} 0 & SSE(k) \succ 1.04 \cdot SSE(k-1) \\ 0.95 & SSE(k) \prec SSE(k-1) \\ mc & \text{other} \end{cases} \tag{23}$$

*k* is the training times, mc is the momentum factor, it is generally 0.95.

Adaptive Variable Learning Rate Back-propagation, aimed at ensuring the algorithm is stable, the convergence speed of network is faster, and the learning time is short. The learning rate is appropriately adjusted by the error surface. The weight of the correction value is checked whether it really lowers the error function. The adjustment formula of adaptive variable learning rate Back-propagation is as follow:

$$\eta(k+1) = \begin{cases} 1.05\eta(k) & SSE(k) \prec SSE(k-1) \\ 0.7\eta(k) & SSE(k) \succ 1.04 \cdot SSE(k-1) \\ \eta(k) & other \end{cases} \tag{24}$$

The selection of initial learning rate η(0) is very arbitrary.

Figure 7 and Figure 8 respectively show error convergence graph in the normal BP algorithm and improved BP algorithm when the network structure parameters are same.

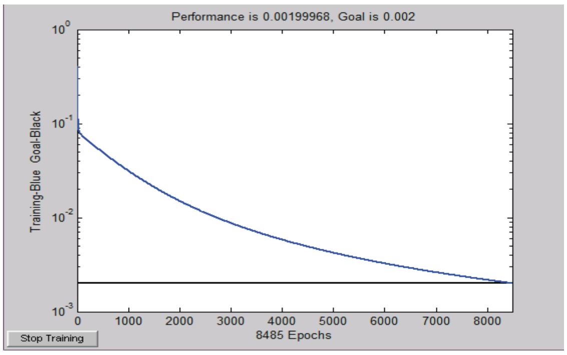


Figure 7. Error convergence graph of ordinary BP algorithm

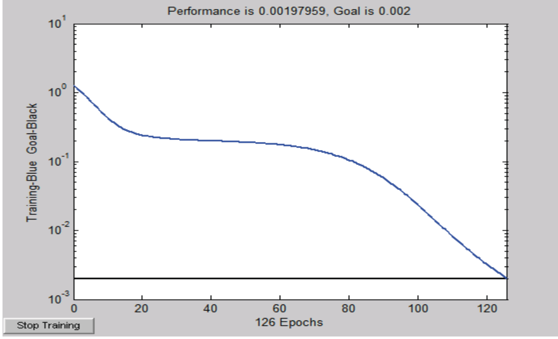


Figure 8. Error convergence graph of improved BP algorithm

It can be seen from Figure 7 and 8, the number of network training improved BP algorithm uses momentum-adaptive learning rate is much less than ordinary BP algorithm, it is the high diagnostic accuracy. It indicates that it is a very effective training method.

### B. The integration diagnosis of wavelet analysis and improved BP network

The magnetic image is resolved by wavelet packet analysis method, the wavelet energy eigenvalues are extracted when the corresponding wavelet coefficients are obtained. Wavelet energy features normalized are as neural network input sample, the network is trained by using improved BP neural network algorithm, and circuit board fault diagnosis is achieved. When the circuit board runs, it will inevitably happen one way or another failure, the effective and timely diagnosis of these main failures are effective means to ensure the normal operation of the board. Table 1 shows 5 common component failure modes of the circuit board.

Because of failure modes and failure causes have one to one relationship, that is, a magnetic image sample corresponds to a failed state. In the paper, the magnetic image features in the circuit board failure mode is as the network input, the corresponding failure cause is as the network output to train the neural network.

TABLE I. PCB FAILURE MODE AND FAILURE MECHANISM ANALYSIS

| fault symptoms | | Fault reason | | |
|---|---|---|---|---|
| **failure mode** | **fault sample** | **fault symbol** | **failure category name** | **failure category encoding** |
| x1 | I1.jpg | y1 | LM8361 fault | 00001 |
| x2 | I2.jpg | y2 | DS-3 U4 fault | 00010 |
| x3 | I3.jpg | y3 | DS-3 U5 fault | 00100 |
| x4 | I4.jpg | y4 | DS-3 U6 fault | 01000 |
| x5 | I5.jpg | y5 | DS-3 U7 fault | 10000 |

Because the pixels of image is too much, and there are many redundant information, magnetic image three layer wavelet energy feature information $X = (x_1, x_2, \cdots, x_9)$ is extracted is as input of neural network, fault causes $Y = (y_1, y_2, \cdots, y_{10})$ is as output, the failure mode is caused different failure is as the training sample and learning, it establishes the corresponding relationship between failure modes and the fault cause. Table 2 is the coefficients energy eigenvalues in different failure modes 3 layer wavelet decomposition of the board magnetic image.

The characteristic values in accordance with equation (21) were normalized. Its failure mode coding is as a neural network corresponding output sample to train the neural network, and the output sample indicates the relative between the input magnetic image information and the component failure. The output samples are shown in Table 3.

Finally, by using the same failure mode of the different magnetic images (non-fault training samples), during the corresponding extracting wavelet energy feature is as the neural network input, judging fault diagnosis, verifying the diagnose ability of network. Test results are shown in Table 4.

TABLE II. THE INPUT SAMPLE OF WAVELET ENERGY EIGENVALUES

| Sample | Level 1 | | | Level 2 | | | Level 3 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Horizontal | vertical | Diagonal | Horizontal | vertical | Diagonal | Horizontal | vertical | Diagonal |
| 1 | 0.2814 | 0.0384 | -0.0112 | 0.3507 | 0.0467 | -0.0016 | 0.2849 | 0.0031 | 0.0077 |
| 2 | 0.5094 | -0.2391 | -0.0350 | 0.6307 | -0.2114 | 0.0230 | 0.5435 | -0.2151 | -0.0058 |
| 3 | 0.4839 | -0.1187 | -0.0131 | 0.7172 | -0.1456 | -0.0034 | 0.3905 | -0.2957 | 0.0495 |
| 4 | 0.4835 | -0.0168 | 0.0060 | 0.3353 | -0.0107 | -0.0281 | 0.2352 | 0.0044 | -0.0086 |
| 5 | 0.1778 | 0.0419 | 0.0120 | 0.2476 | 0.2196 | -0.0127 | 0.2437 | 0.0894 | -0.0193 |

TABLE III. THE DESIRED OUTPUT CHARACTERISTICS OF NEURAL NETWORK

| Circuit State | y1 | y2 | y3 | y4 | y5 |
|---|---|---|---|---|---|
| Element 1 failure | 0 | 0 | 0 | 0 | 1 |
| Element 2 failure | 0 | 0 | 0 | 1 | 0 |
| Element 3 failure | 0 | 0 | 1 | 0 | 0 |
| Element 4 failure | 0 | 1 | 0 | 0 | 0 |
| Element 5 failure | 1 | 0 | 0 | 0 | 0 |

TABLE IV. TEST RESULTS

| Input data | y1 | y2 | y3 | y4 | y5 |
|---|---|---|---|---|---|
| Element 1 failure | 0.0035 | 0.0008 | 0.0072 | -0.0013 | 0.8998 |
| Element 2 failure | 0.0008 | -0.0016 | 0.0049 | 0.9360 | 0.0003 |
| Element 3 failure | -0.0022 | 0.0081 | 0.9221 | 0.0069 | 0.0053 |
| Element 4 failure | -0.0034 | 0.8982 | 0.0071 | -0.0025 | -0.0008 |
| Element 5 failure | 0.9243 | -0.0009 | -0.0046 | 0.0010 | 0.0106 |

Test results from Table 4 can be seen, the recognition results of the network is more accurate in given the different components failure mode, and its deviations with network desired outputs are very small, the SSE (sum of prescribing error)between analog output and the desired results are 0.1000. The identify effect is ideal. Network performance is stable and reliable, the network design is successful. The system also shows it is feasible and reliable method based on the magnetic image analyzing the circuit board fault diagnosis.

## VII. CONCLUSION

In the circuit board fault diagnosis process, the magnetic detection technology is introduced. The gray, median filtering, gray-scale transformation, histogram equalization and related image processing techniques of the magnetic image are studied when the magnetic image is as the test data. To extract the wavelet energy feature of the magnetic image to determine whether it is a failure occurs in pre-test board or not, the improved BP algorithm is melted into the magnetic image fault diagnosis process, which has an advantage of the weight adjustment smooth and a short training time, so the speed is fast, low cost, high accuracy of fault diagnosis. It can play a larger role, it accords with rapid and quick, intelligent direction trend of the fault diagnosis to the circuit board system debugging, monitoring and diagnosis.

## ACKNOWLEDGMENT

This work is sponsored by the Natural Science foundation of Shanxi, project No. 2010011031-1 and Shanxi Scholarship Council of China. The authors would like to thank all members of the research team.

## REFERENCES

[1] Jinyan Ca, Chunhui Han, Yafeng Meng. Analog Circuit Testability for Fault Diagnosis[J]. Tsinghua Science&Technology. 2007,7.112: 270-274.

[2] J. Sijbers, P. Scheunders, N. Bonnet,D. Van Dyck, E. Raman. Quantification and improvement of the signal-to-noise ratio in a magnetic resonance image acquisition procedure[J]. Magnetic Resonance Imaging, 14(10),1996: 1157-1163.

[3] Khalaf Salloum Gaeid and Hew Wooi Ping, Wavelet fault diagnosis and tolerant of induction motor:A review, International Journal of the Physical Sciences Vol. 6(3), pp. 358-376, 4 February, 2011

[4] Arthur Hugues Ball, Thermal and Electrical Considerations for the Design of Highly-Integrated Point-of-Load Converters, faculty of the Virginia Polytechnic Institute and State University[D],November 18, 2008

[5] Sami Ekici, Selcuk Yildirim, Mustafa Poyraz. Energy and entropy-based feature extraction for locating fault on transmission lines by using neural network and wavelet packet decomposition[J]. Expert Systems with Applications, 2008, 34(4): 2937-2944.

[6] Z.K.Peng, M.R.Jackson, J.A.Rongong, F.L.Chu, R.M.Parkin. On the energy leakage of discrete wavelet transform[J] . Mechanical Systems and Signal Processing. 2009, 23(2): 330-343.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# USE OF CONFIDENCE LIMITS IN THE SETTING OF ON-BOARD DIAGNOSTIC THRESHOLDS

Paul J King
Powertrain Control Systems and Calibration
Jaguar Land Rover Limited
Coventry,
CV3 4BJ, U.K

Keith J Burnham
Control Theory and Applications Centre, Coventry University,
Coventry,
CV1 5FB, U.K.

*Abstract*—**This Vehicles sold in the US and Europe have to be equipped with a Diagnostics, called On-Board Diagnostics (OBD), which monitor the performance of various elements of the emission control system. The driver is informed as to any failures by the use of a Check Engine Light on the dashboard of the vehicle and then should return the vehicle to the dealership for rectification. The vehicle manufacturer's aim is to ensure that the Check Engine Light is only illuminated for legitimate failures. For the calibration of an On Board Diagnostics there needs to be sufficient separation between the response of a good sensor and a failed sensor, the setting of this threshold should be based upon a statistical model of the data so that the predicted failures rate can then be determined. By applying confidence limits to the models allows the engineer to understand how additional data points will effect the calculation of the failure threshold. This gives the engineer the ability to determine the tradeoff between the number of data points and a confidence in the estimated statistical model.**

***Keyword-components:*** ***Automobiles, Detection, Diagnosis, Engines***

## I. INTRODUCTION

Vehicles sold throughout the world are subject to an increasingly stringent set of emission thresholds. To achieve certification, all sensors and vehicle sub-systems that may affect vehicle exhaust emissions have to be monitored by an On-Board Diagnostic (OBD) system that is part of the Engine Management System (EMS) or any other embedded controller [1]. This requirement was first introduced in the US in 1988 for OBD1, for open and short circuit faults, and in 1994 for OBD2, for changes in sensor and actuator responses [2]. For Europe this legislation, denoted EOBD, has been introduced for all vehicles built after January 2000 [3]. Both sets of legislation link the performance of the different diagnostics to emission thresholds. In the event of component or sub-system failure, a ‘check engine’ light must be illuminated as an indication to a driver that there is a problem, so corrective action can be taken to minimise the pollution caused by such a fault. As the emissions thresholds are continually reduced, more sophisticated techniques are required to be employed to meet these increasingly tightening thresholds.

As the Vehicle Emission Legislation drives down vehicle pollution the impact on the diagnostics is that they have had to become increasing more complex to determine a failed system. As a consequence the diagnostics are becoming have become increasing more like models of the functions that they are monitoring, if only over a restricted set of operating conditions, and as such the diagnostic results that they produce are becoming more likely to fit to a Gaussian Distribution.

OBD Diagnostic Calibration engineers develop test plans that invoke the worse conditions for the diagnostics, by introduction of variety of test conditions. These are typically different fuel specifications, operations at different ambient conditions (hot, cold and altitude), with different driving styles and tolerance sensors. Each diagnostic will have it’s own set of worse case test conditions which have been developed through the experience of the engineer and lessons learnt. The approach used by the calibration engineer is to collect data for these conditions and then fit a Gaussian distribution to the data to set thresholds to ensure that 'normal' systems does not false flags and that 'failed' system flag in a timely manner.

By making use of these Gaussian models and through the information obtained by the use of confidence interval of these models allows a more conservative and robust threshold to be set. Since the variation in the models reduces as the more data is collected they also provide the Engineer with a way of gauging whether collected any more data will significantly change their results. This is especially useful when collecting Fault condition data which where it is difficult to generate a large amount of data either because it requires specialist hardware setup or a specific environmental condition for which there might be a limited amount of testing time available.

The paper is organised as follows: Section II Problem Formulation which outlines the diagnostic used within this paper, Section III defines the Diagnostic Specification which outlines the set of conditions for setting the diagnostic thresholds, Section IV which defines the calculation of the Confidence Interval, Section V Process Monitoring and

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

Analysis details the implementation and analysis and finally, Section VI details further work and concludes the paper.

## II. Problem Formulation

On modern engines the air fuel ratio (AFR) has to be tightly controlled so that the Three Way Catalyst is operating at it's optimum and providing the correct emissions control [4]. AFR is measured as the ratio between the mass of air and the mass of fuel for pure octane the stoichiometric mixture is approximately 14.7:1. The exact composition of fuels varies seasonally and geographically so modern engines use a more convenient measure of lambda (λ) rather than AFR to allow them to control combustion process. Lambda (λ) is defined as the ratio of measured AFR to stoichiometric AFR for that given mixture. A Lambda of 1.0 is stoichiometry, rich mixtures are less than 1.0, and lean mixtures are greater than 1.0.

In this paper the calibration of an Individual Cylinder Air Fuel Ratio Diagnostic is investigated. This is a diagnostic which was first introduced for vehicles sold as 2010 Model Year vehicles [5,6]. Since this is still a relatively new diagnostic it was decided to carry out a more detailed analysis of the diagnostic results. The data analysed in this paper was collected from a cold ambient environment trip, this condition being deemed as being the worse case condition potentially giving the smallest separation between a fault free and the failed set of data.

Prior to the test trip the failure condition for both a rich and lean shift in λ for individual cylinders was determined by carrying out tests over an Emissions Drive Cycle. A failure condition being determined when the fuelling shift resulted in an emissions test result being 1.5 times the certify emissions standard. The failure threshold is set as the amount of shift in λ. The diagnostic infers the amount of Fuelling shift that each cylinder is experiencing from λ = 1 to determine a failure.

## III. Diagnostic Threshold Setting

Ideally the failure threshold should be set so you can capture all of the diagnostic results. However, in practice this may not always be possible so a minimum target of 90% of the data which contains a fault condition should result in the diagnostic bring on the Check Engine Light. To put on the Check Engine Light in USA requires that a failure is detected on two successive diagnostic operations. So to achieve the 90% detection on two successive tests requires that for a single test the failure threshold should be set at a level approximately 95% (100√0.9). To determine the Failure Thresholds in Section 4 it is more convenient to convert this figure into a Standard Deviation. In the case we are assuming that the threshold will only occur on one side of the distribution closest to the Fault Free set of data. On the Failure Threshold side of the distribution the point at which 45% of the population is represented by a Standard Deviation of 1.64, the other 50% of the data being containing the other half of the distribution.

For the fault free set of data we need to ensure that the fault thresholds are greater than 3×Standard Deviations from either side of the distribution we will refer to this as the Rich or Lean Robustness Threshold. This will then allow 99.74% of the data to be correctly identified as being fault free for a single diagnostic result. For the two successive diagnostic results this would then lead to the possibility of flagging a fault free system as having a fault as being 1 in 148,000 tests.

## IV. Problem Formulation

The set of data collected from the diagnostic cannot precisely define the characteristics of the population. The sample can only define a range of values for both the probable mean position and the probable standard deviation value. Confidence interval calculations are used to define a probable range of values for the population mean $\bar{x}_U$ (Upper) and $\bar{x}_L$ (Lower) and the population standard deviation $\sigma_U$ and $\sigma_L$. This then generates a range of possible statistical models that will include the population model with a given confidence.

Equation (1) [7] is used to Calculation of the Confidence Interval for the Mean

$$\mu = \bar{x}_S \pm t_{\beta,n-1}\frac{\sigma_S}{\sqrt{n}} \tag{1}$$

$\bar{x}_S$ is the sample mean $t$ is the Confidence Factor based upon the Student's t-Distribution, $\sigma_S$ is the Sample Standard Deviation, $n$ is the Sample Size and $\beta$ is either $\alpha$ for a single sided confidence Interval or $\alpha/2$ for two sided confidence interval. Where $\alpha$ is the significance level and for this paper a value of 5% or 0.05 will be used.

Calculation for the Confidence Limit for Standard Deviation is given by

Lower Limit    Upper Limit

$$\sigma_S\sqrt{\frac{n-1}{\chi^2_{\beta,n-1}}} \le \sigma \le \sigma_S\sqrt{\frac{n-1}{\chi^2_{1-\beta,n-1}}} \tag{2}$$

$\chi^2$ is the Confidence Factor based upon the Chi-Squared Distribution.

The results of equation (1) and (2) will produce a range of standard deviation and means values which will include the population model with a 95% confidence. This confidence increases and the range of these values reduces as there is an increase in the amount of data, $n$, as it is collected. Using this range of values it is then possible to choose a combination which will give the worse case Failure or Robustness thresholds. In terms of the Standard Deviation it is the Lower

Limit calculation in (2), $\sigma_L$ , which produces the maximum value of sigma and this will generate a model which produces a greater range of λ values. The means, $\bar{x}_L$ or $\bar{x}_U$ , are chosen so as to give the smallest separation between the Failure threshold and Fault free condition in each case.

## V. Process Monitoring and Analysis

From testing three specific sets of data were collected a the Fault Free, represented by under score F, a set of data in which a rich shift has been introduced to represent the emission failure threshold, represented by an underscore R, and a set of data for the lean shift, represented by an underscore L. This resulted in 6294 sets of results with a normal engine, 102 diagnostic results for the rich shift 106 results for the lean shift.

The first thing to check is that all of the data sets have a Gaussian distribution. This was done using the LilleTest function within MATLAB which performs a Lilliefors test [8, 9] of the default null hypothesis that the samples comes from a Gaussian distribution, against the alternative that it does not come from a Gaussian distribution. The test returns a 1 if it rejects the null hypothesis at the 5% significance level. For the fault free condition the set of data failed this test to future investigate this data is plotted on a Quartile-Quartile (QQ) Plot. This shows the raw data as blue crosses plotted against an ideal Gaussian model which is the red dashed line. This shows that for the bulk of the data between the 5% and 95% percentiles it fits a Gaussian distribution. It is only the tail information of the distribution which does not match this statistical model. From the shape of the tails in Figure 1 it indicates that the data has come from a ‘fat tailed distribution’ where the data extends further than for the tails of a Gaussian distribution. This can be seen by considering the 0.999 and the 0.001 percentile points which lie at points 0.925 and 1.07 respectively and should, for a Gaussian Distribution, lie at 0.94 and 1.06. Even though this statistical model does not accurately fit the data it was decided to assume that the fault free set of data could be considered to be Gaussian initially and then reviewed when the final thresholds for the Rich and Lean have been determined. The risk is by making this assumption that the variance used for the Fault Free Data will under estimate the true risk.

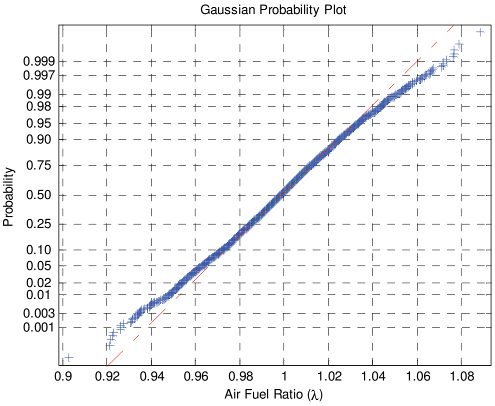


Figure 1: QQ plot for the Fault Free Data

| | $L_F$ | $S_F$ | $U_F$ |
|---|---|---|---|
| $\bar{x}$ | 0.9980 | 0.9985 | 0.9990 |
| Σ | 0.0217 | 0.0213 | 0.0209 |

Table 1: Statistical Model for Fault Free Data ($\beta$ = 0.025)

Table 1 has been calculated for $\beta = \alpha/2 = 0.025$ in Equations (1) and (2). The Standard Deviation $\sigma_{L_F}$ , from Table 1, gives the largest value of sigma and is used to calculate the Robustness Threshold for both the rich and lean sides of the distribution. For the Lean Robustness Threshold the mean $\bar{x}_{U_F}$ was used to give a value of 1.0640, the Rich Robustness Threshold was obtain using $\bar{x}_{L_F}$ which gives a value of 0.9330.

Table 2 shows the Rich Failure model information note that since we are considering a single side threshold then $\beta$ = 0.05. Figure 2 shows the Histogram of the raw Rich Failure data and the red solid line shows the worse case statistical model using $\bar{x}_{U_R}$ and $\sigma_{LR}$ . Using this model the Rich Failure threshold is calculated as 0.8579 and is shown on the graph by the solid vertical red line. The dotted blue vertical line shows the sample Rich Failure threshold of 0.8507 calculated by making use of $\sigma_{SR}$ and $\bar{x}_{SR}$ .

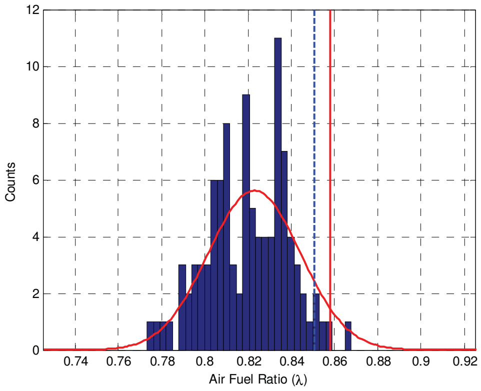


Figure 2: Statistical Results for Rich Failure Data

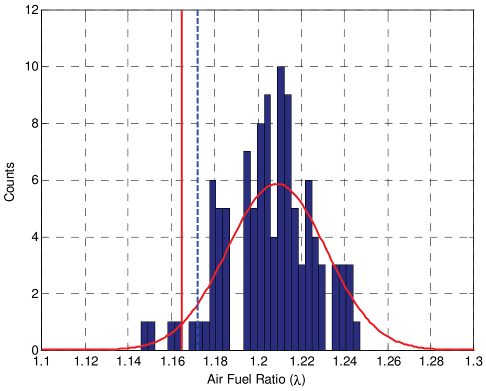


Figure 3: Statistical Results for Lean Failure Data

| | $L_R$ | $S_R$ | $U_R$ |
|---|---|---|---|
| $\bar{x}$ | 0.8166 | 0.8197 | 0.8229 |
| Σ | 0.0214 | 0.0189 | 0.0170 |

Table 2: Statistical Model for Rich Failure Data ($\beta$ = 0.05)

| | $L_L$ | $S_L$ | $U_L$ |
|---|---|---|---|
| $\bar{x}$ | 1.2022 | 1.2055 | 1.2087 |
| Σ | 0.0229 | 0.0203 | 0.0182 |

Table 3: Statistical Model for Lean Failure Data ($\beta$ = 0.05)

Since we most concerned with the issue of falsely flagging a Fault Free System it is useful to take the difference between the Rich Failure Threshold at 0.8579 and the Rich Robustness Threshold at 0.9330 and then determine the amount of sigma separation there is between them. This difference is normalised by dividing by $\sigma_{L_F}$ to determine the amount of separation in terms of sigma and results in a separation of 3.46 $\sigma_{L_F}$

This then provides a clear robust threshold in terms of provide a detection which meets the requirements and provides a significant safety margin against falsely flagging and we have also taken into consideration the variability in the models by making use of the confidence interval

Figure 3 shows the response after introducing a lean failure shift again the graph shows the histogram of the raw data and the solid red line shows the distribution of the worse case statistic model derived from $\bar{x}_{L_L}$ and $\sigma_{L_L}$ in Table 3. The red vertical line shows the Lean Failure Threshold of 1.1647 from this model and the dotted blue vertical line at 1.1723 for derived from $\sigma_{S\,L}$ and $\bar{x}_{S\,L}$. Using the same metric as previously derived the differences between the Lean Failure Threshold and the Lean Robustness Threshold gives a separation of 4.64 $\sigma_{L_F}$ .

From the discussion earlier in the paper it has been highlighted that it can difficult to collect Fault condition so it is useful to be able to assess whether enough data has been collected to ensure that the real distribution of the Fault Condition has been captured. To determine this we have to make the assumption that the Sample model information that we have is the best estimate of the distribution model. So in the Rich Failure case the Sample values of $\sigma_{S\,R}$ and $\bar{x}_{S\,R}$ from Table 2 have been used and the value of n and the subsequent changes made to the t-distribution and Chi-Squared, in equations (1) and (2), can then be used to determine the effect on the threshold calculation.

In Figure 4 the blue solid line shows how the difference between the Rich Failure Threshold and the Nominal Rich Failure Threshold, normalised against the value of $\sigma_{S\,R}$ determined at n = 102, decays as the amount of data increases. The Red dot shows the current point Rich Failure Threshold Delta. In Figure 5 shows the same metric as in Figure 4 but for the Lean Failure Condition.

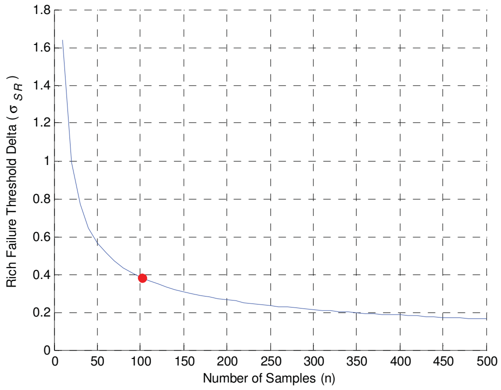


Figure 4: Rich Failure Threshold Delta

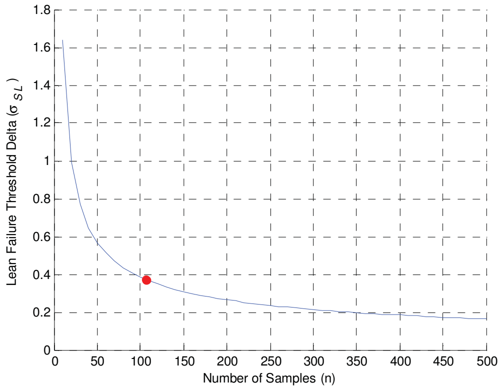


Figure 5: Lean Failure Threshold Delta

From the results shown in Figure 4 and 5 the current set of data obtained at around 100 samples gives an adequate confidence and subsequent testing would not necessarily lead to a significant change in the values of the Failure Thresholds. For example to reduce the current threshold, in both cases, by 50% would require an addition 250 tests.

| | Robust Threshold | Failure Threshold | Separation |
|---|---|---|---|
| Rich | 0.9330 | 0.8579 | $3.46\,\sigma_{L_F}$ |
| Lean | 1.0640 | 1.1647 | $4.64\,\sigma_{L_F}$ . |

Table 4: Summary of the Threshold Information

## VI. Conclusion

The results in Table 4 show the calculated Failure Thresholds and the Robustness Thresholds for Rich and Lean conditions and for this diagnostic there is a significant amount of separation between these two sets of thresholds has enabling the requirements laid out in Section 3 to be met. This amount of separation has reduced the risk of making use of a Gaussian Model to derive the information for the Fault Free set of data. If it were the case that there was not such a separation then further investigation would have to undertaken to understand the reason or determine perhaps a more valid statistical model.

The use of the Confidence Interval in the statistical models and the threshold setting enables the calibrators to determine when they have collected enough data to provide a robust calibration.

## Acknowledgement

To Brian Varney and James Anderson for the inspiration and the data used in this work

## References

[1] P Baltusis. On Board Vehicle Diagnostics, SAE Technical Paper 2004-21-0009, 2004.

[2] OBD-II Title 13, California code regulations, Section 1968

[3] EOBD DIRECTIVE 98/69/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 13 October 1998

[4] R. Stone, Introduction to the Internal Combusiton Engine, Society of Automotive Engineers, U.S., 3rd Revised edition edition, ISBN 978-0-7680-0495-3

[5] J. Fantini, J.F. Burq, Exhaust-Intake Manifold Model for Estimation of Individual Cylinder Air Fuel Ratio and Diagnostic of Sensor-Injector, Electronic Engine Controls 2003 (SP-1749), 2003-01-1059

[6] J.C. Smith, C.W. Schulte, D.D. Cabush, Individual Cylinder Fuel Control for Imbalance Diagnosis, SAE Int. J. Engines, Volume 3 Issue 1, 2010-01-0157

[7] Confidence Interval Presentation - Black Belt Teaching Material, 6 Sigma Academy, 2003

[8] H. W. Lilliefors. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. Journal of the American Statistical Association. Vol. 62, pp. 399–402, 1967.

[9] H. W. Lilliefors. On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown. Journal of the American Statistical Association. Vol. 64, pp. 387–389, 1969

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Discrete Flight Path Angle Tracking Control of Hypersonic Flight Vehicles via Multi-rate Sampling

Bin Xu[*][†], Danwei Wang[†] Chenguang Yang[§], Jing Li[¶] and Shixing Wang[‡]
[*]School of Automation, Northwestern Polytechnical University, Xi'an, China
[†]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[‡]Department of Computer Science, Tsinghua University, Beijing, China
[§]School of Computing and Mathematics, Plymouth University, UK
[¶]Department of Mathematics, Xidian University, China

*Abstract*—**This paper presents the flight path angle tracking control of the longitudinal dynamics of a generic hypersonic flight vehicle(HFV). Due to the use of digital computers and microprocessors for controls applications, the discrete hypersonic flight control is investigated. The altitude command is transformed into the flight path angle information. The back-stepping scheme is applied for the attitude subsystem which includes flight path angle, pitch angle and pitch rate. The virtual control is designed with nominal feedback and Neural Network (NN) approximation. To use the information of throttle setting, the multi-rate sampling method is employed for the two subsystems where the velocity subsystem is considered as slow dynamics. Under the proposed controller, the semiglobal uniform ultimate boundedness (SGUUB) stability is guaranteed. The simulation is presented to show the effectiveness of the proposed control approach.**

## I. Introduction

Given the widespread use of digital computers and microprocessors for controls applications, the discrete-time case is certainly warranted [1]. For the control of flight vehicle and spacecraft, controller on the basis of continuous system is usually implemented by a digital computer with a certain sampling interval [2], [3]. Since modern aircraft are equipped with digital computers, the controller should be designed in discrete-time form [4]. There are two methods for designing the digital controller. One method, called emulation, designs a controller with the continuous-time system, and then discretizing the controller. The other is to design the controllers directly based on the discrete system. In contrast to the emulation method, the discrete controller is designed in a discrete domain so that the performance of the controller may not depend on the sampling rate and the upper bounds of the NN weight update rates guaranteeing the convergence can be estimated analytically while emulation method is otherwise [5].

In this paper, the discrete hypersonic flight control is analyzed. Hypersonic Flight vehicles are intended to present a reliable and more cost efficient way to access space with dramatic reductions in flight times. The longitudinal model of the dynamics is known to be unstable, non-minimum phase with respect to the regulated output, and affected by significant model uncertainty. In the last decade, considerable research focused on robust and adaptive hypersonic flight control [6], [7]. In [8], the control structure combines the inputs from the pilot model, baseline controller and adaptive controller. Based on the input-output linearization using Lie derivative notation, sliding mode control [9] is designed. The sequential loop closure controller design [10] is based on the decomposition of the equations into functional subsystems. The method followed the approach that combined robust adaptive dynamic inversion with back-stepping arguments to obtain control architecture. In [11], one high gain observer based controller is proposed for HFV control with only one NN to compensate the lumped uncertainty. In [12], the attitude states are considered as fast dynamics and the altitude control is transformed into the flight path angle tracking.

It is illustrated that sometimes the controller based on Euler approximate discrete-time model of the plant is superior to back-stepping controllers based on digital control based on continuous-time plant model [13]. In this way, the discrete HFV model is obtained and the dynamic inversion is applied in [14]. By proper assumptions the discrete model is transformed into the strict-feedback form where some theoretical results have been studied in [15]. For the nonlinearity and the coupling, the nominal part of the dynamics should be considered for the feedback control design to provide good performance. The back-stepping neural design is analyzed in [16], [17]. By considering the uncertainty with Gaussian distribution, the Kriging design [18] is studied. In this paper, we focused on the control of the altitude subsystem. Continued with the work in [16], the altitude tracking is done with flight path angle tracking. The controller is designed with nominal value of the control gain. In this way, there is no need to know the upper bound. By multi-rate sampling design, the throttle setting can be viewed as constant during the attitude subsystem controller design. In this way, the assumption of $Tsin\alpha$ in [16] is eliminated.

This paper is organized as follows. Section II describes the longitudinal dynamics of a generic hypersonic flight vehicle. The strict-feedback form is formulated and the discrete analysis model is obtained in Section III. The brief description of HONN is explained in Section IV. Section V presents the adaptive controller design. The simulation result is included in Section VI. Section VII presents several comments and final remarks.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

## II. Hypersonic Aircraft Model

The control-oriented model of the longitudinal dynamics of a generic hypersonic aircraft is considered. This model is comprised of five state variables $X = [V, h, \alpha, \gamma, q]^T$ and two control inputs $U_c = [\delta_e, \beta]^T$ where $V$ is the velocity, $\gamma$ is the flight path angle, $h$ is the altitude, $\alpha$ is the attack angle, $q$ is the pitch rate, $\delta_e$ is elevator deflection and $\beta$ is the throttle setting.

$$\begin{aligned}
\dot{V} &= \frac{T\cos\alpha - D}{m} - \frac{\mu\sin\gamma}{r^2} && (1)\\
\dot{h} &= V\sin\gamma && (2)\\
\dot{\gamma} &= \frac{L + T\sin\alpha}{mV} - \frac{(\mu - V^2 r)\cos\gamma}{Vr^2} && (3)\\
\dot{\alpha} &= q - \dot{\gamma} && (4)\\
\dot{q} &= \frac{M_{yy}}{I_{yy}} && (5)
\end{aligned}$$

where $T$, $D$, $L$ and $M_{yy}$ represent thrust, drag, lift-force and pitching moment respectively, $m$, $I_{yy}$ and $\mu$ represent the mass of aircraft, moment of inertia about pitch axis and gravity constant. $r$ is the radial distance from center of the earth and $r = h + R_E$. The related definition can be found in [9].

This paper focused on the cruise control with no consideration of the reentry process. The main goal of this paper is to design the altitude and velocity controller separately to follow the tracking reference $h_d$ and $V_d$.

## III. System Transformation

### A. Strict-Feedback Formulation

*Assumption 1:* Since $\gamma$ is quite small, we take $\sin\gamma \approx \gamma$ in (2) for simplification.

*Remark 1:* Similar assumption is made in [10] where the value of the flight path angle is set to be inside $[-3°, 3°]$.

The velocity subsystem (1) can be rewritten as

$$\begin{aligned}
\dot{V} &= f_V + g_V u_V\\
u_V &= \beta\\
y_V &= V
\end{aligned} \tag{6}$$

where $f_V = -(D/m + \mu\sin\gamma/r^2) + \bar{q}S \times 0.0224\cos\alpha/m$, $g_V = \bar{q}S \times 0.00336\cos\alpha/m$ if $\beta > 1$. Otherwise $f_V = -(D/m + \mu\sin\gamma/r^2)$, $g_V = \bar{q}S \times 0.02576\cos\alpha/m$.

The tracking error of the altitude is defined as $\tilde{h} = h - h_d$ and the flight path command is chosen as

$$\gamma_d = \arcsin\left[\frac{-k_h\tilde{h} - k_I\int\tilde{h}dt + \dot{h}_d}{V}\right] \tag{7}$$

if $k_h > 0$ and $k_I > 0$ are chosen and the flight-path angle is controlled to follow the reference command $\gamma_d$, the altitude tracking error is regulated to zero exponentially.

Define $X_A = [x_1, x_2, x_3]^T, x_1 = \gamma, x_2 = \theta_p, x_3 = q$ where $\theta_p = \alpha + \gamma$.

Then the strict-feedback form equations of the attitude subsystem (3)-(5) are written as

$$\begin{aligned}
\dot{x}_1 &= f_1(x_1) + g_1(x_1)x_2\\
\dot{x}_2 &= f_2(x_1, x_2) + g_2(x_1, x_2)x_3\\
\dot{x}_3 &= f_3(x_1, x_2, x_3) + g_3(x_1, x_2, x_3)u_A\\
u_A &= \delta_e\\
y &= x_1
\end{aligned} \tag{8}$$

where $f_1 = -(\mu - V^2 r)\cos\gamma/(Vr^2) - \bar{q}S \times 0.6203/(mV) \times \gamma - T sin\alpha$, $g_1 = \bar{q}S \times 0.6203/(mV)$, $f_2 = 0$, $g_2 = 1$, $f_3 = \bar{q}S\bar{c}[C_M(\alpha) + C_M(q) - 0.0292\alpha]/I_{yy}$, $g_3 = 0.0292\bar{q}S\bar{c}/I_{yy}$.

In the analysis [11], the altitude is mainly up to elevator deflection while the velocity is controlled by throttle setting. By command transformation, we know attitude subsystem is controlled by elevator deflection. To obtain the information of $\beta$ from the velocity subsystem, the multi-rate sampling design is employed so that velocity and $\beta$ can be considered as constant during the altitude subsystem controller design. The similar idea is studied in [19] where the airspeed, altitude and flight path angle are selected as slow-dynamics variables and considered invariant during the controller design of the fast dynamics.

The control objective of system (8) is to design an adaptive controller, which makes $\gamma \to \gamma_d$, further $h \to h_d$ and all the signals involved are bounded.

*Assumption 2:* $f_i$ and $g_i$ are unknown smooth functions which can be decomposed into the nominal part $f_{iN}$, $g_{iN}$ and the unknown part $\Delta f_i$, $\Delta g_i$, $i = 1, 3, V$.

### B. Discrete-time Model

By Euler approximation with different sample time period $T_V$ and $T_s$, systems (6) and (8) can be approximated as

$$V(k+1) = V(k) + T_V[f_V(k) + g_V(k)u_V(k)] \tag{9}$$

$$\begin{aligned}
x_1(k+1) &= x_1(k) + T_s[f_1(k) + g_1(k)x_2(k)]\\
x_2(k+1) &= x_2(k) + T_s[f_2(k) + g_2(k)x_3(k)]\\
x_3(k+1) &= x_3(k) + T_s[f_3(k) + g_3(k)u_A(k)]
\end{aligned} \tag{10}$$

Here $T_V$ is selected four times as $T_s$. So during the design of attitude subsystem, the velocity and throttle setting could be considered as constant.

## IV. HONN Approximation

Higher order neural network (HONN) is one kind of linearly parameterized NNs. The structure of HONN is expressed as follows:

$$U(\omega, X) = \omega^T\theta(X) \quad \omega, \theta(X) \in R^N \tag{11}$$

$$\theta_i(X) = \prod_{j\in I_i}[s(X_j)]^{d_{ji}} \tag{12}$$

where $X \subset R^m$ is the input to HONN, $N$ is the NN nodes number, $\{I_1, I_2, ..., I_N\}$ is a collection of $N$ not-ordered subsets of $\{1, 2, ..., m\}$, specified by the designer, $d_{ji}$'s are prescribed nonnegative integers, $\omega$ is an adjustable synaptic weight vector, and $s(X_j)$ is a monotonically increasing and differentiable

sigmoidal function. In this paper, it is chosen as a hyperbolic tangent function, i.e.,$s(X_j) = (e^{X_j} - e^{-X_j})/(e^{X_j} + e^{-X_j})$.
For a desired function $U^*$, it is assumed there exists an ideal weight vector $\omega^*$ such that the smooth function vector can be approximated by an ideal NN on a compact set

$$U^* = \omega^{*T}\theta(X) + \varepsilon(X), \|\varepsilon(X)\| < \varepsilon_M \tag{13}$$

where $\varepsilon(X)$ is the bounded NN approximation error vector and $\varepsilon_M$ is the supreme of $\varepsilon(X)$.

## V. Discrete Control Design

### A. Adaptive NN Control for Attitude Subsystem

The errors are defined as

$$\begin{aligned} z_1(k) &= x_1(k) - x_{1d}(k) && (14)\\ z_2(k) &= x_2(k) - x_{2d}(k) && (15)\\ z_3(k) &= x_3(k) - x_{3d}(k) && (16) \end{aligned}$$

where $x_{2d}(k)$, $x_{3d}(k)$ are the virtual control inputs to be designed.
***Step 1***. From (14),

$$z_1(k+1) = x_1(k) + T_s[f_1(k) + g_1(k)x_2(k)] - x_{1d}(k+1) \tag{17}$$

where $x_{1d}(k+1)$ is acquired from (7).

Since $g_1(k)$ and $f_1(k)$ are unknown, the uncertainty is defined as

$$\begin{aligned} U_1(k) &= -\frac{1}{T_s g_1(k)}[-T_s f_1(k) - x_1(k) + x_{1d}(k+1)]\\ &+ \frac{1}{T_s g_{1N}(k)}[-T_s f_{1N}(k) - x_1(k) + x_{1d}(k)]\\ &= \omega_1^{*T}\theta_1(X_1(k)) + \varepsilon_1(X_1(k)) && (18) \end{aligned}$$

where $X_1(k) = [V(k), x_1(k), h(k), h_d(k), h_d(k+1)]^T$, $f_{1N}(k)$ and $g_{1N}(k)$ are the nominal parts of $f_1(k)$ and $g_1(k)$, $\omega_1^*$ is the optimal parameters for NN to approximate $U_1(k)$ and $\varepsilon_1(X_1)$ is the NN reconstruction error.
Take $x_2(k)$ in (17) as the virtual control input and design its desired value as

$$\begin{aligned} x_{2d}(k) &= \frac{1}{T_s g_{1N}(k)}[c_1 z_1(k) - T_s f_{1N}(k) - x_1(k) + x_{1d}(k)]\\ &+ \hat{\omega}_1^T(k)\theta_1(X_1(k)) && (19) \end{aligned}$$

where $\hat{\omega}_1$ is the estimation of $\omega_1^*$.

Combining (15), (17) and (19), the following equation can be obtained.

$$\begin{aligned} z_1(k+1) &= x_1(k) + T_s[f_1(k) + g_1(k)x_2(k)] - x_{1d}(k+1)\\ &= T_s g_1(k) z_2(k) + T_s g_1(k)[\tilde{\omega}_1^T(k)\theta_1(X_1(k)) - \varepsilon_1(X_1(k))]\\ &+ \frac{g_1(k)}{g_{1N}(k)} c_1 z_1(k) \end{aligned} \tag{20}$$

where $\tilde{\omega}_1(k) = \hat{\omega}_1(k) - \omega_1^*$. The robust updating algorithm for the NN weights is

$$\hat{\omega}_1(k+1) = \hat{\omega}_1(k) - \lambda_1 z_1(k+1)\theta_1(X_1(k)) - \delta_1\hat{\omega}_1(k) \tag{21}$$

where $\lambda_1 > 0$ and $0 < \delta_1 < 1$.

***Step 2***. From(15),

$$z_2(k+1) = x_2(k) + T_s[f_2(k) + g_2(k)x_3(k)] - x_{2d}(k+1) \tag{22}$$

Define $X_2(k) = [X_1^T(k), x_2(k), h_d(k+2)]^T$. From (19), $x_{2d}(k+1)$ involves $x_1(k+1)$, $f_1(k+1)$ , $z_1(k+1)$ and $x_{1d}(k+2)$. It can be concluded that $x_{2d}(k+1)$ is the function of $X_2(k)$ . The uncertainty $U_2(k)$ is defined and can be approximated by NN as

$$U_2(k) = x_{2d}(k) - x_{2d}(k+1) = \omega_2^{*T}\theta_2(X_2(k)) + \varepsilon_2(X_2(k)) \tag{23}$$

where $\omega_2^*$ is the optimal parameters and $\varepsilon_2(X_2(k))$ is the NN reconstruction error.

Take $x_3(k)$ in (22) as the virtual control input and design its desired value as

$$x_{3d}(k) = \frac{1}{T_s g_2(k)}[-x_2(k) + x_{2d}(k) + c_2 z_2(k)] + \hat{\omega}_2^T(k)\theta_2(X_2) \tag{24}$$

where $\hat{\omega}_2$ is the estimation of $\omega_2^*$. The result of (23) and (24) is due to the fact that $f_2 = 0$ and $g_2 = 1$.

The robust updating algorithm for the NN weights is

$$\hat{\omega}_2(k+1) = \hat{\omega}_2(k) - \lambda_2 z_2(k+1)\theta_2(X_2(k)) - \delta_2\hat{\omega}_2(k) \tag{25}$$

where $\lambda_2 > 0$ and $0 < \delta_2 < 1$.

***Step 3***. From (16),

$$z_3(k+1) = x_3(k) + T_s[f_3(k) + g_3(k)u_A(k)] - x_{3d}(k+1) \tag{26}$$

Define $X_3(k) = \left[X_2^T(k), x_3(k), h_d(k+3)\right]^T$. Similarly we can deduce that the uncertainty $U_3(k)$ is the function of $X_3(k)$ and it can be approximated by NN as

$$\begin{aligned} U_3(k) &= -\frac{1}{T_s g_3(k)}[-T_s f_3(k) - x_3(k) + x_{3d}(k+1)]\\ &+ \frac{1}{T_s g_{3N}(k)}[-T_s f_{3N}(k) - x_3(k) + x_{3d}(k)]\\ &= \omega_3^{*T}\theta_3(X_3(k)) + \varepsilon_3(X_3(k)) && (27) \end{aligned}$$

where $\omega_3^*$ is the optimal parameters for NN to approximate $U_3(k)$ and $\varepsilon_3(X_3(k))$ is the NN reconstruction error. The actual control input is designed as

$$\begin{aligned} u_A(k) &= \frac{1}{T_s g_{3N}(k)}[-T_s f_{3N}(k) - x_3(k) + x_{3d}(k) + c_3 z_3(k)]\\ &+ \hat{\omega}_3^T(k)\theta_3(X_3(k)) && (28) \end{aligned}$$

where $\hat{\omega}_3$ is the estimation of $\omega_3^*$. where $\tilde{\omega}_3(k) = \hat{\omega}_3(k) - \omega_3^*$. The update law for the NN weights is

$$\hat{\omega}_3(k+1) = \hat{\omega}_3(k) - \lambda_3 z_3(k+1)\theta_3(X_3(k)) - \delta_3\hat{\omega}_3(k) \tag{29}$$

where $\lambda_3 > 0$ and $0 < \delta_3 < 1$.

*Theorem 1:* Considering system (10) with the controller (28), virtual design (19), (24) and the update law (21), (25), (29), all the signals involved are semiglobal uniform ultimate bounded.

The proof is quite similar to [16] and thus omitted here to save space.

### B. Adaptive NN Control for Velocity Subsystem

Define $X_V(k) = [V(k), x_1(k), x_2(k), x_3(k), x_4(k), V_d(k+1)]^T$ and $z_V(k) = V(k) - V_d(k)$

$$\begin{aligned} z_V(k+1) &= V(k+1) - V_d(k+1) \\ &= V(k) + T_V[f_V(k) + g_V(k)u_V(k)] - V_d(k+1) \end{aligned} \tag{30}$$

The control input is designed as

$$\begin{aligned} u_V(k) &= \frac{1}{T_V g_{VN}}[-T_V f_{VN}(k) - V(k) + V_d(k+1)] \\ &+ \hat{\omega}_V^T(k)\theta_V(X_V(k)) \end{aligned} \tag{31}$$

where $\hat{\omega}_V$ is the estimation of $\omega_V^*$.

The robust updating law for NN weights is

$$\hat{\omega}_V(k+1) = \hat{\omega}_V(k) - \lambda_V z_V(k+1)\theta_V(X_V(k)) - \delta_V\hat{\omega}_V(k) \tag{32}$$

where $\lambda_V > 0$ and $0 < \delta_V < 1$.

*Theorem 2:* Considering system (9) with the controller (31) and the update law (32), the velocity is semiglobal uniform ultimate bounded. The proof is omitted here.

## VI. Simulations

In this section, we verify the effectiveness and performance of the proposed adaptive neural controller. The flight of the vehicle is at trimmed cruise condition $M = 15$, $V = 15,060$ft/s, $h = 110,000$ft. Reference commands are generated by the filter:

$$\frac{h_d}{h_c} = \frac{\omega_{n1}\omega_{n2}^2}{(s+\omega_{n1})(s^2 + 2\varepsilon_c\omega_{n2}s + \omega_{n2}^2)} \tag{33}$$

$$\frac{V_d}{V_c} = \frac{\omega_{n3}}{(s+\omega_{n3})} \tag{34}$$

where $\omega_{n1} = 0.2$, $\omega_{n2} = 0.2$, $\varepsilon_c = 0.7$, $\omega_{n3} = 0.1$.

The parameters for the controller are selected as $k_h = 0.2$, $k_I = 0.1$, $\lambda_1 = 0.05$, $\lambda_2 = 0.05$, $\lambda_3 = 0.05$, $\delta_1 = 0.02$, $\delta_2 = 0.02$, $\delta_3 = 0.02$, $T_s = 0.03$s, $T_V = 0.12$s, $c_1 = 0.9$, $c_2 = 0.8$, $c_3 = 0.2$, $c_V = 0.6$.

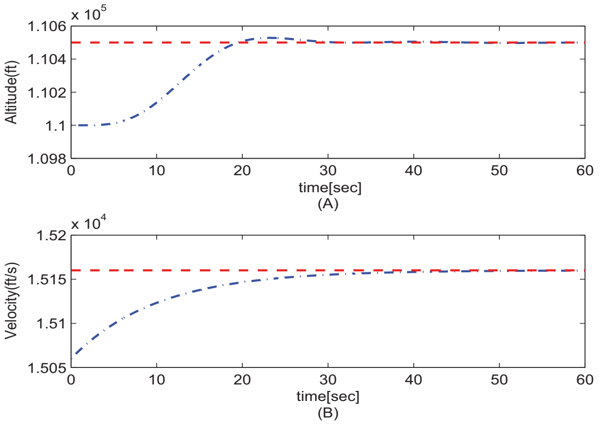


Fig. 1. Step Tracking: Altitude and Velocity Response

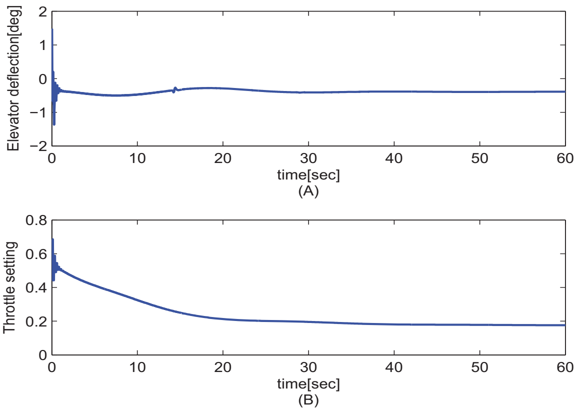


Fig. 2. Step Tracking: Control Inputs

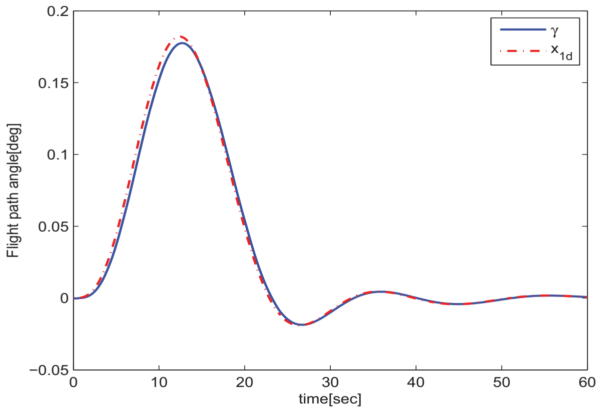


Fig. 3. Step Tracking: Flight Path Angle Tracking

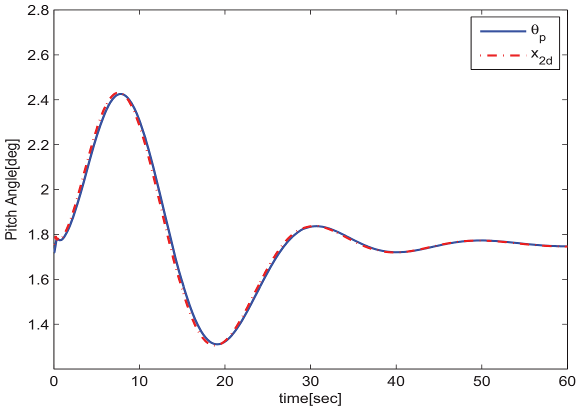


Fig. 4. Step Tracking: Attack Angle

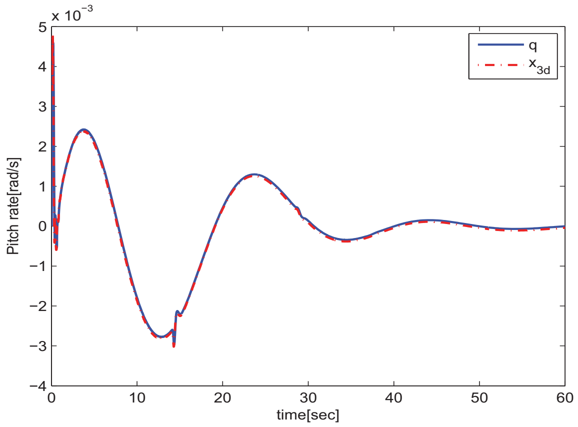


Fig. 5. Step Tracking: Pitch Rate

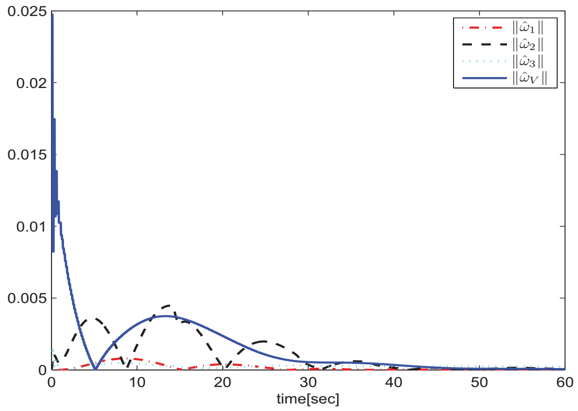


Fig. 6. Step Tracking: Trajectories of NN Weights

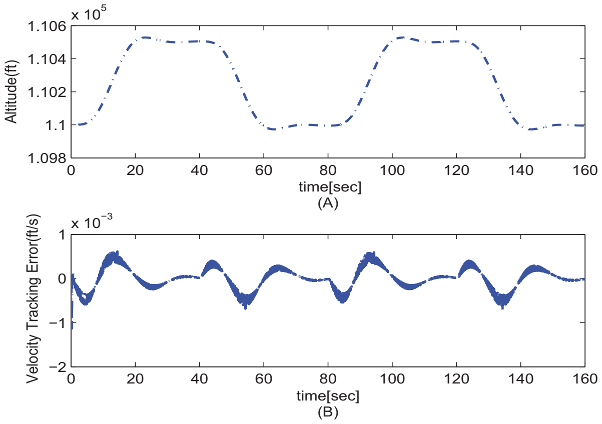


Fig. 7. Square Signal Tracking: Altitude and Velocity Response

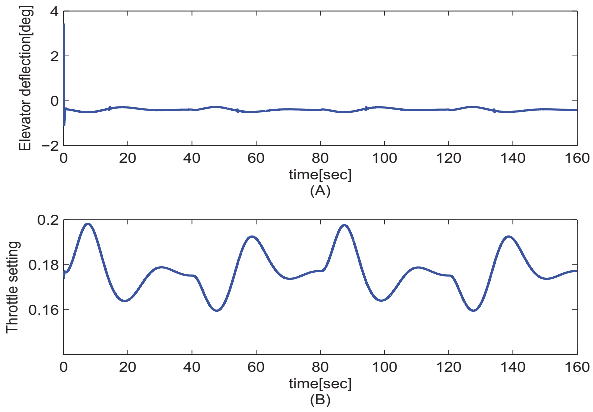


Fig. 8. Square Signal Tracking: Control Inputs

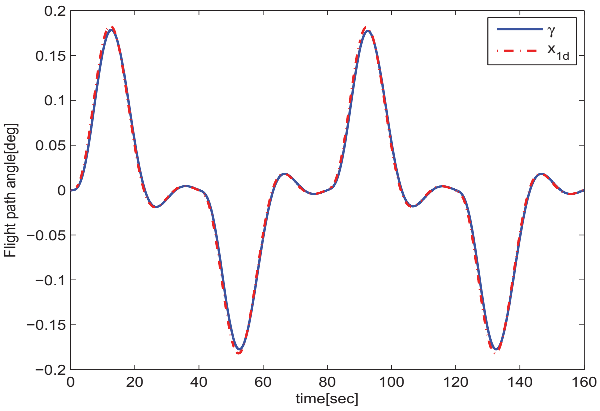


Fig. 9. Square Signal Tracking: Flight Path Angle Tracking

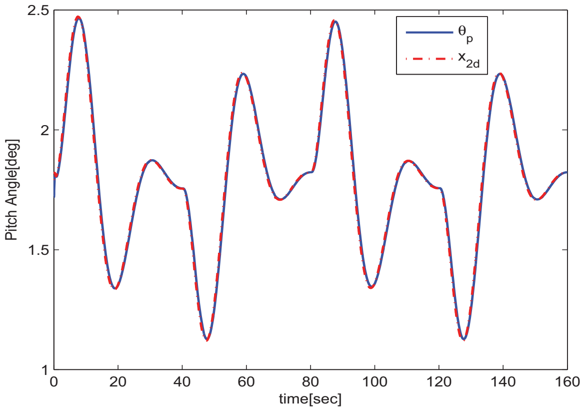


Fig. 10. Square Signal Tracking: Attack Angle

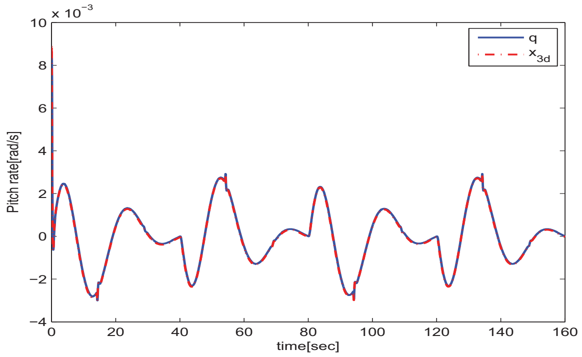


Fig. 11. Square Signal Tracking: Pitch Rate

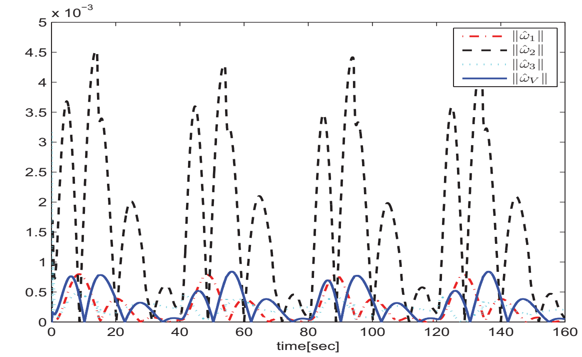


Fig. 12. Square Signal Tracking: Trajectories of NN weights

### A. Step Tracking

Fig.1 depicts the response performance that the altitude controller tracks the step change with magnitude 500ft while the velocity steps from 15060ft/s to 15160ft/s. The control inputs of the elevator deflection and the throttle setting are shown in Fig.2. Flight path angle tracks the reference command very well in Fig.3 so that the system tracks the altitude step change. From the pitch angle in Fig.4 and pitch rate response in Fig.5, we know the related system states are bounded. Also we find the NN weights are bounded in Fig.6.

### B. Square Signal Tracking

The system tracks the square signal with amplitudes 500ft and period 80 seconds while velocity is maintained in the neighborhood of 15060ft/s. The results are referred to Fig.7-12. It can be observed that all the system states are bounded and the velocity is regulated with a small error around the initial value. Also it is noted that the system states flight path angle, pitch angle, pitch rate perform good tracking response of altitude command, $x_{2d}$, $x_{3d}$ separately.

## VII. Conclusions

The altitude control of HFV is transformed into the flight path angle tracking. The two dynamics are sampled by different rate. In this way, the velocity and throttle setting from the slow dynamics can be employed for the design of attitude subsystem. Simulation results show the effectiveness of the method.

## Acknowledgements

This work was supported by DSO National Laboratories of Singapore through a Strategic Project Grant (Project No. DSOCL10004), National Science Foundation of China (Grants No: 61134004,61004002), NWPU Basic Research Funding (Grant No: JC20120236) and Aeronautical Science Foundation of China(Grant No: 20110184).

## References

[1] C. Yang, S. Ge, and T. Lee, "Output feedback adaptive control of a class of nonlinear discrete-time systems with unknown control directions," *Automatica*, vol. 45, no. 1, pp. 270–276, 2009.

[2] R. Stengel, J. Broussard, and P. Berry, "Digital controllers for VTOL aircraft," *IEEE Transactions on Aerospace and Electronic Systems*, no. 1, pp. 54–63, 1978.

[3] N. Lincoln and S. Veres, "Application of discrete time sliding mode control to a spacecraft in 6DoF with parameter identification," *International Journal of Control*, vol. 83, no. 11, pp. 2217–2231, 2010.

[4] K. Kanai, N. Hori, and P. Nikiforuk, "A discrete-time multivariable model-following method applied to decoupled flight control," *Journal of Guidance, Control, and Dynamics*, vol. 9, no. 4, pp. 403–407, 1986.

[5] D. Shin and Y. Kim, "Nonlinear discrete-time reconfigurable flight control law using neural networks," *IEEE Transactions on Control Systems Technology*, vol. 14, no. 3, pp. 408–422, 2006.

[6] Y. Hu, F. Sun, and H. Liu, "Neural network-based robust control for hypersonic flight vehicle with uncertainty modelling," *International Journal of Modelling, Identification and Control*, vol. 11, no. 1, pp. 87–98, 2010.

[7] H. Buschek and A. Calise, "Uncertainty modeling and fixed-order controller design for a hypersonic vehicle model," *Journal of Guidance, Control, and Dynamics*, vol. 20, no. 1, pp. 42–48, 1997.

[8] Z. Dydek, A. Annaswamy, and E. Lavretsky, "Adaptive control and the NASA X-15-3 flight revisited," *IEEE Control Systems Magazine*, vol. 30, no. 3, pp. 32–48, 2010.

[9] H. Xu, M. Mirmirani, and P. Ioannou, "Adaptive sliding mode control design for a hypersonic flight vehicle," *Journal of Guidance, Control, and Dynamics*, vol. 27, no. 5, pp. 829–838, 2004.

[10] L. Fiorentini, A. Serrani, M. Bolender, and D. Doman, "Nonlinear robust adaptive control of flexible air-breathing hypersonic vehicles," *Journal of Guidance, Control, and Dynamics*, vol. 32, no. 2, pp. 401–416, 2009.

[11] B. Xu, D. Gao, and S. Wang, "Adaptive neural control based on HGO for hypersonic flight vehicles," *SCIENCE CHINA Information Sciences*, vol. 54, no. 3, pp. 511–520, 2011.

[12] D. Gao and Z. Sun, "Fuzzy tracking control design for hypersonic vehicles via TS model," *SCIENCE CHINA Information Sciences*, vol. 54, no. 3, pp. 521–528, 2011.

[13] D. Nešić and A. Teel, "Stabilization of sampled-data nonlinear systems via backstepping on their Euler approximate model," *Automatica*, vol. 42, no. 10, pp. 1801–1808, 2006.

[14] D. Gao, Z. Sun, and T. Du, "Discrete-time controller design for hypersonic vehicle via back-stepping," *Control and Decision*, vol. 24, no. 3, pp. 459–463, 2009.

[15] C. Yang, S. Ge, C. Xiang, T. Chai, and T. Lee, "Output feedback NN control for two classes of discrete-time systems with unknown control directions in a unified approach," *IEEE Transactions on Neural Networks*, vol. 19, no. 11, pp. 1873–1886, 2008.

[16] B. Xu, F. Sun, C. Yang, D. Gao, and J. Ren, "Adaptive discrete-time controller design with neural network for hypersonic flight vehicle via back-stepping," *International Journal of Control*, vol. 84, no. 9, pp. 1543–1552, 2011.

[17] B. Xu, D. Wang, F. Sun, and Z. Shi, "Direct neural discrete control of hypersonic flight vehicle," *Nonlinear Dynamics*, 2012,DOI:10.1007/s11071-012-0451-x.

[18] B. Xu, F. Sun, H. Liu, and J. Ren, "Adaptive Kriging controller design for hypersonic flight vehicle via back-stepping," *IET Control Theory & Applications*, vol. 6, no. 4, pp. 487–497, 2012.

[19] A. Ataei and Q. Wang, "Non-linear control of an uncertain hypersonic aircraft model using robust sum-of-squares method," *IET Control Theory & Applications*, vol. 6, no. 2, pp. 203–215, 2012.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Graph Theory : Application to System Recovery

Ahmed Mekki
Univ Lille Nord de France F-59000 Lille, France,
EC LILLE, LAGIS, F-59651, Villeneuve d'Ascq, France
Email: ahmed.mekki@ec-lille.fr

Simon Collart-Dutilleul
Univ Lille Nord de France F-59000 Lille, France,
EC LILLE, LAGIS, F-59651, Villeneuve d'Ascq, France
Email: simon.collart_dutilleul@ec-lille.fr

*Abstract*—**The aim of the work presented in this paper is to introduce a method for assisting the recovery of a given system in case of failure detection. The proposed method rely on a graph-based algorithm that allows the identification of the alternative system configuration. In this way, the method guarantees the system under study functionalities/missions even in case of fault. The method is detailed in the sequel. Furthermore, in order to provide user with automated means which are at the same time simple, intuitive and rigorous, the whole of the developed mechanisms have been implemented in a prototype tool with an intuitive graphical interface that offers interesting facilities in terms of system recovery. The method is illustrated using an intelligent and autonomous vehicle case study.**

## I. Introduction

Life-critical systems are systems that its dysfunction could cause human-life death as well as an important equipment damage or loss. Thereby, this kind of system (e.g. transportation systems, intelligent and autonomous vehicle, nuclear plants, manufacturing systems, medical devices) must achieve a high level of robustness, availability, reliability and safety. Usually, system life-cycle could be divided into two main phases : before implementation and after implementation. The former phase rely on some research topics such as specification, modelling design and V&V (validation and verification). For the later phase, one can find maintenance and system recovery in case of failure detection. In our case, we use system recovery to design a re-configuration or a new arrangement of the system functional units in case of failure. Various are the causes of failure: performance problems like access to shared resource (resource contention), hardware faults, software bugs, system operators misconfiguration, . . . [1]. Nevertheless, all failure causes are often due to a material, soft dysfunction or operator manipulation.

Given the human/material impact of critical system dysfunction, such systems must guarantee the availability of its services/missions. Availability express the quality of being at hand when needed. This including the case when failure occur. Therefore, a recovery technique is strangely recommended. The aim of the work presented here is to guide the user during the system recovery and control phase.

Intelligent and autonomous vehicle (IAV) is a vehicle that is expected to achieve different tasks without the intervention of a human operator. Several projects rely on IAV but the most famous ones are the NASA's rovers, for instance the Mars Exploration Rover Mission (MER) [3]. MER is an ongoing robotic space mission involving two rovers, Spirit and Opportunity, exploring the planet Mars. It began in 2003 and its cost raises to more than US$ 900 million. Actually, to guarantee expected tasks, IAV rely on set of services provided by several hardware components (sensors, actuators, . . . ) as well as software components. Nevertheless, due to failures, it is possible that one or more services are no longer available and thereby the achievement of some tasks becomes no longer guaranteed. IAV design rely on fault tolerant control procedures in order to define strategies allowing the system to continue its operations with the required performances despite component faults [4].

As mentioned previously in this paper, we focus on the system recovery step in case of failure detection. In fact, the idea is to propose assisting means that are easy to manipulate and, at the same time, accurate. However, the more accurate and rigorous a notation is, the more abstract and difficult to handle and understand it becomes. Therefore, one of the challenges that we faced while dealing with this work was to look at both intuition/simplicity and rigour/accuracy. To deal with this, we propose a graph-based algorithm able to cover system configuration and able to identify the alternative configuration in case of system-down. However, in order to hide all the formal aspects -met when dealing with system recovery- from user, a GUI tool have been implemented to automate this step.

The paper is organized as follows: in Section II, an overview of the context and related works is given. In Section III, we present the algorithm that we have proposed. The developed tool is introduced in Section IV. The method is illustrated using an intelligent and autonomous vehicle case study in section V before concluding and suggesting some future works in Section VI.

[1] This research has been supported by CISIT project.

## II. Context and Related Work

Availability of a system is its capacity to achieve its missions/services in the occurrence of the failure of (or one or more faults within) some of its components [5]. One of the main properties that characterize availability is fault-tolerance [6]. Here, we focus on life-critical systems and thereby, given their failure cost, fault-tolerance is particularly sought-after such systems. In practice, some fault can cause a system failure by propagating the fault to the rest of the system. Therefore, fault tolerant systems (FTS) must deal with multiple failure types and thereby, must achieve fault isolation capacity and reversion modes availability (redundancy). Actually, FTS are typically based on three main concepts namely replication, redundancy and diversity [6]. These concepts could been defined as follows:

- Replication means providing multiple instances of the same system. Then, tasks/jobs are directed to all system instances in parallel. The correct result is determined by a quorum;
- Redundancy means providing multiple instances of the same system/service and switching to one of the remaining non-faulty instances, in case of a failure-detection;
- Diversity means providing various system implementations in order to deal with with errors in some specific implementation.

It should noticed that before recovery can be carried, one must first detect and diagnose the failure. Indeed, failure detection is to determine the instances while a system is facing dysfunctions. Afterwards comes the failure diagnosis which allow to locate the source of detected failure. Failure detection as well as failure diagnosis are not in the scope of this paper.

Severals studies dealing with system recovery have been proposed and published [6], [7], [8], [9], [10], [11]. Unlike this meantionned studies, our work propose a graph-based approach to tackle the system recovery issue. Furthermore, a GUI tool have been implemented to automate this step.

## III. Graph-based Method for System Recovery

### A. Idea

Let us here recall the aim of our study: we want to provide the user with simple means for assisting the system-recovery in case of failure detection. Our idea is to elaborate a supporting approach which hides the formal foundation to the user. Concretely, we will develop a GUI-tool that automate all the steps of system-recovery procedure.

The idea is to express the service-architecture of a system with a (directed) graph. In this graph the root element represents the system while the leaf represent the elementary services. We assume that a system is defined as a (non-empty) set of exploitation modes. At a given instance, only one exploitation mode is active. Each exploitation mode is composed of a (non-empty) set of missions. All missions of active exploitation mode should be active. Multiple versions are defined for each mission and thereby, at a given instant, only one version is active. A mission version is composed of (non-empty) set of service that, all of them, should be active. Various version of each service could be defined but only one service version is active at a given instance. A version service is (non-empty) set of elementary services that should be active, all of them. This description could be expressed as follow:

| | |
|---|---|
| **system** | = OR({Exploitation-Mode}) |
| **Exploitation-Mode** | = AND({Mission}) |
| **Missions** | = OR({Version-Mission}) |
| **Version-Mission** | = AND({Service}) |
| **Service** | = OR({Version-Service}) |
| **Version-Service** | = AND({Elementary-Service}) |

Where

- $E = AND$(s) means that an element $E$ is active if all the elements composing it are active and vice-versa;
- $E = OR$(s) means that an element $E$ is active if one and only one of elements composing it is active and vice-versa.

### B. Foundations and Problem's Formalization

A graph $\mathcal{G}$ [12] is a pair $\mathcal{G}$ = ($\mathcal{V}$,$\mathcal{E}$) where

- $\mathcal{V}$ is a finite set of vertices (nodes), and
- $\mathcal{E}$ is the set of edges, formed by pairs of vertices, $\mathcal{E}$ is a subset of $\mathcal{P}_2$(V) ($\mathcal{E} \subseteq (\mathcal{V} \times \mathcal{V})$).

Graphically, a graph is pictured by drawing a four-square (or point) for each vertex and representing each edge by a curve joining its endpoints.

Usually, *a simple graph* denotes a graph having no loops or multiple edges where each edge $e \in \mathcal{E}$ can be specified by its endpoints $(a, b) \in \mathcal{V}$ and is denoted $e = ab$. In this case, we say $a$ and $b$ are *adjacent*. Thereby, a path is defined as a set of consecutive and ordered nodes so that two nodes are adjacent if and only if they are consecutive in the ordering.

A graph variant is called directed graphs $\mathcal{D}$ = ($\mathcal{V}$, $\mathcal{E}$), where the edges have a direction. In other words, edges are ordered and $ab \neq ba$. Graphically, the edges are drawn as arrows.

### C. System description

In our case, a system is defined as directed graph $\mathcal{DS}$=($\mathcal{V}$, $\mathcal{E}$) where

- $\mathcal{V}$ is a finite set of nodes, $\mathcal{V} = Sys \cup EM \cup M \cup VM \cup S \cup VS \cup ES$:
    1) $Sys$ is the root element of the system under study,
    2) $EM$ is the set of exploitation modes of the system under study,
    3) $M$ is the set of missions defined by $EM$,

4) *VM* is the set of mission version (for each mission, one can define at least one mission version),
5) *S* is the set of services composing missions,
6) *VS* is the set of service version (for each service, one can define at least one service version),
7) *ES* is the set of all elementary services proposed by the system.

- $\mathcal{E}$ is a finite set of edges, $\mathcal{E} \subseteq (Sys \times EM) \cup (EM \times M) \cup (M \times VM) \cup (VM \times S) \cup (S \times VS) \cup (VS \times ES)$.

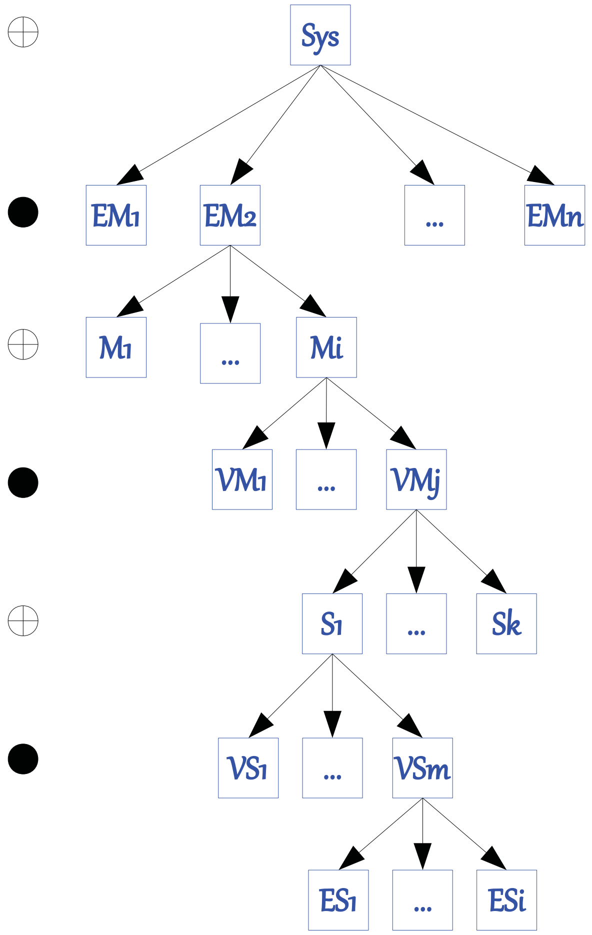


Fig. 1. System description as a directed graph

Figure 1 depicts graphically the system definition given in this section. To this definition, we add a set of binary relations (using AND and OR binary operators) between elements. Indeed, these relations define the type of composition between nodes, given above in section III-A. Actually, for each element type (node), a symbol is linked. Two kind of symbols are used: ⊕ and ●. Node linked to ⊕ is defined as OR-composition of its children. In the same way, node linked to ● is defined as AND-composition of its children.

### D. Algorithm

Failure detection is the task of identifying system dysfunction. Different source of failure could be identified in a given system. A recovery step is switching to an exploitation mode where all (or a part) of system capabilities/facilities are guaranteed in case of failure. Thereby, the exact cause of error is often not required for recovery to take place. Nevertheless, an automated approach for the recovery step is necessary in order to improve system availability. In this section, we propose an algorithm that automate this step. Given a graph model of the system under study and given the detected failure, the algorithm determine and update the list of the alternative exploitation modes. Actually, based on the system graph, the algorithm determine a sub-graph (recovery tree) that represent the non-faulty nodes.

In practice, for each node, we add two attributes: *statut* and *activity*. *Statut* denotes the statement of a node: *functional* or *not-functional*, where *activity* denotes the activity statement of a node: *active* or *not-active*. Furthermore, the approach that we propose require that the failure lists all the element that are no longer available. In other word, the failure should be defined as the set of nodes that are no longer available in the current system description model.

The recovery algorithm rely on two parameters, namely the system under study graph and the failure (a list of faulty elements), and is composed of three main functions and is conducted as follow:

1) the *statut* attribute of all faulty element is set to *not-functional* within the graph model of the system under study,
2) For each node n in the list of faulty elements (the detected failure), the failure spread function, *FS(n)*, is carried,
3) Determine alternative exploitation mode

**Failure spread function of a node n: *FS(n)***

**Step 1:** Set the *statut* attribute all children of n to *non-functional*,
**Step 2:** Set the *statut* attribute of n to *non-functional*,
**Step 3:** If parent of n is an AND-composition type element, then

1) n ← Parent_of(n),
2) go to **Step 2**.

**Step 4:** If parent of n is an OR-composition type element and if all sibling of node "n" are *non-fonctionnel*,then

1) n ← Parent_of(n),
2) go to **Step 2**.

**Determine alternative exploitation mode**

Once the failure is spread, the next step is to determine the alternative exploitation modes that could be activated. Actually, the aim of this task is to determine the sub-graph where all nodes have a *statut* attribute value as *functional*. In other words, the alternative graph is composed only by nodes that statut is *functional*.

Let $\mathcal{G} = (\mathcal{V},\mathcal{E})$ be the graph of system under study and let $\mathcal{G}' = (\mathcal{V}',\mathcal{E}')$ be the sub-graph of $\mathcal{G}$ where all nodes are *functional* and is determined as follow:

**Step 0:** $\mathcal{V}' = \emptyset \wedge \mathcal{E}' = \emptyset$,
**Step 1:** if root of $\mathcal{G}$ is faulty, exit algorithm
**Step 2:** the root of $\mathcal{G}$ is chosen and is added to $\mathcal{G}'$,
**Step 3:** a walk[1] $\omega$ that starts from the chosen node is constructed by adding only *functional* nodes.
**Step 4:** if all nodes of $\mathcal{G}$ are tested, then the walk $\omega$ represents the graph $\mathcal{G}'$ (break algorithm) else go to step 5,
**Step 5:** Select the parent of the last node of $\omega$, go to **Step 6**
**Step 6:** A walk $\omega'$ that starts from the chosen node is constructed by adding only *functional* nodes.
**Step 7:** $\omega \leftarrow$ concatenate($\omega$,$\omega'$). Go to **Step 4**

It is worthy to notice that the obtained walk is a connected graph $\mathcal{G}'$ since the start node of this function is the root node. This connected graph is equivalent to the *spanning tree* for the undirected graph representation of graph $\mathcal{G}$. Once the *spanning tree* is returned, a product of all possible exploitation mode configuration is computed based on it. Thereby, from the list of all possible configuration, the user (operator) could choose one (and only one) to activate.

## IV. Recovery Tool

The various mechanisms we have developed have been implemented within a software tool. This tool guides the user during the recovery step, it offers a quite intuitive graphical user interface. Actually, the tool rely on two main elements:

- Recovery algorithm: first, the developed tool displays the system description and lists its various parameters. Indeed, as shown above in the paper, the system description is based on a graph theory. A graphic representation of this graph is done ones its description (graph description of the system to check) is loaded. This representation is made on a tab within the recovery tool. Then, based on the algorithm discussed in section III-D and given a failure, the tool can automatically determine all possible alternative exploitation modes. Finally, the switch to the selected exploitation mode is carried.
- Data structure: since we aim that the developed tool can be used in combination with other tools within a global system supervision and control approach, we have chosen XML (Extensible Markup Language) [13] standard format for the input/output files. The main advantage of using the standard XML is that format of the encoding documents is both human-readable and machine-readable. Actualy, our tool rely on three different files: a system description which is an input/output file, the initial configuration file and the failure file, both are input files. The corresponding structure of each file has been defined by an XML schema. Listings 1, 2 and 3 define respectively the XML schema for the system description, the initial configuration and the failure.

Listing 1. System XML schema

```
<!ELEMENT Systeme (mode+)>
<!ELEMENT mode (nom?, Mission+)>
<!ELEMENT Mission (nom?, VMission+)>
<!ELEMENT VMission (nom?, service+)>
<!ELEMENT service (nom?, vservice+)>
<!ELEMENT vservice (nom?, eservice+)>
<!ELEMENT nom (#PCDATA)>
<!ELEMENT eservice (#PCDATA)>
<!ATTLIST mode
        activity ( Active | notActive ) #REQUIRED
        statut ( nonFunctional | Functional ) #REQUIRED>
<!ATTLIST Mission
        activity ( Active | notActive ) #REQUIRED
        statut ( nonFunctional | Functional ) #REQUIRED>
<!ATTLIST VMission
        activity ( Active | notActive ) #REQUIRED
        statut ( nonFunctional | Functional ) #REQUIRED>
<!ATTLIST service
        activity ( Active | notActive ) #REQUIRED
        statut ( nonFunctional | Functional ) #REQUIRED>
<!ATTLIST vservice
        activity ( Active | notActive ) #REQUIRED
        statut ( nonFunctional | Functional ) #REQUIRED>
<!ATTLIST eservice
        activity ( Active | notActive ) #REQUIRED
        statut ( nonFunctional | Functional ) #REQUIRED>
```

Listing 2. Initialization XML schema

```
<!ELEMENT Init (mode?)>
<!ELEMENT mode (nom?, eservice+)>
<!ELEMENT nom (#PCDATA)>
<!ELEMENT eservice (#PCDATA)>
```

Listing 3. Failure XML schema

```
<!ELEMENT Erreur
        (mode*,
        Mission*,
        VMission*,
        service*,
        vservice*,
        eservice*)>
```

In the sequel we will discuss the steps of a recovery process and we show how the tool facilities are helpful and useful when carrying this process (figure 2). From the menu bar (green box),

1) First, the user should start by loading the system specification. The uploaded description will be then drawn in the shape of tree (blue box),
2) Once the system description is uploaded, an initial configuration should be selected. In the same way, the initial configuration is then represented in the shape of tree (black box),

[1] A walk is a list $v_0, e_1, v_1 \dots e_n, v_n$ of nodes and edges such that for $1 \leqslant i \leqslant n$, the edge $e_i$ has endpoints $v_{i-1}$ and $v_i$.

3) Then, relying on the uploaded failure file, the recovery algorithm is carried automatically. Afterwards, a box showing (red box) all possible exploitation mode is shown and from which a system mode configuration could be selected (to update the current configuration.

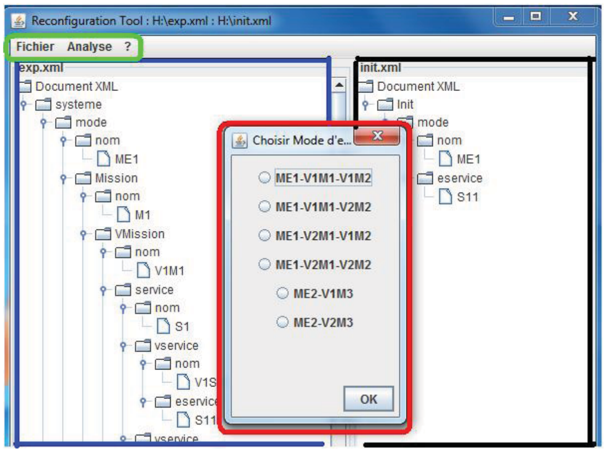


Fig. 2. Main GUI of the developed tool

In order to illustrate all this steps an IAV case study is presented hereafter.

## V. Application to IAV system

### A. System description

In real environment, an IAV could be repaired after the fault detection and isolation is carried. Nevertheless, within some critical environments, where the intervention (the necessary repairs) of a human operator is impossible, fault tolerance property must be provided to the IAV. In this case, IAV must be kept continuously operating even with a fault, such as when IAV operating in distant and/or dangerous areas. Consequently, IAV systems rely on service (component) redundancy in order to guarantee fault tolerance property. For illustration, let us consider an IAV described (Figure 3[2]) as follows :

- RoBuCar can carry two different exploitation modes : normal mode and degraded mode.
- Unlike normal mode where the four wheels are used, degraded mode rely on three or two wheels. Two main differences that distinguish the two exploitation modes: the maximum authorized speed and the weight of the transported goods for each mode.
- RoBuCar is expected to provide several missions : moving on, moving back, curbing, . . .
- Each wheel is equipped with two electrical motors : only one is used to move the wheel. The second motor is used in case of dysfunction of the first motor.
- To move a wheel, four services are required : (1) the generation of an electrical power, (2) the conversion of the electrical power to a mechanical one, (3) the transmission of the mechanical energy to the tire and (4) a measurement service required to control the motor in closed loop.
  1) The generation of an electrical power rely on three elementary services namely: (a) supply the motor, (b) induce an electrical power and (c) limit the electrical current.
  2) The conversion of the electrical power to a mechanical one is composed by one elementary service.
  3) The transmission of the mechanical energy to the tire is composed of two elementary services namely: (a) keep all the pieces in rotation and (b) load service.
  4) The measurement service is composed of three elementary services namely: (a) measure the electrical current, (b) measure the velocity and (c) measure distances.

Let us recall that, in order to keep system operating even in case of fault, service (component) redundancy is essential. Therefore, different versions of the same service as well as of the same mission are available.

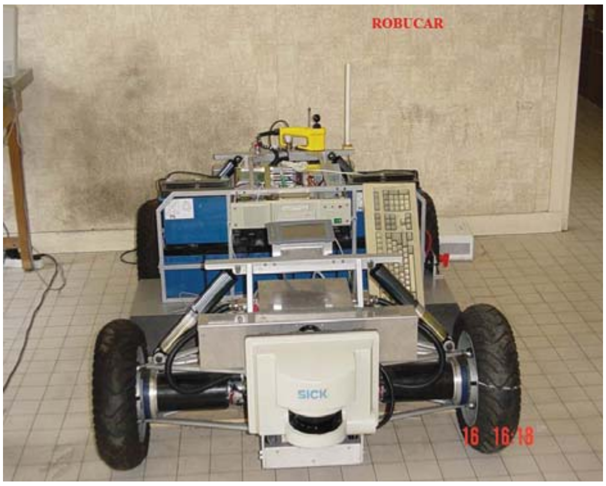


Fig. 3. RoBuCar

### B. Recovery procedure

The service graph associated to the RoBuCar system described above is depicted graphically in figure 4 (for simplicity reason, some details are omitted). The XML file describing this graph is used to carry on the recovery tool. Afterwards, in case of a failure detection, the recovery tool is carried.

For example, suppose that one of the four wheels is no longer available. In other words one of the two

[2] RoBuCar : is an IAV at LAGIS laboratory built by the Robosoft Company http://lagis.ec-lille.fr/

electrical motors that equippe the wheel is no longer available. The XML file describing this failure is defined according to the XML schema given above. The failure XML file is needed by the recovery tool. Once the failure file is loaded, the tool, automatically, finds alternative exploitation modes to keep the vehicle working despite the failure. In our case, two alternative exploitation modes are retuned :

1) Keep the "normal mode" by proposing to activate the second electrical motor of the wheel.
2) Skip to the "degraded mode" where only three wheel are used.

The main advantage is that the procedure is automatic and rely on standard format (XML) for systems exchange (the system description, the failure description, . . . ). Therefore, its integration in operating system supervision approach can easily be done.

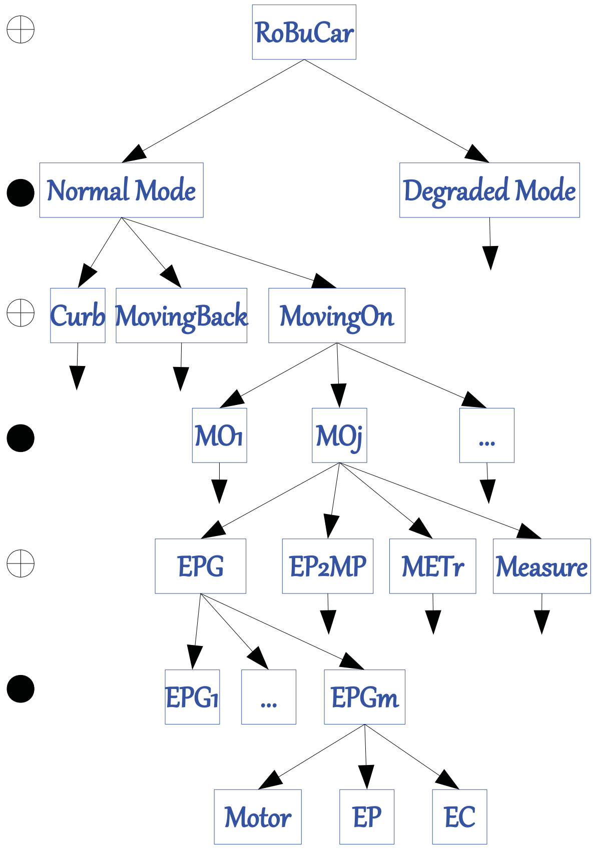


Fig. 4. Simplified version of the graph service of the RoBuCar

Therefore, the method allows to keep IAV continuously operating even with failures. In practice, this is very interesting specially for IAV operating in distant or/and dangerous areas where human cannot operate.

## VI. Conclusions

In this paper we aim at introducing a means to assist and guide the user while carrying the system recovery in case of failure detection. The proposed approach rely on the graph theory. The main advantage of this notation that is formal and thereby, automatic tool using this notation could implemented. Actually, we started by proposing a graph-based description of the study system. Then, we defined an algorithm for the recovery system in case of failure detection. All proposed mechanisms we have developed have been implemented within a software tool. In order to illustrate the developed method an intelligent and automatic vehicle (IAV) case study was presented. First, a textual description of system was given. Then, the service graph is depicted on which the recovery tool is carried. Although, within some critical environments, where the intervention of a human operator is not possible, IAV must be kept continuously operating even with a fault, such as operating in distant and/or dangerous areas. Here we showed how the developped approach could, automatically, be used to detremine alternative exploitation modes.

## References

[1] S. Pertet and P. Narasimhan, *Causes of failure in web applications*, Carnegie Mellon University Parallel Data Lab, Tech Rep CMU-PDL- 05-109, Dec 2005.
[2] http://esj.com/blogs/enterprise-insights/2011/06/it-systems-failure-costs-quantified.aspx
[3] J. Carsten, A. Rankin, D. Ferguson, and A. Stentz, *Global Planning on the Mars Exploration Rovers: Software Integration and Surface Testing*, Journal of Field Robotics, 26(4), April 2009, 337-357.
[4] N. Chatti, *Online supervision of intelligent vehicle using functional and behavioral models*, Intelligent Vehicles Symposium (IV), 2011 IEEE, p. 827-832, 2011.
[5] DP. Siewiorek and RS. Swarz, *The theory and practice of reliable system design*, Digital Press, Bedford, Massachusetts 1982.
[6] J. Bowen and V. Stavridou, *Safety-critical systems, formal methods and standards*, Software Engineering Journal, 8(4), p. 189-209, IET.
[7] N.G. Leveson, *Software safety: Why, what and how*, ACM Computing Surveys, 18, p. 125–163, 1986.
[8] E. Hammami, *Déploiement sensible au contexte et reconfiguration des applications dans les sessions collaboratives*, Thèse à l'Université de Toulouse, 2007.
[9] R. Sirdey, *Modèles et algorithmes pour la reconfiguration de systèmes répartis utilisés en téléphonie cellulaire*, Thèse à l'Université de Technologie de Compiègne, 2007.
[10] M. Staroswiecki and A-L. Gehin, *From control to supervision*, Annual Reviews in Control, vol. 25, p. 1-11, 2001.
[11] G. Bajpai, H.G. Kwatny and B.C. Chang, *Control systems perspective on safety critical systems*, 8th Asian Control Conference (ASCC) p. 413 -417, 2011.
[12] R. Diestel, *Graph Theory*, Springer-Verlag, Graduate Texts in Mathematics, Third edition, 173 pages, 2005.
[13] W3C, *XML specification 1.0*, http://www.w3.org/TR/xml/

# Detection of Low Adhesion in the Railway Vehicle Wheel/Rail Interface: Assessment of Multi-Bodied Simulation Data

Christopher Ward, Roger Goodall, Roger Dixon
School of Electronic, Electrical
and Systems Engineering
Loughborough University
Loughborough, Leicestershire
UK, LE11 3TU
Email: c.p.ward@lboro.ac.uk, r.m.goodall@lboro.ac.uk,
r.dixon@lboro.ac.uk

Guy Charles
Faculty of Engineering
Coates Building
University of Nottingham
Nottingham
UK, NG7 2RD
Email: guy.charles@nottingham.ac.uk

***Abstract*—Low adhesion in the wheel/rail interface of railway vehicles creates safety and punctuality issues in terms of missed station stops and signals passed at danger. RSSB project T959 is tasked with developing advanced monitoring techniques for the detection of adhesion in this key interface. A number of techniques were developed and initially tested on simplified models of a rail vehicle. The efficacy of these techniques is now being tested with more representative data produced by multi-bodied physics simulation package Vampire. This paper therefore covers the outcomes of the Vampire testing, initial application of a Kalman-Bucy filter creep force estimator to the Vampire data, and application of a data comparison method based upon the Sprague and Geers method, also to the Vampire data.**

## I. Introduction

Low adhesion in the wheel/rail contact of railway vehicles is a current issue occupying the railway industry. This is commonly reported as the ‘leaves on the line’ issue and can create large safety and punctuality issues as rail vehicles fail to stop at stations or pass signals at danger. RSSB managed project T959 [7] is tasked with finding methods of estimating the available adhesion in the wheel/rail interface using modest cost vehicle-mounted sensor sets and advanced filtering applied to in-service vehicles. Knowledge of this would allow numerous commercial benefits such as targeting mitigation methods more efficiently and scheduling rail services to make best potential use of the available adhesion.

Early stages of project T959 investigated a number of low adhesion estimation techniques as applied to simplified plan view dynamics models of typical railway vehicles and were highlighted in [9]. The primary amongst these methods was application of a Kalman-Bucy filter (KBF) [3] which was used to estimate creep forces (longitudinal and lateral forces arising from contact mechanics of the wheel-rail interface), that were then post-processed to imply an adhesion level. Additional techniques were multiple/interacting Kalman filters and data comparison techniques. The efficacy of these methods is now being tested on more representative data produced through a multi-bodied simulation (MBS) package Vampire, success of which will lead to full scale physical testing on a fully instrumented rail vehicle.

This paper therefore covers: brief outcomes of the Vampire simulations and how these compare to the MATLAB/Simulink modelling; initial application of the KBF technique and current issues; and finally a data driven method of non-model based comparison currently being developed using the Sprague and Geers metric [8].

## II. Vampire multi-bodied simulation

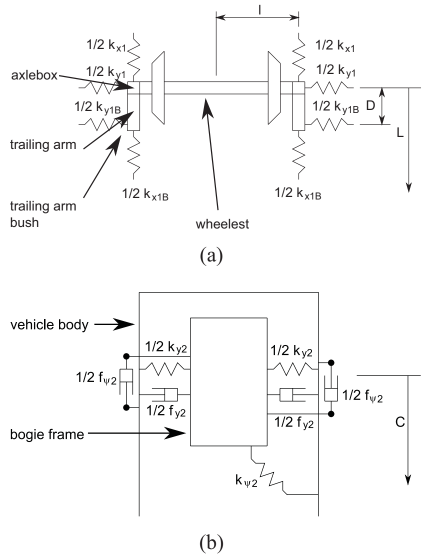
The current phase of the project is using data produced by the MBS package Vampire, [1]. The data produced from the package includes the full nonlinearity of the suspension system, as well as the nonlinearity in the wheel/rail contact. It also encapsulates any interaction of the vertical suspension components with the lateral and yaw suspension.

### A. Vehicle selection

The vehicle selected for testing is the British Rail Mk.3 coach. This vehicle was selected due to: vehicle models being readily available; and the physical testing will be likely to take place with this vehicle.

Parameters were interpreted from the MBS model to fit with the simpler plan view lateral and yaw models used for the filter design (Figure 1(a) for the primary suspension and Figure 1(b) for the secondary suspension, with parameter values shown in Table I). The Mk.3 coach is an older form of coach with many features than are no longer incorporated in vehicle design. In particular the secondary yaw damper (that increases the critical speed of bogie instability [11]) is a friction damper rather than the more common viscous damper. Due to the discontinuous nature of this component it adds a significant nonlinearity to the system that cannot be incorporated easily in a state space model.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE



(a)

(b)

Fig. 1. Suspension layouts, (a) primary suspension, (b) secondary suspension

| Parameter | Symbol | Value | Units |
|---|---|---|---|
| Body mass | $m_v$ | 24380 | $kg$ |
| Body yaw inertia | $I_v$ | 1129740 | $kgm^2$ |
| Bogie mass | $m_b$ | 2130 | $kg$ |
| Bogie yaw inertia | $I_b$ | 2870 | $kgm^2$ |
| Wheelset mass | $m_w$ | 1475 | $kg$ |
| Wheelset yaw inertia | $I_w$ | 910 | $kgm^2$ |
| Nominal radius | $r_0$ | 0.4570 | $m$ |
| Lateral stiffness (2nd) | $k_{y2}$ | 197000 | $N/m$ |
| Lateral damping (2nd) | $f_{y2}$ | 40000 | $Ns/m$ |
| Yaw stiffness (2nd) | $k_{\psi 2}$ | 175000 | $Nm$ |
| Yaw damper friction breakout (2nd) | $f_{\psi 2}$ | 11860 | $N$ |
| Longitudinal stiffness (1st) | $k_{x1}$ | 204800 | $N/m$ |
| Lateral stiffness (1st) | $k_{y1}$ | 204800 | $N/m$ |
| Longitudinal damping (1st) | $f_{x1}$ | 0 | $Ns/m$ |
| Lateral damping (1st) | $f_{y1}$ | 0 | $Ns/m$ |
| Longitudinal stiffness, bush (1st) | $k_{x1b}$ | 15696000 | $N/m$ |
| Lateral stiffness, bush (1st) | $k_{y1b}$ | 15456000 | $N/m$ |
| Longitudinal damping, bush (1st) | $f_{x1b}$ | 19400 | $Ns/m$ |
| Lateral damping, bush (1st) | $f_{y1b}$ | 13200 | $Ns/m$ |

TABLE I
MK.3 COACH INTERPRETED PARAMETERS

### B. Wheel/rail contact adhesion conditions and test runs

As highlighted in [9] the shape of the creep curves is critical to the detection of areas of low adhesion. It is assumed that the initial slope of the creep curve is constant for all adhesion conditions and that the differentiating factor is the level of creep saturation, [2]. However, as first highlighted in [5] and subsequently verified using the University of Sheffield SUROS twin-disk machine [10], the initial slope of the creep curve reduces as the adhesion conditions reduce, meaning changes in adhesion can be determined in 'normal' running and not just when the contact forces are saturated. Therefore four adhesion condition creep curves were set at dry, wet, low and very low levels for the Vampire simulation testing, Figure 2. These can be thought of as relating to friction coefficients of 0.56, 0.32, 0.072 and 0.038 respectively.

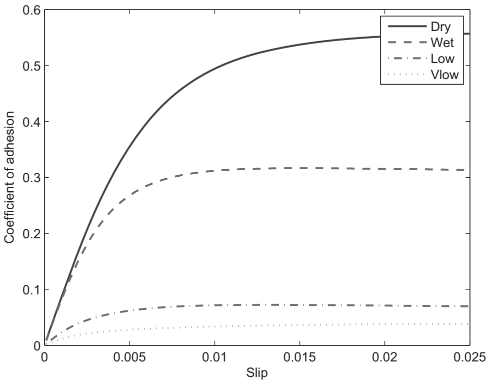


Fig. 2. Creep curves developed from the University of Sheffield SUROS twin disk machine

As with any such form of simulation there are a huge number of potential test combinations. The available variables were narrowed down to: track conditions (straight line, 200 $km/h$ design speed); vehicle speed (100 $km/h$ and 200 $km/h$); track irregularity sizes (full scale and half scale); and adhesion levels (constant, step changes and continuously varying).

### C. Vampire modelling observations

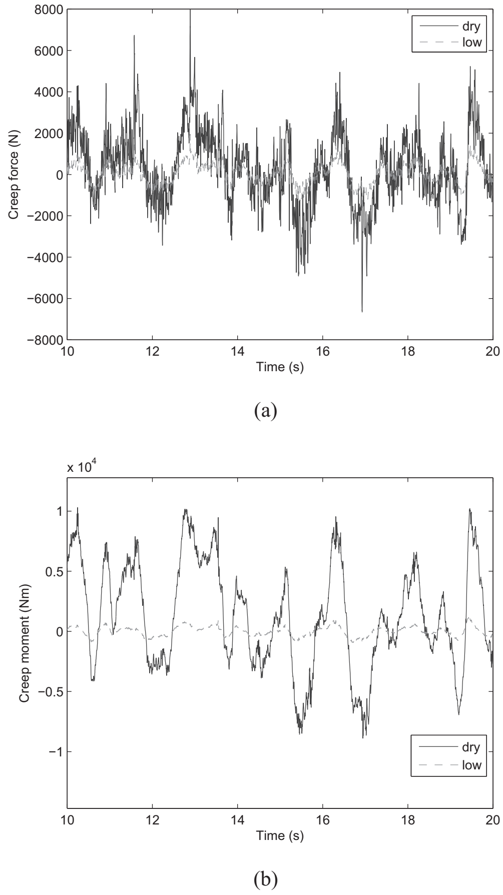
MATLAB/Simulink modelling in [9] demonstrated that the creep forces in the wheel/rail contact reduce as the adhesion level reduces. This trend is repeated with the Vampire simulation, Figure 3(a), for the lateral creep forces of the front wheelset of the front bogie. These tests were performed at a vehicle speed of 200 $km/h$ and with full sized track irregularity where there is a drop in the RMS of the lateral creep force from 1340 $N$ for the dry adhesion case to 300 $N$ for the very low adhesion case. However in the lateral case there is a constant 'gravitational' stiffness force that arises from the profiling of the wheel and varies little with adhesion conditions. This has an RMS value of 1300 $N$, therefore masking changes in the creep forces when estimated in combination. The creep moment demonstrates an even larger change in RMS between the dry adhesion case, 2900 $Nm$, and the very low adhesion case, 310 $Nm$, shown in Figure 3(b). Here gravitational moment is negligible.

## III. CREEP FORCE ESTIMATION

The KBF method of creep force estimation uses a simplified full vehicle model or half vehicle model. The latter approach is favourable due to potential reduction in sensors required and associated reduced order of the model.

### A. Open loop model comparison, half vehicle model

The open loop estimator model is first validated against the Vampire simulation outputs. The creep force estimator model is output only and does not include any terms from the track irregularity that would be costly to measure in practice. In order to test the open loop estimator model the



(a)

(b)

Fig. 3. Vampire modelling creep reductions, (a) lateral, (b) yaw

state space format is subtly modified to include the track irregularities as generated by the Vampire simulation model. The suspension models are linear formations and follow the simplified suspension layouts of Figures 1(a) and 1(b). The lateral and yaw dynamic equations of each wheelset are

$$m_w \ddot{y}_w = F_s + F_g + F_c \tag{1}$$

$$I_w \ddot{\psi}_w = M_s + M_g + M_c \tag{2}$$

where $m_w$ is the mass of the wheelset, $\ddot{y}_w$ is the lateral acceleration of the wheelset, $F_s$ is the lateral primary suspension forces, $F_g$ is the lateral gravitational stiffness, $F_c$ is the lateral creep force, $I_w$ is the moment of inertia of the wheelset, $\ddot{\psi}_w$ is the yaw acceleration of the wheelset, $M_s$ is the primary suspension yaw moment, $M_g$ is the gravitational moment and $M_c$ is the creep moment. For this linear estimator model, the open loop tests use the creep forces and moments of Kalker [2]

$$F_c = 2f_{22}\psi_w - \frac{2f_{22}}{V}\dot{y}_w \tag{3}$$

$$M_c = -\frac{2l\lambda f_{11}}{r_0}\left(y_w - d_r\right) - \frac{2l^2 f_{11}}{V}\dot{\psi}_w \tag{4}$$

where $f_{11}$ is the longitudinal creep coefficient (5770000 $N$ for the dry adhesion case, and 339000 $N$ for the very low adhesion case), $f_{22}$ is the lateral creep coefficient (5770000 $N$ for the dry adhesion case, and 339000 $N$ for the very low adhesion case), $V$ is the vehicle speed (200 $km/h$), $\lambda$ is wheelset conicity (0.131) and $d_r$ is rail lateral irregularity.

| Gain | Phase | $R^2$ | Gain | Phase | Combined |
|---|---|---|---|---|---|
| 0.5 | 0 | 0.75 | -0.5 | 0 | 0.5 |
| 1 | 0 | 1 | 0 | 0 | 0 |
| 2 | 0 | 0 | 1 | 0 | 1 |
| 1 | -90 | -1.09 | 0 | 0.52 | 0.52 |
| 1 | 90 | -0.86 | 0 | 0.48 | 0.48 |
| 1 | 180 | -3 | 0 | 1 | 1 |

TABLE II
SPRAGUE AND GEERS METRIC TEST

Figure 4(a) shows a section of half vehicle open loop model data excited by the track irregularity file compared to the Vampire outputs for the dry adhesion case for the front wheelset yaw, yaw rate and yaw accelerations. Visual inspection shows some differences between the outputs in terms of gain, but that the general trend is that the frequency content is followed. Numerically this is assessed using the Sprague and Geers metric.

*1) Sprague and Geers metric:* This metric was initially used for the comparison of different wave patterns in fluid flows. If $m(t)$ is the measured history and $c(t)$ is the estimated history, then a number of time integrals can be defined

$$\upsilon_{mm} = (t_2 - t_1)^{-1}\int_{t_1}^{t_2} m^2(t)dt \tag{5}$$

$$\upsilon_{cc} = (t_2 - t_1)^{-1}\int_{t_1}^{t_2} c^2(t)dt \tag{6}$$

$$\upsilon_{mc} = (t_2 - t_1)^{-1}\int_{t_1}^{t_2} m(t)c(t)dt \tag{7}$$

where $t_1 < t < t_2$ is the time step of interest, the error in the magnitude is given as

$$M_{SG} = \sqrt{\frac{\upsilon_{cc}}{\upsilon_{mm}}} - 1 \tag{8}$$

the phase error is given by

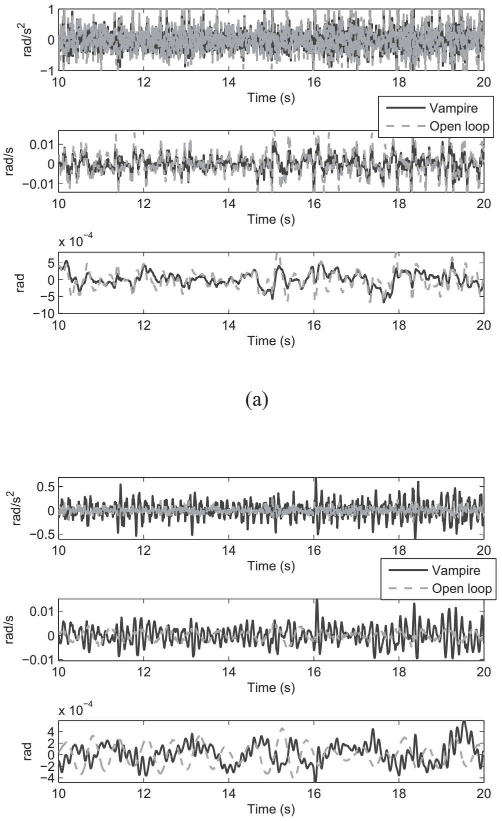
$$P_{SG} = \frac{1}{\pi}cos^{-1}\left(\frac{\upsilon_{mc}}{\sqrt{\upsilon_{mm}\upsilon_{cc}}}\right) \tag{9}$$

these two errors can be combined to give an overall global error

$$C_{SG} = \sqrt{M_{SG}^2 + P_{SG}^2} \tag{10}$$

this is comparable to the $R^2$ method of [4], but is able to cope with a degree of phase lag in the signals. This is illustrated in Table II for a series of sine wave comparison tests where the wave is scaled and phased. This demonstrates that the Sprague and Geers Metric can give information about the size and direction of the gain comparison between the original and estimated value, whereas the $R^2$ metric begins to fail when the scaling is past 2. The Sprague and Geers metric also gives good metrics information when there is phase difference between the signals as will be the case in this application where the signals may not be perfectly aligned, under which conditions the $R^2$ metric fails to see a correlation.

The Sprague and Geers metrics for the open loop model correlation are summarised in Table III. These values show for the acceleration and position signals that they are approximately 1.2 times larger that those from Vampire and the rate signal is 1.5 times larger. There is also mainly agreement in the phase signals of the analysis, with phases of less that $40^0$, though this is more difficult to interpret with more widely spaced spectrum signals. The modelling gives some confidence that the robust properties of the KBF would be able to accommodate model mismatches of this order. Figure 4(b) shows a section of open loop and Vampire output data for the very low adhesion case. This shows that the open loop model now is no longer correlated with the Vampire data, either in terms of gain or phase. This is reflected in the Sprague and Geers metrics shown in Table III, where the phase equivalent is considerably larger for the dry case and gain content is mostly much lower than that observed from the Vampire simulation with the exception of the yaw angle.



(a)

(b)

Fig. 4. Open loop estimator model comparison with Vampire output data, (a) dry case, (b) very low

This therefore represents a significant model mismatch at the lower adhesion levels and may be due to a number of reasons that have not full been understood: stick/slip dynamics in the the secondary yaw damper of the vehicle body and bogie, poor model extraction from the Vampire modelling, etc. Steps are currently being undertaken to understand these dynamics.

| Condition | Parameter | Gain | Phase | Combined |
|---|---|---|---|---|
| Dry | $\ddot{\psi}_{FF}$ | 0.1971 | 0.1332 | 0.2379 |
| Dry | $\dot{\psi}_{FF}$ | 0.5727 | 0.1528 | 0.5928 |
| Dry | $\psi_{FF}$ | 0.1990 | 0.2497 | 0.3193 |
| Very low | $\ddot{\psi}_{FF}$ | -0.5923 | 0.4922 | 0.7701 |
| Very low | $\dot{\psi}_{FF}$ | -0.4603 | 0.5585 | 0.7238 |
| Very low | $\psi_{FF}$ | -0.0390 | 0.5730 | 0.5744 |

TABLE III

SPRAGUE AND GEERS METRIC COMPARISONS FOR THE OPEN LOOP ESTIMATOR MODEL

### B. Creep force estimation example

The open loop suspension modelling from the previous section is now applied to the KBF creep force estimation method. It is noted at the outset that due to the discrepancies in the estimator modelling of the Mk.3 coach at low adhesion levels it is expected that the KBF performance will be affected. The size of this performance deficit requires assessment due to the robust qualities of the KBF.

For this example a full possible measurement vector is used

$$\begin{aligned} y = [&y_{FF}\ \dot{y}_{FF}\ \psi_{FF}\ \dot{\psi}_{FF}\ y_{FR}\ \dot{y}_{FR} \cdots \\ &\cdots\ \psi_{FR}\ \dot{\psi}_{FR}\ y_{BF}\ \dot{y}_{BF}\ \psi_{BF}\ \dot{\psi}_{BF} \cdots \\ &\cdots y_V\ \dot{y}_V]^T \end{aligned} \tag{11}$$

where the subscript $_{FF}$ refers to the front wheelset of the front bogie, subscript $_{FR}$ refers to the rear wheelset of the front bogie, subscript $_{BF}$ refers to the front bogie and subscript $_V$ refers to the vehicle body. It should be noted that this case represents the highest number of measurements possible and will not be practical in a long term application. The state vector is defined as

$$\begin{aligned} x = [&y_{FF}\ \dot{y}_{FF}\ \psi_{FF}\ \dot{\psi}_{FF}\ y_{FR}\ \dot{y}_{FR} \cdots \\ &\cdots\ \psi_{FR}\ \dot{\psi}_{FR}\ y_{BF}\ \dot{y}_{BF}\ \psi_{BF}\ \dot{\psi}_{BF} \cdots \\ &\cdots y_V\ \dot{y}_V\ F_{FF}\ F_{FR}\ M_{FF}\ M_{FR}]^T \end{aligned} \tag{12}$$

where $F$ is the lateral creep force and gravitational stiffness combined, and $M$ is the combined gravitational and creep moment. No physics of the creep forces are included in the estimator model, instead this is defined as

$$\dot{F}_{FF} = \dot{F}_{FR} = \dot{M}_{FF} = \dot{M}_{FR} = 0 \tag{13}$$

The $Q$ and $R$ covariance matrices where selected heuristically through multiple iterations. The Q matrix essentially defines that the state matrix has a high level certainty for the wheelset models, with less certainty assigned to the bogie and vehicle dynamics due to the use of a half vehicle estimator model. The creep force and moment sections are assigned the highest level uncertainty due to the assumptions of Equation 13, where

$$Q = diag[1e^{-10},\ 1e^{-10},\ 1e^{-10},\ 1e^{-10},\ 1e^{-10},\ 1e^{-10},\cdots$$
$$\cdots 1e^{-10},\ 1e^{-10},\ 1e^{5},\ 1e^{5},\ 1e^{10},\ 1e^{10},\cdots$$
$$\cdots 1e^{5},\ 1e^{5},\ 1e^{20},\ 1e^{20},\ 1e^{20},\ 1e^{20}] \tag{14}$$

The measurement covariance $R$ is defined as

$$R = diag[1e^{-3},\ 1,\ 1e^{-3},\ 1,\ 1e^{-3},\ 1,\ 1e^{-3},\cdots$$
$$\cdots 1,\ 1e^{5},\ 1e^{5},\ 1e^{5},\ 1e^{5},\ 1e^{10},\ 1e^{10}] \tag{15}$$

where the wheelset measurement are scaled in relation to their variance. The bogie and vehicle measurements are again treated as a higher level uncertainty due to the use of a half vehicle estimator.

Figure 5(a) shows an example of the creep moment estimation for the front wheelset of the front bogie at the dry adhesion level. Visual inspection shows that the estimator is identifying the correct frequency content of the creep moment but is over estimating the gain of the signal. This is reinforced by the Sprague and Geers metric of the estimation shown in Table IV.

As with the open loop estimator the KBF estimator shows poor convergence to the creep moment of the front wheelset of the front bogie at the very low adhesion level, Figure 5(b). The KBF in this case has failed to account for any discrepancies in the modelling. This is again reinforced by the poor values provided by the Sprague and Geers metric in Table IV.

| Condition | Parameter | Gain | Phase | Combined |
|---|---|---|---|---|
| Dry | $M_{FF}$ | 0.8671 | 0.2493 | 0.9023 |
| Very low | $M_{FF}$ | 6.5419 | 0.3912 | 6.5536 |

TABLE IV
CREEP MOMENT ESTIMATION SPRAGUE AND GEERS METRIC

### C. Development areas

There are clearly discrepancies between the modelling as performed in the MATLAB/Simulink phase of the project and the Vampire simulation package. Currently work is on-going to determine the cause of these differences as it is still hoped that the creep force estimation method of low adhesion estimation offers a high performance and robust solutions once these initial issues are rectified.

## IV. NON-MODEL BASED COMPARISON TECHNIQUE

An alternative method to the model based KBF creep force estimation of the previous section and as proposed in [9] is the use of known adhesion level ‘training’ data sets and real time advanced comparison computational methods. The basic concept of the idea is shown in Figure 6, that of comparing data gathered when the adhesion levels are considered acceptable and comparing this to the current measured data to determine any changes. This method’s success rests upon a number of factors: that the vehicle being tested runs at consistent speed profile down the same section of track on numerous occasions (ideal for service vehicles); that the system excitation (the track irregularity) doesn’t vary too greatly with time; and that spatial data can be stored and recalled in an efficient and accurate manner. The advantage with this type of system is that the processing requirements are much reduced and the limiting computational factor is now essentially one of storage.

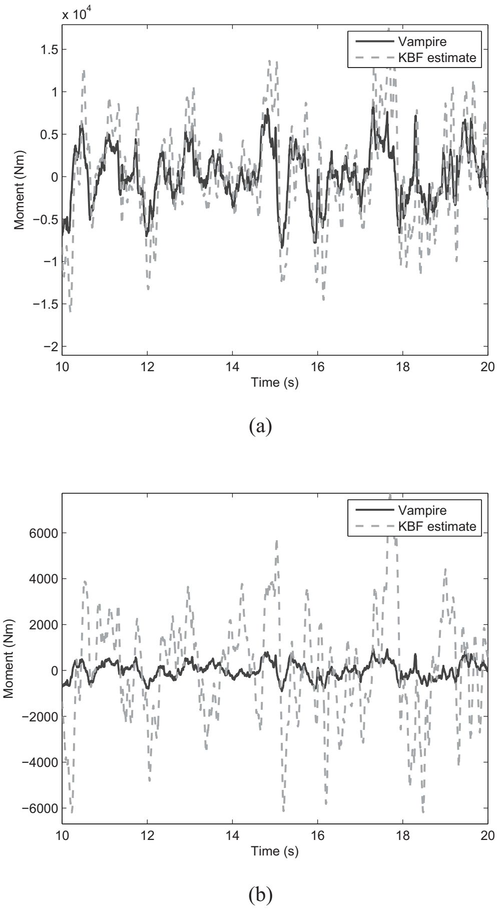


Fig. 5. KBF estimation of the creep moment using a half vehicle estimator for the front wheelset of the front bogie, (a) dry case, (b) very low

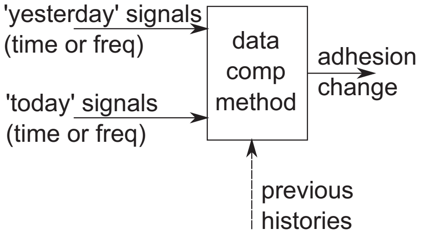


Fig. 6. Sensor signal comparison method

The processing algorithm used is the Sprague and Geers metric, [8]. In this simple demonstration of the method, the signal from an simulated lateral accelerometer mounted on the front bogie frame away from the centre of mass is used. Mounting on the bogie rather than the wheelset is much more advantageous due to the large accelerations (+/-300 $g$) experienced at the wheelset level. A resolution adequate for the application is assumed and the sample rate is 100 $Hz$.

### A. Threshold setting

The 'training' data is defined here as Vampire data run at the constant dry adhesion level and is compared to the constant adhesion data runs performed at the wet, low and very low levels. These are compared in 5-second sections of moving time window data, the size of which was determined in this simulation as to be sensitive enough to identify changes.

Figure 7 demonstrates the clearly defined levels for the gain metric of the Sprague and Geers metric and that they are almost linearly decreasing with the corresponding reduction in adhesion level. Simple thresholds can therefore be set around these levels to determine the current adhesion level.

It should be noted that such high quality data at the lower adhesion levels may not be available in application. However the technique will give a clear indication that the adhesion level has varied for a particular section of track.

### B. Step tests and signal delay

Vampire data was also created for a step change in the adhesion level from the dry condition to the very low at the half way point of the simulation at 30 seconds. The comparison to the 'training' data of the previous section is shown in Figure 7, which shows a clear reduction of the Sprague and Geers metric. Some time lag is evident due to the 5 second time windowing of the data, but this demonstrates that the technique can produce usable comparisons from signal based data alone.

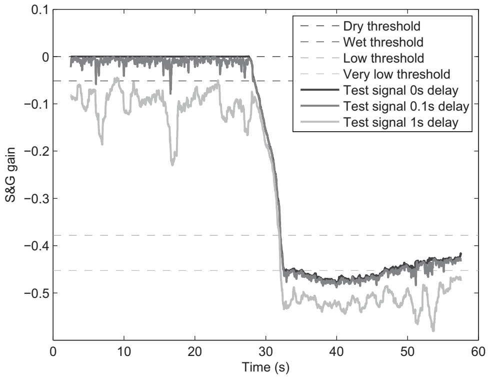


Fig. 7. Data comparison method adhesion threshold setting and step reduction tests at 30 seconds, for 0s time delay, 0.1s time delay and 1s time delay

The method will rely upon precise synchronisation of the 'training' and test data sets. The robustness of the algorithm was checked via delaying the 'training' data. This is also shown in Figure 7 for a 0.1 second and 1 second delay. The first case can be still seen clearly to demonstrate correlation though the quality is reduced. In the second case there is now a significant misalignment in the data, but the algorithm copes, not as clearly defining variations in adhesion but clearly demonstrating changes.

### C. Fuzzy logic reasoning

Multiple signals offer the opportunity for more comparisons of the changes in the dynamics of the vehicle, once thresholds are set this essentially becomes a problem of data fusion. This can be accomplished through simple logic processes or more complex fuzzy logic reasoning the basic architecture of which is a current being developed.

## V. Conclusion

Low adhesion causes punctuality and safety issues to rail operators and users alike due to vehicles failing to stop at stations or vehicles passing signals at danger. The RSSB managed project T959 is developing a number of practical processing options to determine the level of adhesion on in-service vehicles using relative modest cost sensors. The techniques' efficacies are now being tested on more representative modelling data from the multi-bodies physics simulation package Vampire. To date the Kalman-Bucy filtering method of estimating creep forces has proven to work at the higher ends of the adhesion spectrum but fails to converge at the low adhesion levels, which is mostly likely due to a filter model mismatch and research is continuing to resolve the problem. A non-model based pragmatic data comparison method utilising the Sprague and Geers metric has so far proven positive in the estimation of low adhesion provided high quality comparison data is available and that any signal phase is within acceptable limits.

## Acknowledgment

The authors would like to thank RSSB who manage project T959 with funding provided from the Rail Industry Strategic Research Programme. Thanks also goes to DeltaRail for the production of the Vampire simulation data.

## References

[1] DeltaRail: Vampire software, http://vampire-dynamics.com/, accessed 11th April 2012

[2] J. Kalker, On the rolling contact of two elastic bodies in the presence of dry friction, PhD Thesis, Delft University of Technology, Delft, Netherlands, (1967).

[3] R. Kalman. A new approach to linear filtering and prediction, Transactions of the ASME Journal of Basic Engineering, volume 82 (series D), pp. 35-45, (1960).

[4] L. Ljung, *System Identification: Theory for the User*, Prentice Hall, 1999.

[5] T. Pearce and K. Rose, Measured force-creep relationships and their use in the vehicle response calculation, In proceedings of the 9th IAVSD symposium, Linkoping, (1985).

[6] O. Polach, Creep forces in simulation of traction vehicles running on adhesion limit, Wear, volume 258(1), pp 992-1000, 2005.

[7] RSSB: Research and development, http://www.rssb.co.uk/RESEARCH/, accessed 12th April 2012

[8] L.E. Schwer, Validation metrics for response histories: perspectives and case studies, Engineering with Computers, 2007

[9] C.P. Ward, R.M. Goodall and R.Dixon. Use of real time creep force estimation data for assessment of low adhesion in the wheel/rail contact, In proceedings of the The 5th IET conference on Railway Condition Monitoring and Non-Destructive Testing 29-30 November 2011, Derby Conference Centre(2011).

[10] C.P.Ward, R.M.Goodall, R.Dixon, D.Fletcher, S.Lewis and G.Charles, RSSB Project T959 On-Board Detection of Low Adhesion, Interim Research Report, June 2011

[11] A.H. Wickens. Fundamentals of rail vehicle dynamics, Guidance and Stability, Swets & Zeitlinger, Lisse, Netherlands, 2003

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# A Novel Collaboration Compensation Strategy of Railway Power Conditioner for a High-Speed Railway Traction Power Supply System

Chenmeng Zhang(Student), Baichao Chen,
Chao Cai, Mengkui Yue, Cuihua Tian, Bo
Chen, Jiaxin Yuan*
(*Corresponding Author)
Wuhan University
Wuhan, Hubei, China

Jiabin Jia
University of Leeds
Leeds, UK

*Abstract*— **High-speed train traction power supply system causes serious negative current problem. Railway power conditioner (RPC) is efficient in negative sequence compensation. A novel power quality collaboration compensation system and strategy based on RPC is proposed in this paper. The minimum capacity conducted is 1/3 smaller than traditional single station compensation. Simulation results have confirmed that the collaboration compensation system proposed can achieve a good performance at the negative sequence compensation with capacity and cost efficient.**

***Keywords-RPC; Collaboration compensation; Unbalance compensation; Minimum capacity***

## I. INTRODUCTION

With the rapid development of high-speed railway in China, power quality has become a major concern for traction supply system [1]. Compared with normal electrification railway locomotive load, high-speed locomotive load has some characteristics, such as big instantaneous power, high power factor, low harmonic components and high negative sequence component. A large amount of negative current is injected into grid [2], which causes serious adverse impact on power system, such as increasing motor vibration and additional loss, reducing output ability of transformers and causing relay protection misoperation [3]. These adverse impacts threaten the safety of high-speed railway traction supply system and power system. Therefore, it’s necessary to take measures to suppress negative current.

Many methods and power quality compensators are studied in order to solve the issue of power quality. The traditional methods adopted to suppress negative current are as follows: (1) Connect unbalanced load to different supply terminals;(2) Adopt phase sequence rotation to make unbalanced load distributed to each sequence reasonably;(3) Connect unbalanced load to higher voltage level supply terminals; (4) Use balanced transformers such as Scott transformer and impedance balance transformer [4]. These methods have some effects on reducing unbalance degree, but they are lack of flexibility and can't adjust dynamically.

Recent years, high-voltage, large-capacity Static Var Compensator (SVC), Active Power Filter (APF) and Static Compensator (STATCOM) have become focus on power quality compensation of electrified railway [5]-[7]. However, these methods all need high-voltage transformers which increase cost. APF is effective in suppressing harmonic currents in electrified railway but rarely used in negative sequence compensation [8]. An active power quality compensator (APQC) with a impedance-matching balance transformer or a Scott transformer is proposed in [9] to compensate negative-sequence current, harmonics and reactive current. Reference [10] and [11] put forward a proposal of Railway Power Conditioner (RPC), RPC can make comprehensive compensation of negative sequence components, harmonics and reactive power. Reference [12] carries a dual-loop control strategy in order to improve the control effect and performance of RPC. Taken into account the disturbance and variation of electrified railway environment, a recursive proportional-integral control based on fuzzy algorithm is adopted to realize a fast and smooth tracking to reference current. Reference [13] raises a method of setting up two groups of thyristor control reactors (TCR) and two groups of thyristor control 3rd harmonic wave filter besides RPC. The RPC is used to transfer active power; the reactive power is supplied by the TCR and the filter. These works prove that RPC is a effective way to solve the power quality problems in railway system. But the compensator capacity is still too big to make RPC into practice.

To reduce the high compensator capacity, this paper puts forward a new railway negative unbalance compensation system based on the thought of multiple RPC collaboration compensation. This method realizes a minimum compensation capacity which is strictly proved, which reduces 1/3 capacity compared with traditional single station RPC compensation method. The simulation results have verified the correctness of the method proposed in this paper.

## II. RPC Structure and Analysis of Compensation Principle

The structure of RPC is shown in Fig.1. Three phase 220kV voltage is stepped down into two single-phase power supply voltage at the rank of 27.5kV by V/V transformer. RPC is made of back-to-back voltage source converters and a common dc capacitor, which can provide stable dc-link voltage. Two converters are connected to secondary arms of V/V transformer by step down transformer. Two converters can transfer active power from one power supply arm to another, supply reactive power and suppressing harmonic currents.

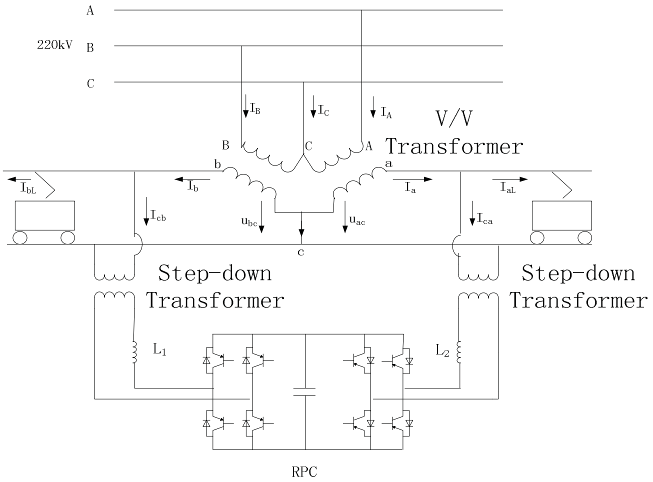


Figure 1. Traction power system with a three-phase V/V transformer and a RPC

The right feeder section in Fig.1 is denoted as *a*-phase power arm, while that the left side is *b*-phase power arm. The corresponding phases on the primary side are denoted as Phase A and Phase B, respectively. Since using four-quadrant pulse rectifiers to feed electrical locomotives, the power factor of high speed electrical locomotive is close to 1. Set $U_A$ as the reference value. Assume that the fundamental current vector of *a*-phase power arm is $\dot{I}_{aL}$ and the fundamental current vector of *b*-phase power arm is $\dot{I}_{bL}$ . $\dot{I}_{aL}$ and $\dot{I}_{bL}$ are shown as follows:

$$\begin{cases} \dot{I}_{aL} = I_{aL} e^{-j30^\circ} \\ \dot{I}_{bL} = I_{bL} e^{-j90^\circ} \end{cases} \tag{1}$$

The turns ratio of V/V transformer is *K*, so the three currents of the high-voltage side are shown as follows:

$$\begin{cases} \dot{I}_A = \dfrac{\dot{I}_{aL}}{K} = \dfrac{I_{aL}}{K} e^{-j30^\circ} \\ \dot{I}_B = \dfrac{\dot{I}_{bL}}{K} = \dfrac{I_{bL}}{K} e^{-j90^\circ} \\ \dot{I}_C = -(\dot{I}_A + \dot{I}_B) \end{cases} \tag{2}$$

Before RPC compensation, *a*-phase power arm has load current $\dot{I}_{aL}$ and the *b*-phase power arm has load current $\dot{I}_{bL}$ . Assume that $I_{aL} \geq I_{bL}$ , the three phase current is shown in Fig.2 .

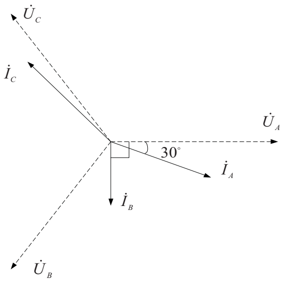


Figure 2. Three-phase current phase diagram without compensation

It is obvious that three phase current is unbalance before compensation. Use RPC to shift $\frac{1}{2}(I_{aL} - I_{bL})$ from *a*-phase to *b*-phase. Then, the current of two power arms are compensated to $I'_{aL}$ and $I'_{bL}$ , and they have an equal amplitude of $\frac{1}{2}(I_{aL} + I_{bL})$ and an angle difference of $\pi/3$ . The unbalance level is 50% now.

On the basis of active power transfer, RPC should compensate a certain quantity of capacitive reactive current $I_{caq}$ on the power arm *a* and a certain quantity of inductive reactive current $I_{cbq}$ on the power arm *b*, which can make the current of *a*-phase power arm lead the corresponding voltage $\pi/6$ . At this point, the reactive current should be calculated as follows:

$$I_{caq} = I_{cbq} = \frac{1}{2}(I_{aL} + I_{bL}) \tan 30^\circ \tag{3}$$

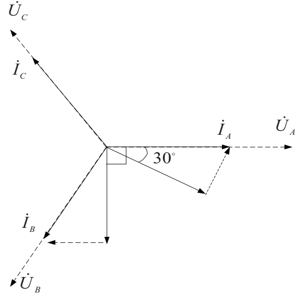


Figure 3. Three-phase current phase diagram after adjusting active and reactive power by RPC

After the compensation, the currents $I_A$ and $I_B$ have the same amplitude, as shown in Fig.3, and their angle difference is $2\pi/3$ . The C phase current $I_C$ can be obtained as $I_C = -I_A - I_B$ . The primary side of traction transformer has a balance three-phase current after active power shift and reactive power compensation. It is similar when $I_{aL} < I_{bL}$ . The common expression of RPC compensation current is:

$$\begin{cases} \dot{i}_{ca} = \dfrac{1}{2}(I_{bL} - I_{aL}) e^{-j30^\circ} + \dfrac{1}{2\sqrt{3}}(I_{aL} + I_{bL}) e^{j60^\circ} \\ \dot{i}_{cb} = \dfrac{1}{2}(I_{aL} - I_{bL}) e^{-j90^\circ} + \dfrac{1}{2\sqrt{3}}(I_{aL} + I_{bL}) e^{j180^\circ} \end{cases} \tag{4}$$

$\dot{I}_{ca}$ , $\dot{I}_{cb}$ --the equivalent current of RPC converters of *a*-phase arm and *b*-phase arm at the voltage of 27.5 kV.

## III. Principle of Collaboration Compensation

Since phase sequence rotation is widely adopted in traction power supply system, 3 stations collaboration compensation is mainly discussed in this paper. The structure of 3 stations collaboration compensation is shown in Fig.4.

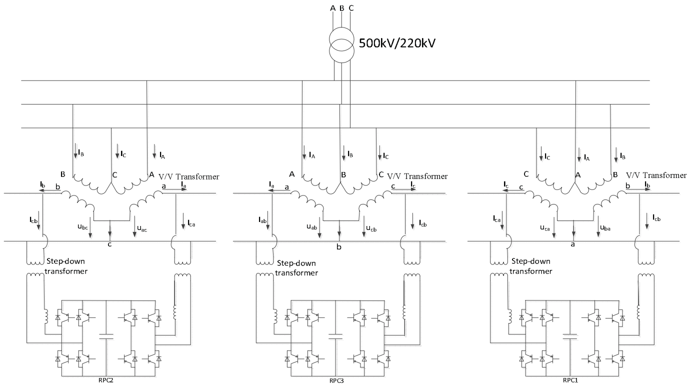


Figure 4. Schematic diagram of collaboration compensation of three stations

The capacity in phase CA, AB and BC is *x,y,z* , which has a relationship of *x>y>z*. The network of *x,y,z* can be divided into two parts, the one is a balanced network of *z,z,z* , the other is an unbalanced network of *x-z, y-z, 0*. Assume that $X = x - z$ , $Y = y - z$ , the original network is simplified as $X, Y, 0$ . Set *X*/2 as the reference value, the p.u. value of the simplified network is 2, $Y'$,0. $Y'$ is varying from 0 to 2.

The extreme case is $Y'$ =0. The optimize compensation strategy is shown below:

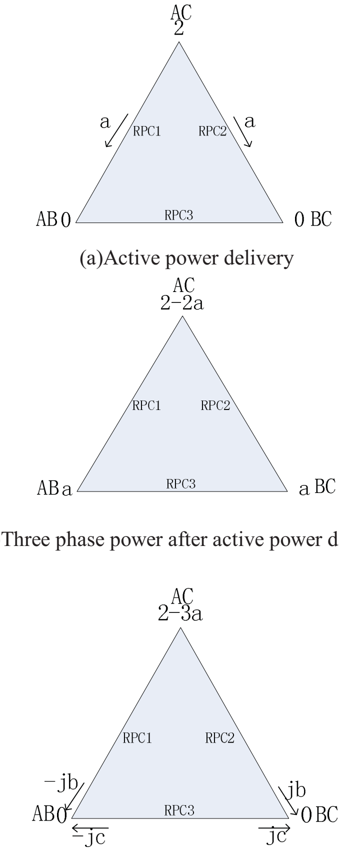
### *A. Single RPC compensation*

Based on the compensation strategy of RPC, when there is a maximum capacity in one of the traction feeder arms, RPC transfers $\frac{1}{2} * \frac{X}{2}$ active power from one traction feeder arm to another. And then compensates $\frac{1}{2\sqrt{3}} * \frac{X}{2}$ reactive power to both traction feeder arms based on Steinmetz theory. So the compensation capacity of single RPC is:

$$S = \sqrt{\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2\sqrt{3}}\right)^2} \frac{X}{2} = 0.2885X \quad (5)$$

### *B. Three stations collaboration compensation*

The simple model of 3 stations structure is shown in Fig.5. Since RPC could transfer a quantity of active power and compensate reactive power, a triangle is applied to illustrate the principle of collaboration compensation: apexes of the triangle are regarded as active load in Phase-AC, Phase-BC and Phase-AB, and edges of the triangle are regarded as three railway power conditioners. The arrows mean the delivery of active power (real part) and compensation of reactive power (imaginary part) . There are three steps to compensate. Firstly, transfer a quantity of active power. Secondly, separate the network into two parts: a balanced network and an unbalanced network. And last, make compensation to the unbalanced network based on the Steinmetz theory.



(a)Active power delivery

(b)Three phase power after active power delivery

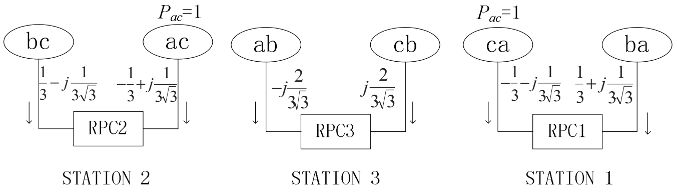
(c) Reactive power compensation based on Steinmetz theory

Figure 5. Compensation strategy under the condition of 2,0,0

According to the Steinmetz theory, fully compensation should satisfy the relationship of $b + c \geq \frac{2 - 3a}{\sqrt{3}}$. The capacity of three RPC is $\sqrt{a^2 + b^2}$ , $\sqrt{a^2 + b^2}$ , $c$ , separately. The installed capacity will be the maximum of the three RPC capacities above. So we can obtain the minimum installed capacity when $\sqrt{a^2 + b^2} = c \cdot$

The results can be conducted that $a = \frac{1}{3}, b = \frac{1}{3\sqrt{3}}$ , and the minimum capacity is $S_{\min} = \sqrt{a^2 + b^2} = c = \frac{2}{3\sqrt{3}}$ . This is a fully compensation but the station where RPC2 installed is capacitive. To avoid this condition, RPC1 supply inductive reactive power with the value of *b*, and RPC2 supply capacitive reactive power with the value of *b*, too. So the capacitive condition is avoided and the system keeps balance at the same time.

Working condition of three stations is shown in Fig.6. The ellipses stand for different traction feeder arms, the squares stand for RPC which connect to traction feeder arms. The arrows stand for active power transfer and reactive power compensation.



$P_{ac}$=1
$P_{ac}$=1
bc
ac
ab
cb
ca
ba
$\frac{1}{3}-j\frac{1}{3\sqrt{3}}$
$-\frac{1}{3}+j\frac{1}{3\sqrt{3}}$
$-j\frac{2}{3\sqrt{3}}$
$j\frac{2}{3\sqrt{3}}$
$-\frac{1}{3}-j\frac{1}{3\sqrt{3}}$
$\frac{1}{3}+j\frac{1}{3\sqrt{3}}$
RPC2
RPC3
RPC1
STATION 2
STATION 3
STATION 1

Figure 6. Working condition of three stations which supply active power and reactive power.

Three stations collaboration compensation minimum capacity is :

$$S_3=\sqrt{\left(1\big/3\right)^2+\left(1\big/3\sqrt{3}\right)^2}\,\frac{X}{2}=\frac{2}{3\sqrt{3}}*\frac{X}{2}=0.1925X\,,\quad(6)$$

which is 2/3 of the capacity of single RPC compensation. Tab.1 shows the compensation capacity of the two strategies.

TABLE I. COMPARISON OF TWO COMPENSATION METHOD

| **Compensation mode** | **Single station** | **Three station collaboration compensation** |
|---|---|---|
| RPC capacity | 0.2885*X* | 0.1925*X* |

It can be proved that this installed capacity (0.1925*X*) can satisfy any condition when *Y'* varying from 0 to 2.

If there is *N* stations connect to one 220kV bus, *N* may be 3*n*, 3*n*+1 or 3*n*+2 (*n*=0,1,2…). When *N*=3*n*, it means there are n sets of 3-stations compensation. When *N*=3*n*+1, it means there are n sets of 3-stations compensation and a single station compensation. When *N*=3*n*+2, it means there are *n* sets of 3-stations compensation and 2 single station compensation.

## IV. SIMULATION RESULTS

Simulation is done to proof the correctness of the theory by MATLAB/Simulink.

### A. Single RPC Compensation

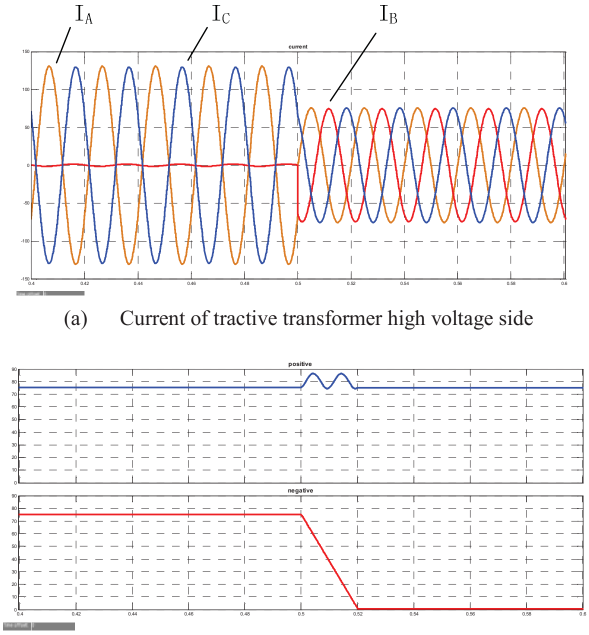
Assume the maximum load capacity appears at *a*-phase power arm, that is $P_{AC}=1$. The base capacity is $P_{base}=20MW$, and the short-circuit capacity is 750MVA. The power of *b*-phase locomotive load is 0. The *a*-phase load was switch on at 0*s*, the compensation system ran at 0.5*s*. The simulation schematic diagram is shown in Fig.1. The simulation parameters are as follows: three phase voltage of the system is 220kV; the frequency is 50Hz; the ratio of V/V transformer is 8:1; the ratio of step down transformer is 40:1; the capacitor of RPC at DC side is 100000 μF, and the value of $L_1$ and $L_2$ is 3mH and 2mH respectively.

Fig.7 (a) is the simulation current waveforms before and after three-phase negative sequence current compensation at 220kV side when locomotive load is under *a*-phase power arm. Fig.7 (b) is sequence analysis of current waveform. It can be seen from Fig.7 that before the compensation the current of Phase B $I_B$ is zero, and the phase current $I_A$ and $I_C$ have the same amplitude and an angle difference of 180° . Meanwhile, the negative sequence component is equal to positive sequence component. The unbalance level is defined :

$$\varepsilon_I=\left|\dot{I}_-\right|\Big/\left|\dot{I}_+\right|\times100\%\qquad(7)$$

$\left|\dot{I}_-\right|,\left|\dot{I}_+\right|$ --modulus of negative and positive current

The unbalance level before the compensation is 100%. The three phase currents become balanced after the compensator was carried out and the unbalance level was reduced to 0.



(a) Current of tractive transformer high voltage side

(b) Positive sequence and negative sequence current

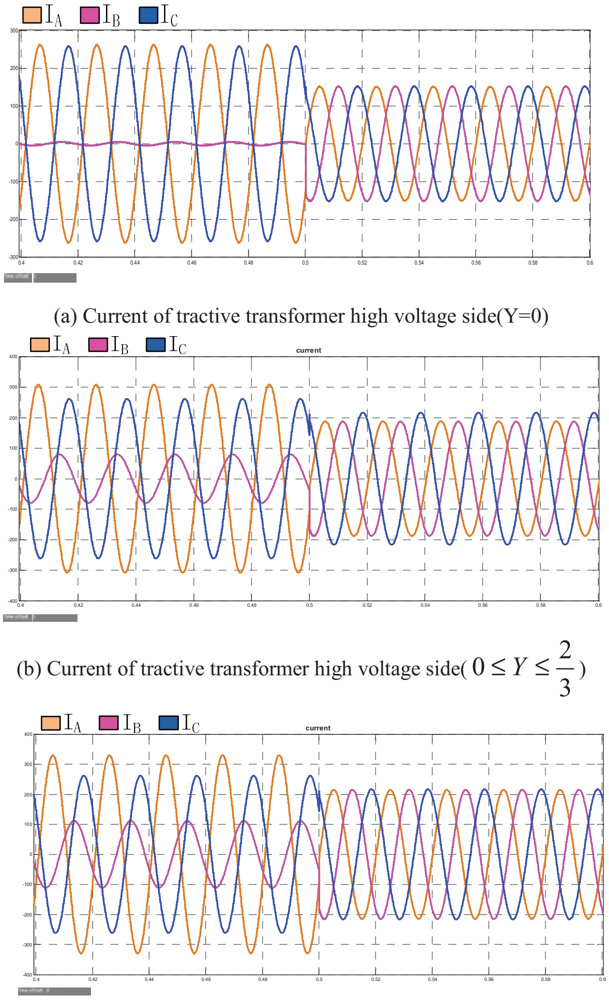
Figure 7. Compensation result under the condition of single station

### B. Three Station Collaboration Compensation

Set three station collaboration compensation for example. Fig.4 shows the schematic diagram of three station collaboration compensation. Three typical conditions are taken into consideration :

$$1)\mathrm{Y}{=}0;\ \ 2)\,0\le Y\le\frac{2}{3};\ \ 3)\,\frac{2}{3}\le Y\le1\,.$$

The simulation parameters are the same as single station compensation. Simulation results are shown in Fig.8. Fig.8 (a) shows the waveform before and after compensation when the maximum locomotive load appears at the Phase-AC at the condition of *Y*=0. The situation before compensation is almost the same as single station compensation, except for that the load is twice as much as single station locomotive load. With the use of compensator, the unbalance level was changed from 100% to 1%.Fig.8 (b) is the waveform when $0 \le Y \le \frac{2}{3}$, $Y$ appears at Phase AB. Compensator was put into operation at 0.5*s*. The unbalance level was reduced from 71% to 7%. The waveform when $\frac{2}{3} \le Y \le 1$ is shown in (c). The unbalance level was reduced from 60% to 2%. It can be seen from the simulation that there is a serious unbalanced condition before the compensation. The collaboration compensation network is effective in reducing the negative current(the unbalance level is reduced below 8%). The error may come from the loss of power electronic components and isolation transformers. The unbalance level before and after the compensation is listed in Tab. 2.



(a) Current of tractive transformer high voltage side(Y=0)

(b) Current of tractive transformer high voltage side( $0 \le Y \le \frac{2}{3}$ )

(c) Current of tractive transformer high voltage side( $\frac{2}{3} \le Y \le 1$ )

Figure 8. Three station collaboration compensation result under the condition of 2,Y,0

Table II. Unbalance Level Before And After Compensation

| | $Y$=0 | $0 \le Y \le \frac{2}{3}$ | $\frac{2}{3} \le Y \le 1$ |
|---|---|---|---|
| **Before Compensation** | 100% | 71% | 60% |
| **After Compensation** | 1% | 7% | 2% |

## V. Conclusion

This paper proposes a new power quality compensation system which is composed of several railway power conditioners. The proposed system can be used to compensate negative sequence current in high speed electrified railway. A minimum installed capacity is conducted which is 2/3 of the traditional single station compensation capacity. A new compensation strategy is raised Simulation results show that the proposed collaboration compensation of railway power conditioners is effective. It can reduce compensation capacity and has a good performance at negative sequence current compensation.

## Acknowledgment

The authors wish to express their gratitude to the National Foundation of China and school of electrical engineering of Wuhan University, for support of this research effort (National Natural Science Foundation of China under Grant 50807041).

## References

[1] X. Huang, L. Zhang, M He, X.You, and Q. Zheng, "Power electronics used in Chinese electrical locomotives", in Proc. IEEE 6th Int. Conf. Power Electron. Motion Control, pp.1196-1200，May, 2009, .

[2] S. L. Chen, R. J. Li, and P. H. Hsi, “Traction system unbalance problem-analysis methodologies,” IEEE Trans. Power Del, vol. 19, no. 4,pp. 1877–1883, Oct. 2004.

[3] B.Wang, X. Z. Dong, Z. Q. Bo, and A. Klimek, “Negative-sequence pilot protection with applications in open-phase transmission lines,” IEEE Trans. Power Del., vol. 25, no. 3, pp. 1306–1313, Jul. 2010.

[4] Z.W. Zhang, B.Wu, J. S. Kang, and L. F. Luo, “A multi-purpose balanced transformer for railway traction applications,” IEEE Trans. Power Del.,vol. 24, no. 2, pp. 711–718, Apr. 2009.

[5] P.-C. Tan, P. C. Loh, and D. G. Holmes, “A robust multilevel hybrid compensation system for 25-kV electrified railway applications,” IEEE Trans. Power Electron., vol. 19, no. 4, pp.

1043–1052, Jul. 2004.

[6] H. L. Ginn and G. Chen, “Flexible active compensator control for variable compensation objectives,” IEEE Trans. Power Electron., vol. 23, no. 6, pp. 2931–2941, Nov. 2008.

[7] M. Jianzong,W.Mingli, and Y. Shaobing, “The application of SVC for the power quality control of electric railways,” in Proc. Int. Conf. Sustainable Power Gener. Supply, pp. 1–4，2009.

[8] A. Luo, Z. K. Shuai, W. J. Zhu, and Z. J. Shen, “Combined system for harmonic suppression and reactive power compensation,” IEEE Trans.Ind. Electron., vol. 56, no. 2, pp. 418–518, Feb. 2009.

[9] Zhuo Sun, Xinjian Jiang, Dongqi Zhu, et al. “A novel active power quality compensator topology for electrified railway,” IEEE Trans. On Power Electron., vol.19, pp. 1036-1042, July, 2004.

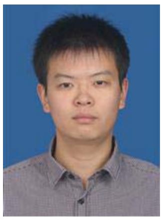
[10] Uzuka T，Ikedo S，Ueda K．A static voltage fluctuation compensator for AC electric railway[C]．Power Electronics Specialists Conference，Aachen，German, pp. 1869–1873, 2004.

[11] Morimoto H, Ando M, Mochinaga Y, et al. "Development of railway static power conditioner used at substation for Shinkansen,"[C]. Power Conversion Conference, Osaka, Japan ,pp. 1108–1111, 2002.

[12] Luo An，Fujun Ma，Chuanping Wu，Shi Qi Ding,"A dual-loop control strategy of railway static power regulator under V/V electric tranction system,"IEEE Trans. Power Electron.，vol. 26, pp. 2079-2090, 2011.

[13] Lu Fang, An Luo, Xiaoyong Xu, Houhui Fang,"A novel power quality compensator for negative-sequence and harmonic currents in high-speed electric railway, "Power and Energy Engineering Conference (APPEEC),pp. 1-5, 2011.

## BIOGRAPHIES



**Chenmeng Zhang** was born in Xiangyang in Hubei province, China, on October 1, 1988. He received the B.S.degree in the school of electrical engineering from Wuhan University, Wuhan, China in 2011.
From 2011, he is a graduate student for a Master's degree in Wuhan University, where he was engaged in research and development of railway power quality issues.

**Baichao Chen** received the B.Sc. degree in electrical engineering from the Huazhong University of Science and Technology, Wuhan, China, in 1982 and the M.Sc. and Ph.D. degrees from the College of Electrical Engineering, Wuhan University, Wuhan, in 1986 and 1993, respectively.
From 1998 to 1999, he was a Visiting Researcher with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY. He is currently a Professor of electrical engineering with the College of Electrical Engineering, Wuhan University. His main research interests include high-voltage engineering, power quality, and power electronic applications in high-voltage engineering.

**Jiaxin Yuan** was born in Nanchang in Jiang-xi province, China, on June 10, 1981. He received the B.S. and PH.D. degree in the school of electrical engineering from Wuhan University, Wuhan, China in 2002 and 2007 respectively.
From 2007 to 2009, he was a lecturer with Wuhan University, where he was engaged in research and development of STATCOM and DSP inverter control. in 2007 as, where he has been engaged in power electronics system control, power quality issues, application and control of inverters. In 2010, he was an Associate Professor in the school of electrical engineering of Wuhan University. Dr. Jiaxin is a member of IEEE

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# *Isochronal synchronization in complex networks*

## *The Lyapunov-Krasovskii theorem and stability in the network parameter space*

J. M. V. Grzybowski, T. Yoneyama
Divison of Computer Science and Electronic Engineering
ITA – Technological Institute of Aeronautics
São José dos Campos, Brazil
zzmariovic@yahoo.com.br, takashi@ita.br

E. E. N. Macau
Laboratory for Applied Mathematics and Computing
INPE – National Institute for Space Research
São José dos Campos, Brazil
elbert@inpe.lac.br

***Abstract*—Isochronal synchronization is a unique phenomenon in which physically distant oscillators wired together relax into zero-lag synchronous behavior over time. Such behavior is observed in natural processes and, recently, has been considered for promising applications in communication. Towards technological development of devices that explore isochronal sync, stability issues of the phenomenon need to be considered, both in the context of a pair or a network of coupled oscillators. This study concerns such stability issues by using the Lyapunov-Krasovskii theorem to propose a framework to study synchronization stability by using accessible parameters of the network coupling setup. As a result, relations between stability and network parameters are unveiled and the comprehension of roads leading to stability is enhanced.**

***Keywords: complex networks, isochronal synchronization, Lyapunov-Krasovskii.***

## I. INTRODUCTION

The communication among physically distant entities is subject to time delays. This is a result of the finite time a signal requires to travel through a physical media the distance from an emitter to a receiver. Curiously, despite of such condition, chaotic oscillators were revealed to overcome time delay and synchronize with zero-lag [1-10]. Under bidirectional coupling and adequate conditions, coupling setups of chaotic oscillators are somehow capable of absorbing coupling delays and synchronize as if no delays were present in the communication process at all [1, 8].

The knowledge of such fact brought promising applications for chaos in communication, such as those presented in [1, 8]. In this context, a fundamental requirement for communication is the stability of isochronal synchronization under small disturbances [6, 7]. This is so because information itself is introduced in the communication process as a disturbance to the synchronous dynamics [1]. As such, the communication process per se is a sequence of desync and resync episodes which can be understood as the transmission of bits through the coupling link between chaotic oscillators [1]. Resync, specifically, is possible if sync is stable, and the need for stability is justified.

In the scenario featuring pairs of delay-coupled oscillators, analytical results may establish frameworks that allow the systematic study of stability of isochronal sync in rather straightforward manners. This paper presents analytical results and their use towards the comprehension of the relations between network sync stability and the network parameters. The phenomenon of isochronal sync is considered, which implies the presence of time delays in the coupling among the oscillators. On this basis, the network parameter space is swept in a search process to map regions of sync stability. The relation between the number of nodes of the network and the number of links among its nodes is shown to be critical to the stability of isochronal sync, as it was observed to be for other types of synchronization as well [12].

The paper is organized as follows: section 2 presents the theoretical basis upon which the developments and explorations are based; section 3 explores the theoretical results towards the understanding of the underlying mechanisms of sync stability; section 4 presents some final remarks.

## II. ISOCHRONAL SYNC STABILITY OVER THE NETWORK PARAMETER SPACE

### *A. Isochronal sync and the network error equations*

Consider the systems

$$\dot{x}_i(t) = Ax_i(t) + g(x_i(t)) + u_i(t) \tag{1}$$

where $x(t) \in R^n$ is a state vector, $A \in R^n \times R^n$ is a constant matrix and $g(.) \in R^n \to R^n$ is a continuous vector field, such that the nodes are coupled exclusively by the control function $u_i(t)$ which is a feedback control function given by

$$u_i(t) = \tfrac{1}{G_{ii}} K \left( G_{ii} x_i(t) - \sum_{\substack{j=1 \\ j \neq i}}^{N} G_{ij} x_j(t-\tau) \right) \tag{2}$$

and $G_{ij}$ are entries of the Laplacian matrix $G$ and $K = K^T$ is a feedback matrix to be designed. The symmetry property of the feedback matrix is desirable as it allows some simplifications later on the development. Note that equation (2) can be rewritten as

$$u_i(t) = KC_{ii} \sum_{j=1}^{N} G_{ij} x_j(t-\tau_{ij}) \tag{3}$$

and the equations for the *ith* node of the network ( $i = 1, 2, ..., N$ ) as

$$\dot{x}_i(t) = Ax_i(t) + g(x_i(t)) + KC_{ii}\sum_{j=1}^{N} G_{ij}x_j(t-\tau_{ij}) \qquad (4)$$

where $\tau_{ij} = \tau$, for $i \neq j$, $\tau_{ij} = 0$, otherwise, and

$$C_{ii} = -\frac{1}{G_{ii}} \qquad (5)$$

for node balance. Alternatively, collecting the *ith* term of the summation (3), the self-coupling, for which $\tau = 0$, and reorganizing the terms of equation (4), it now reads

$$\dot{x}_i(t) = (A-K)x_i(t) + g(x_i(t)) + KC_{ii}\sum_{\substack{j=1 \\ j\neq i}}^{N} G_{ij}x_j(t-\tau) \quad (6)$$

In the following, the state vectors of the network nodes, $x_1(t), x_2(t), ..., x_N(t)$ are collected into the network state vector $X(t) \in \mathbb{R}^{Nn}$, given by

$$X(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_N(t) \end{pmatrix} \qquad (7)$$

Thus, the equations of the delay-coupled network can be written in the compact form

$$\dot{X}(t) = I_N \otimes (A-K)X(t) + \bar{G}(X(t)) + CA_d \otimes KX(t-\tau) \quad (8)$$

where $I_N$ is an *N-dimensional* identity matrix, ⊗ is the Kronecker product,

$$\bar{G}(X(t)) = \begin{pmatrix} g(x_1(t)) \\ \vdots \\ g(x_N(t)) \end{pmatrix} \qquad (9)$$

is a nonlinear vector field and $A_d$ is the network adjacency matrix that assigns $A_{d_{ij}} = A_{d_{ji}} = -1$ if nodes $i$ and $j$ are connected, $A_{d_{ij}} = A_{d_{ji}} = 0$ otherwise and, $A_{d_{ii}} = 0$.

At this point, the dynamical equations of the network are available and the formulation of the synchronization problem requires the definition of an error vector function $e(t)$, such that, if $\|e(t)\| \to 0$ as $t \to \infty$, then the $N$ systems of the network are asymptotically synchronized. Towards that end, considering the definition of isochronal synchronization, one can define

$$e(t) = X(t) - X_s(t) \qquad (10)$$

where

$$X_s(t) = SX(t) \qquad (11)$$

being $S \in \mathbb{R}^{Nn} \times \mathbb{R}^{Nn}$ an appropriate vector coordinate transformation defined as $S = T \otimes I_n$ and being $T$ defined as

$$T = \begin{pmatrix} T_{11} & T_{12} & \cdots & T_{1N} \\ T_{21} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ T_{N1} & \cdots & \cdots & T_{NN} \end{pmatrix} \qquad (12)$$

where either $T_{ij} = 1$ or $T_{ij} = 0$ and $\sum_i T_{ij} = \sum_j T_{ij} = 1$, such that $rank(T) = N$ and

$$rank(I_N - T) = N - 1 \qquad (13)$$

It can be verified that (10) yields

$$\begin{aligned} \dot{e}(t) \;=\; & \left(A-K+M_{(X(t),X_s(t))}\right)e(t) + \left(CA_d \otimes K - S[CA_d \otimes K]\right)X(t-\tau) \\ & \left(A-K+M_{(X(t),X_s(t))}\right)e(t) + \left(I_{Nn} - S\right)\left(CA_d \otimes K\right)X(t-\tau) \end{aligned} \qquad (14)$$

where $M_{(X(t),X_s(t))}e(t) = \bar{G}(X(t)) - S\bar{G}(X(t))$. Note that $S\bar{G}(X(t)) = \bar{G}(X_s(t))$.

At this point, note that system (14) depends both on the error states $e(t)$ and the network states $X(t)$, which is undesirable, since the synchronization stability evaluation must be performed in the error equations alone. However, considering equations (10), (11), one can rewrite systems (14) in terms of the error variable $e(t)$ by designing an adequate matrix of coefficients $E$ such that

$$\begin{aligned} (I_{Nn} - S)(CA_d \otimes K)X(t-\tau) &= E\left(X(t-\tau) - X_s(t-\tau)\right) \\ &= E(I_{Nn} - S)X(t-\tau) \\ &= Ee(t-\tau) \end{aligned} \qquad (15)$$

and, finally,

$$(I_{Nn} - S)(CA_d \otimes K) = E(I_{Nn} - S) \qquad (16)$$

Since the matrix $(I_{Nn} - S)$ is singular due to the definition of $S$, the equation cannot be solved for $E$ analytically. However, it is possible to prove that a solution exists, such that equation (14) can be rewritten in the form

$$\dot{e}(t) = I_N \otimes \left(A - K + M_{(X(t),X_s(t))}\right)e(t) + Ee(t-\tau) \qquad (17)$$

which depends only on the error variables $e(t)$. Note that $\|e(t)\| = 0$ implies $x_1(t) = x_2(t) = ... = x_N(t)$, which imply that the network is asymptotically synchronized. In other words, the stability of network synchronization requires that the error system (17) asymptotically establishes at the trivial fixed point.

Towards the determination of conditions for synchronization stability, consider the identity $e(t-\tau) = e(t) - \int_{t-\tau}^{t} \dot{e}(\theta)d\theta$, such that the error equations (17) can be rewritten as

$$\dot{e}(t) = I_N \otimes \left(A - K + M_{(X(t),X_s(t))}\right)e(t) + E\left(e(t) - \int_{t-\tau}^{t} \dot{e}(\theta)d\theta\right) \quad (18)$$

and, one step ahead,

$$\begin{aligned} \dot{e}(t) = & \left[I_N \otimes \left(A - K + M_{(X(t),X_s(t))}\right) + E\right]e(t) + \\ & -E\int_{t-\tau}^{t}\left[I_N \otimes \left(A - K + M_{(X(t),X_s(t))}\right)e(\theta) + Ee(\theta-\tau)\right]d\theta \end{aligned} \qquad (19)$$

such that the error system is obtained.

Notice that according to our formulation, the definition of the error system dismisses the use of reference signals. It follows that the asymptotic stability of the error system (19) means that of the synchronization of the network of oscillators

(8). In the next section, a stability criterion for the trivial fixed point of the error system (19) is derived, by means of the Lyapunov-Krasovskii stability theorem [11].

### B. Criterion for Isochronal sync stability

Consider the upper bound of the Lipschitz constant of the nonlinear vector field $g(.)$ over the invariant set $\Omega$ constituted by the trajectory of $x(t)$ to be given by $\ell$, such that $L=\ell I$ and $I$ is the $n$-dimensional identity matrix. The following result can be established [7]:

***Theorem [Isochronal synchronization of delay-coupled complex networks].*** *If there exists a constant matrix $P=P^T>0$ a positive constant $\varepsilon>0$ and a feedback gain matrix $K$ such that*

$$W=Q-2\varepsilon I_N\otimes A^T A-2\varepsilon E^T E-\frac{\tau}{\varepsilon}E^T P^2 E>0 \qquad (20)$$

*holds for a matrix $Q=Q^T>0$, where $Q$ is given by equation*

$$-Q=\left[I_N\otimes(A-K+L)+E\right]^T P+P\left[I_N\otimes(A-K+L)+E\right] \qquad (21)$$

*then the delay-coupled network whose error system given by equation (19) achieves isochronal synchronization for coupling delay $\tau$.*

## III. Investigating Isochronal Sync Stability

This section explores the relations between (i) synchronization stability and network topology, (ii) synchronization stability and coupling delays, based on the analytical results presented in [7]. The objective is to trace the influence of such network parameters in synchronization stability.

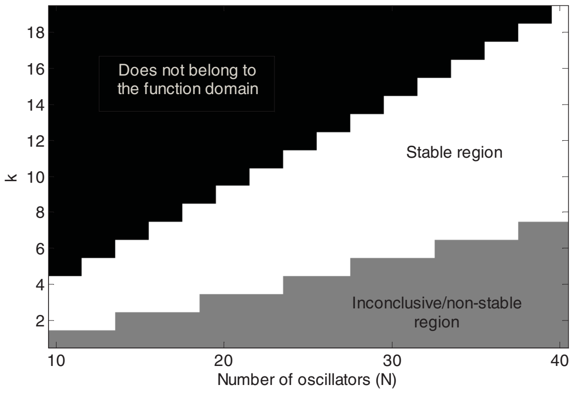


Figure 1 - Example of application of the network isochronal sync stability criterion as a stability function over the parameter domain: stability regions can be mapped from accessible network parameters. Since the stability criterion is sufficient (but not necessary), a the gray region is traced for the parameter values for which (20) is not satisfied.

### A. Sync stability versus network topology

In this example, the regions of stable isochronal synchronization are established for *k-cycle* networks on the basis of the inequality (20). Such regions are subsets of the cartesian product $N\times k$. A similar systematic approach can be used to trace such map for other kinds of network, as other parameters are chosen. To make the visual analysis simple, a $2-$ *dimensional* parameter meshgrid is generated. From Figure 1, Figure 2, Figure 3 and Figure 4, the loss of stability can be traced towards the identification of the parameters that are most influential to its occurrence. It can be recognized that smaller proportions of links among nodes relatively to the number of nodes in the network has negative effect on stability of isochronal sync. This effect is illustrated by arrows which show the direction of the loss of stability. Further analysis and insights into the nature of network stability are possible as other parameters are considered in the generation of the domain for evaluation of other similar stability functions resulting from the analytical results presented in this paper. In the following, similar studies considering stability under different values of coupling delay are considered.

### B. Time-delays versus number of links in k-cycle networks

At first, the form that the delay term appears in the inequality (20) suggests that stability is degraded for larger values of delay. Although this is not a general rule, the qualitative behavior of DDEs is shown to change as delays increase, due to the occurrence of Hopf bifurcations [13], which induce oscillatory and unstable behavior. The stability map in the parameter space $\tau\times k$ shows this relation, and it is illustrated in Figure 4 and Figure 5.

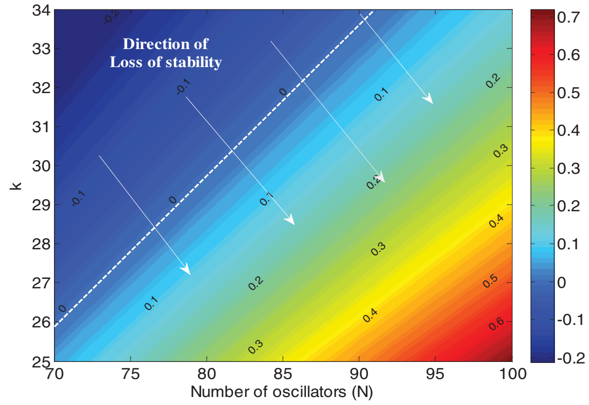


Figure 2 – Stability function over the set $N\times k$ of parameter space of the of the k-cycle network of Lorenz oscillators: the stability region has negative maximum eigenvalue of $-W$ and sync stability is lost as the number of oscillators increases and/or the number of links decreases.

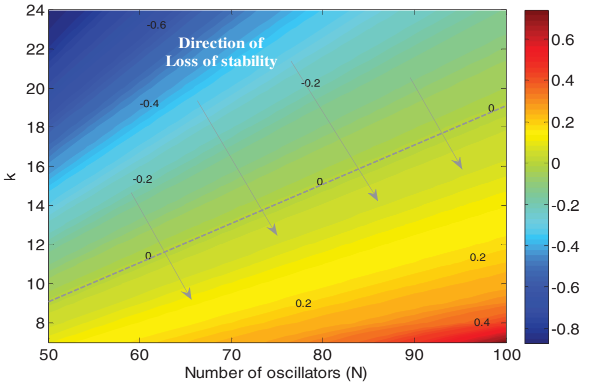


Figure 3 - Stability function over the set $N \times k$ of the parameter space of the k-cycle network of Rössler oscillators: the stability region has negative maximum eigenvalue of $-W$ and sync stability is lost as the number of oscillators increases and/or the number of links decreases.

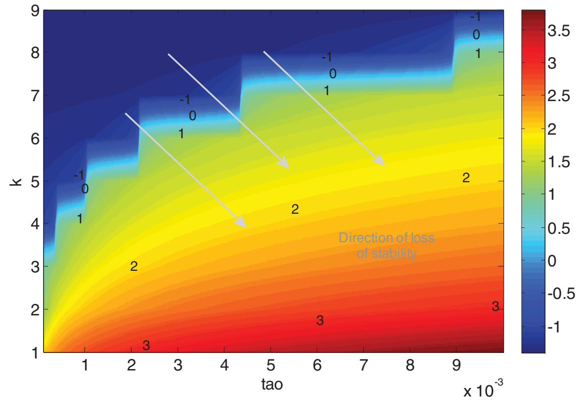


Figure 4 – Stability function over the set $\tau \times k$ of the parameter space of a k-cycle network with N = 20 Rössler oscillators: the region of stable sync has negative maximum eigenvalue of $-W$ . In this case, sync stability is lost as the number of links decreases and the time-delay increases.

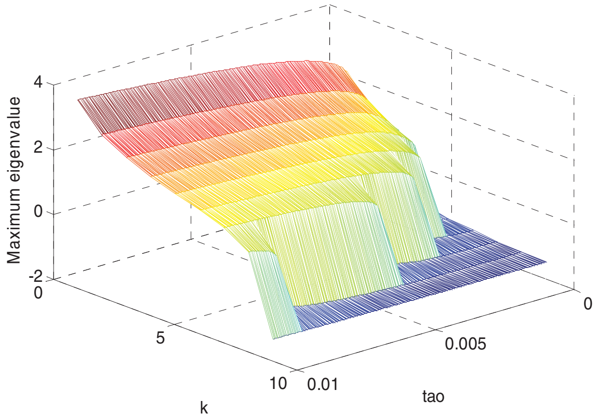


Figure 5 – Mesh of the stability function plotted in Figure 4: few links among nodes and larger values of time-delays render network synchronization less prone to synchronization.

### C. *Sync stability versus individual node dynamics*

A careful observation of the stability criterion (20) reveals that larger values of the Lipschitz $\ell$ constant also contribute negatively to stability. Moreover, it can be concluded that the substitution of the nonlinear terms of the nodes' dynamics with their upper bound, performed in the development of the stability criterion, is a source of restrictiveness, since it considers a worst-case scenario.

From a different perspective, this also illustrates the fact that some chaotic networks are more likely to synchronize than others, depending on the intrinsic dynamics of its nodes. Moreover, larger upper bounds of the nonlinear vector field $g$ over $x(t)$ reflect on the entries of the feedback matrix $K$ , which are supposed to lead the network nodes to isochronal synchronization, and thus counteract the effect of perturbations. Note that the value of $K$ also affects the delay term and, as such, may as well act as a perturbation to stability. Thus, not necessarily larger values of the entries of $K$ imply enhancement of stability, as observed also in non-delayed networks [12].

## IV. Final Remarks

The underlying mechanisms of network synchronization are topics of great interest within the context of complex networks. This paper presented analytical results that allow the determination of stability of isochronal synchronization in complex networks of delay-coupled chaotic oscillators. The relations between stability and the parameters of the network, as developed in this paper, help in the understanding of such mechanisms.

## Acknowledgment

The authors thank FAPESP and CPNq for the support on this research.

## References

[1] A. Wagemakers, J.M. Buldú, M.A.F. Sanjuán, Experimental demonstration of bidirectional chaotic communication by means of isochronal synchronization, *Europhys. Lett.* 81 (2008) 40005.

[2] E. Klein, N. Gross, M. Rosenbluh, W. Kinzel, L. Khaykovich, I. Kanter, Stable isochronal synchronization in mutually coupled chaotic lasers, *Phys. Rev. E* 73 (2006) 066214.

[3] A. Landsman, I. B. Schwartz, Complete chaotic synchronization in mutually coupled time-delay systems, *Phys. Rev. E* 75 (2007) 026201.

[4] A. Wagemakers, J.M. Buldu, and M.A.F. Sanjuan. Isochronous synchronization in mutually coupled chaotic circuits. *Chaos*, 17:023128, 2007.

[5] J.M.F. Avila and J.R. Rios Leite. Time delays in synchronization of chaotic coupled lasers with feedback. Opt. Lett., 17:21442–21451, 2009.

[6] J.M.V. Grzybowski, E.E.N. Macau, T. Yoneyama, Isochronal synchronization of time delay and delay-coupled chaotic systems, *J. Phys. A: Math. Theor.* 44 (2011) 175103.

[7] J.M.V. Grzybowski, E.E.N. Macau, T. Yoneyama. "A criterion for isochronal synchronization stability in complex networks", unpublished.

[8] B. B. Zhou, R. Roy, Isochronal synchrony and bidirectional communication with delay-coupled nonlinear oscillators, *Phys. Rev. E* 75 (2007) 026205.

[9] I. Kanter, E. Kopelowitz, W. Kinzel. "Public channel cryptography: chaos synchronization and Hilbert's tenth problem". Phys. Rev. Lett. 101 (2008) 084102.

[10] L. Illing, C. D. Panda, L. Shareshian. *Isochornal chaos synchronization of delay-coupled optoelectronic oscillators.* Phys. Rev. E 84 (2011) 016213.

[11] N. N. Krasovskii, *Stability of motion*, Stanford University Press, Chicago, 1963.

[12] Pecora, L. M., Carroll, T. L. *Master stability function for synchronized coupled systems.* Phys. Rev. Lett. 80 (1998) 2109-2112.

[13] T. Erneux. *Applied delay differential equations*. Berlim: Springer, 2009.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Synchronized tracking control of multiple Euler-Lagrange systems

Z-Jiang Yang and Yoshiyuki Shibuya
Department of Intelligent Systems Engineering
Ibaraki University, Hitachi, Ibaraki 316-8511, Japan
Email: yoh@mx.ibaraki.ac.jp

Pan Qin
Faculty of Mathematics, Kyushu University
744 Motooka, Nishi-ku 812-8581, Japan
pan@math.kyushu-u.ac.jp

***Abstract*—In this paper, we propose a distributed robust control method for synchronized tracking of multiple Euler-Lagrange systems, where the time-varying reference trajectory is sent to only a subset of the agents. It is assumed that the agents can exchange information with their local neighbors on an undirectionally connected communication graph. The controllers are not only distributed on the network, but also decentralized for each generalized coordinate within each agent. Theoretical analysis is performed. And simulation results are provided to support the theoretical results.**

## I. Introduction

Motivated by applications in physics, biology and engineering the study of synchronized control of collections of locally connected dynamic systems has become an important topic in control theory. Examples of interesting research directions include coverage control, consensus, formation control, flocking, and leader-follower tracking [1]. In recent years, there have been some remarkable works on synchronized tracking problem for multiple multiple Euler-Lagrange (EL) systems when only a portion of the agents can access the leader. In [2], a method of finite time synchronization tracking control of multirobot systems is proposed. The agent models are assumed to be known and each agent's controller requires its neighbors' control signals. In [3], a model-independent sliding mode control algorithm is proposed. However, the algorithm is discontinuous and requires the availability of the information of both the neighbors and the neighbors' neighbors. In [4], an adaptive robust control algorithm is proposed for multiple uncertain EL systems. In [5], the problem of position synchronization of multiple EL systems is studied. However, the proposed method considers the tracking of a stationary leader which sends a piece-wisely constant reference position signal.

In our recent work [6], we proposed a decentralized adaptive robust controller for trajectory tracking of robot manipulators. In this paper, the work of [6] is modified and extended to develop a new distributed robust control method for synchronized tracking of multiple EL systems, where the time-varying reference trajectory is sent to only a subset of the agents. It is assumed that the agents can exchange information with local neighbors on an undirectionally connected communication graph. In the local controller equipped in each generalized coordinate of each agent, a disturbance observer (DOB) is introduced to compensate for the low-passed coupled uncertainties, and a sliding mode control term is employed to handle the uncertainties that the DOB cannot compensate for sufficiently. By some damping terms, the boundedness of the signals of the overall multiple nonlinear systems is first ensured. Then we show how the DOB and sliding mode control play in a cooperative way in each coordinate to achieve an excellent synchronized tracking performance. Simulation results are provided to support the theoretical results.

## II. Background and problem statement

### A. Graph Theory

Consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$ with a finite set of $N$ nodes $\mathcal{V} = \{v_1, v_2, \cdots, v_N\}$ and a set of edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. Let $i$ denotes the $i$th agent. An edge of $\mathcal{E}$ is denoted as $e_{ij} = (v_i, v_j) \in \mathcal{E}$ where agent $j$ can receive information from agent $i$. In a directed graph, agent $j$ does not send information to agent $i$, whereas in an undirected graph, if $(v_i, v_j) \in \mathcal{E}$, then $(v_j, v_i) \in \mathcal{E}$. A graph is called connected if there exists a path between any two distinct agents. Denote the adjacency matrix as $\mathcal{A} = [a_{ij}] \in \mathcal{R}^{N\times N}$ with $a_{ij} > 0$ if $(v_j, v_i) \in \mathcal{E}$, and $a_{ij} = 0$ otherwise. Note $a_{ii} = 0$. For an undirected graph, we have $a_{ij} = a_{ji}$. The set of neighbors of a node $v_i$ is $\mathcal{N}_i = \{v_j \in \mathcal{V} | (v_j, v_i) \in \mathcal{E}\}$, i.e., the set of nodes with information incoming to $v_i$. The Laplacian matrix $\mathcal{L} = [l_{ij}] \in \mathcal{R}^{N\times N}$ is then defined as $l_{ii} = \sum_{j=1,j\neq i}^{N} a_{ij}$, and $l_{ij} = -a_{ij}, i \neq j$.

### B. EL System models

Consider $N$ agents governed by the following EL vector equations for $i = 1, \cdots, N$.

$$M_i(\theta_i)\ddot{\theta}_i + C_i(\theta_i, \dot{\theta}_i)\dot{\theta}_i + g_i(\theta_i) + f_i(\dot{\theta}_i) = u_i \qquad (1)$$

where $\theta_i = \theta_i(t) \in \Re^n$ is the generalized coordinate vector; $u_i \in \Re^n$ is the input torque vector; $M_i(\theta) = M_i^T(\theta) \in \Re^{n\times n}, M_i(\theta_i) > 0$ is the inertia matrix; $C_i(\theta_i, \dot{\theta}_i)\dot{\theta}_i \in \Re^n$ is the centrifugal and Coriolis torque; $g_i(\theta_i) \in \Re^n$ is the grativity torque; $f_i(\dot{\theta}) \in \Re^n$ is the friction force torque.

We first impose the following assumption.

*Assumption 1:* The reference trajectory $\theta_d(t) \in \mathcal{R}^n$ and the time derivatives $\dot{\theta}_d(t)$ and $\ddot{\theta}_d(t)$ are bounded signals.

Define an auxiliary error vector $r_i = [r_{i1}, \cdots, r_{in}]^T$ as

$$r_i = \dot{e}_i + \phi_i e_i \qquad (2)$$

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

where $e_i = \theta_i - \theta_d$ is the local tracking error vector, $\phi_i = \mathrm{diag}(\phi_{i1}, \cdots, \phi_{in}) > 0$ with constant entries.

Substituting $e_i$ and $r_i$ into (1), we have

$$M_i(\theta_i)\dot{r}_i + C_i(\theta_i, \dot{\theta}_i)r_i = u_i + \xi_i \tag{3}$$

where $\xi_i = [\xi_{i1}, \cdots, \xi_{in}]^T$ is considered to be an uncertain term of (3), and

$$\xi_i = -M_i(\theta_i)(\ddot{\theta}_d - \phi_i\dot{e}_i) - C_i(\theta_i, \dot{\theta}_i)(\dot{\theta}_d - \phi_i e_i) - g_i(\theta_i) - f_i(\dot{\theta}_i) \tag{4}$$

The global dynamics of the multiple EL systems is

$$M(\theta)\ddot{\theta} + C(\theta, \dot{\theta})\dot{\theta} + g(\theta) + f(\dot{\theta}) = u \tag{5}$$

$$\begin{aligned} M(\theta) &= \mathrm{diag}[M_1(\theta_1), .., M_N(\theta_N)] \\ C(\theta, \dot{\theta}) &= \mathrm{diag}[C_1(\theta_1, \dot{\theta}_1), .., C_N(\theta_N, \dot{\theta}_N)] \\ g(\theta) &= [g_1(\theta_1), \cdots, g_N(\theta_N)]^T \\ f(\dot{\theta}) &= [f_1(\dot{\theta}_1), \cdots, f_N(\dot{\theta}_N)]^T \\ u &= [u_1, \cdots, u_N]^T \end{aligned} \tag{6}$$

where $\theta = [\theta_1^T, \cdots, \theta_N^T]^T$.

And the global version of (3) is given as

$$M(\theta)\dot{r} + C(\theta, \dot{\theta})r = u + \xi \tag{7}$$

$$\xi = -M(\theta)(1_N \otimes \ddot{\theta}_d - \Phi\dot{e}) - C(\theta, \dot{\theta})(1_N \otimes \dot{\theta}_d - \Phi e) - g(\theta) - f(\dot{\theta}) \tag{8}$$

where $\Phi = \mathrm{diag}(\phi_1, .., \phi_N)$, $\xi = [\xi_1, \cdots, \xi_N]^T$ is the global uncertain term, $r = [r_1, \cdots, r_N]^T$ is the global auxiliary error vector, and $e = \theta - 1_N \otimes \theta_d$ is the global tracking error vector.

Then the following properties hold [7].

*Property 1:*

$$\mu_{\min}(M)I \leq M(\theta) \leq \mu_{\max}(M)I \tag{9}$$

where $\mu_{\max}(\cdot), \mu_{\min}(\cdot) > 0$ denote respectively the maximal and minimal eigenvalues of a matrix.

*Property 2:*

$$\| C(\theta, \dot{\theta}) \|_2 \leq c_H \| \dot{\theta} \|_2 \tag{10}$$

for some constant $c_H > 0$.

*Property 3:*

$$\| g(\theta) \|_2 \leq c_g, \quad \| f(\dot{\theta}_i) \|_2 \leq c_{f1} + c_{f2} \| \dot{\theta} \|_2 \tag{11}$$

for some constants $c_g, c_{f1}, c_{f2} > 0$.

*Property 4:*

$$x^T \left[ \frac{1}{2}\dot{M}(\theta) - C(\theta, \dot{\theta}) \right] x = 0, \quad {}^{\forall}x \neq 0 \tag{12}$$

### *C. Synchronized tracking problem*

The control objective is to design a controller for each agent to track a time-varying reference trajectory exerted by a leader, with the aid of the neighbor agents' information obtained by certain communication protocol. That is, $\| e_i \|_2$ and $\| \dot{e}_i \|_2$ $(i = 1, \cdots, N)$ should be controlled to be small.

To construct a feedback controller, we define the following local synchronization error vector of agent $i$ which will be used as a feedback signal:

$$e_{si} = \sum_{j \in \mathcal{N}_i} a_{ij}(\theta_i - \theta_j) + b_i(\theta_i - \theta_d) \tag{13}$$

where the scalar pinning gain $b_i \geq 0$. If agent $i$ receives information directly from the leader then $b_i > 0$, otherwise $b_i = 0$. $a_{ij}$ is the $(i, j)$ entry of the adjacency matrix $\mathcal{A}$ which defines the communication topology of the network.

Then the global synchronization error vector is given as $e_s = \mathcal{H}e$, where $\mathcal{H} = (\mathcal{L} + B) \otimes I_n$, $e_s = [e_{s1}^T, \cdots, e_{sN}^T]^T$, $B = \mathrm{diag}(b)$, $b = [b_1, \cdots, b_N]^T$.

We impose the following assumption on the communication topology of the network [4], [3].

*Assumption 2:* The communication graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$ of the multiple EL systems is undirected and connected.

Then, the following lemma is useful [3], [4].

*Lemma 1:* If the information interchange graph $\mathcal{G}$ is undirected and connected, and if at least one of the elements of $b$ is nonzero, then $\mathcal{L} + B$ is a positive definite symmetric matrix.

According to the aforementioned definitions, the global auxiliary synchronization error vector $r_s$ is given as

$$r_s = \mathcal{H}r = \mathcal{H}(\dot{e} + \Phi e) = \dot{e}_s + \Phi e_s \tag{14}$$

where $r_s = [r_{s1}, \cdots, r_{sN}]^T$, $r_{si} = [r_{si1}, \cdots, r_{sin}]^T$.

For the EL system model (3), it is well known that the uncertainty term is bounded by the following relation [8]:

$$\| \xi \|_2 \leq \alpha_1 + \alpha_2 \| e \|_2 + \alpha_3 \| \dot{e} \|_2 + \alpha_4 \| e \|_2 \| \dot{e} \|_2 \tag{15}$$

where $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ are some positive constants.

For a vector-valued signal $x(t)$, we define a truncated norm for $T > 0$ as $\| x \|_T \equiv \sup_{t \in [0,T]} \| x(t) \|_2$. Then we have [8]

*Lemma 2:* Let Assumption 1 and Properties 1∼4 hold. If there is a constant $T$ such that $\| r \|_T$ exists, then for all $t \in [0, T]$ we have for some $\beta_1, \beta_2, \beta_3 > 0$,

$$\| \xi \|_2 \leq \beta_1 + \beta_2 \| r \|_T + \beta_3 \| r \|_T^2 \tag{16}$$

## III. Controller design

### *A. Introduction of DOB*

Replacing $r$ in (7) by $r_s$, we have

$$\mathcal{H}^{-1}\dot{r}_s = M^{-1}u + M^{-1}(\xi - C(\theta, \dot{\theta})\mathcal{H}^{-1}r_s) \tag{17}$$

For the sake of design a distributed and decentralized controller $u_{ij}$ for each generalized coordinate, we define $\mathcal{H}_0^{-1}$ and $M_0$ as some diagonal nominal matrices of $\mathcal{H}^{-1}$ and $M$ respectively, and then we have

$$\begin{aligned} M_0\mathcal{H}_0^{-1}\dot{r}_s = {} & u + M_0M^{-1}(\xi - C(\theta, \dot{\theta})\mathcal{H}^{-1}r_s) \\ & + (M_0M^{-1} - I)u + M_0(\mathcal{H}_0^{-1} - \mathcal{H}^{-1})\dot{r}_s \end{aligned} \tag{18}$$

Let $M_0\mathcal{H}_0^{-1} = M_{s0} = \mathrm{diag}(m_{s01}, \cdots, m_{s0N})$ with $m_{s0i} = \mathrm{diag}(m_{s0i1}, \cdots, m_{s0in}) > 0$. Then we have the following vector-valued nominal linear system model.

$$M_{s0}\dot{r}_s = u + w \tag{19}$$

where $w = [w_1^T, \cdots, w_N^T]^T$ with $w_i = [w_{i1}, \cdots, w_{in}]^T$ is the global lumped disturbance vector expressed as

$$\begin{aligned} w = {} & M_0M^{-1}(\xi - C(\theta, \dot{\theta})\mathcal{H}^{-1}r_s) + (M_0M^{-1} - I)u \\ & + M_0(\mathcal{H}_0^{-1} - \mathcal{H}^{-1})\dot{r}_s \end{aligned} \tag{20}$$

Writing the dynamics of the $j$th generalized coordinate of the $i$th agent, we have

$$m_{s0ij}\dot{r}_{sij} = u_{ij} + w_{ij} \tag{21}$$

where $w_{ij}$ is considered as a lumped disturbance term. Since calculation of $\dot{r}_{sij}$ by direct differentiation is usually contaminated with high frequency noise, we may pass $w_{ij}$ through a low-pass filter to obtain its estimate as

$$Q_{ij}(s)w_{ij} = Q_{ij}(s)(m_{s0ij}sr_{sij} - u_{ij}) \tag{22}$$

In this paper, for convenience of expression, $s$ denotes not only the Laplace operator, but also a differential operator. This is the so-called DOB studied extensively in the literature. In this a study, we adopt a simple second-order filter

$$Q_{ij}(s) = \frac{1}{(1+\lambda_{ij}s)^2} \tag{23}$$

where $\lambda_{ij} > 0$.

However, we can only expect $Q(s)w_{ij} \approx w_{ij}$ at low-frequencies due to limited pass-band of the DOB. Moreover, the DOBs' outputs $\widehat{w}_{ij}(i = 1, \cdots, n,\ j = 1, \cdots, N)$ may disturb the signals of the other generalized coordinates. To ease the analysis shown later, a straightforward and simple idea is to saturate the output of the DOB as

$$\widehat{w}_{ij} = \begin{cases} \overline{w} & \text{for } |Q_{ij}(s)(m_{s0ij}sr_{ij} - u_{ij})| \geqslant \overline{w} \\ Q_{ij}(s)(m_{s0ij}sr_{ij} - u_{ij}) & \\ & \text{for } |Q_{ij}(s)(m_{s0ij}sr_{ij} - u_{ij})| < \overline{w} \\ -\overline{w} & \text{for } |Q_{ij}(s)(m_{s0ij}sr_{ij} - u_{ij})| \leqslant -\overline{w} \end{cases} \tag{24}$$

where $\overline{w} > 0$ is a selected upper bound of $|\widehat{w}_{ij}|$. Usually, it is recommended to choose a sufficiently large $\overline{w}$. However, even when $\overline{w}$ is not so large such that $\widehat{w}_{ij}$ is really saturated and hence the estimation error $(w_{ij} - \widehat{w}_{ij})$ is not sufficiently small, the control performance is still satisfactory, owing to the sliding mode control term included in the local controller (25) given later. The key point is that the DOB and the sliding mode control term work in a cooperative manner as suggested in [6]. Owing to their cooperative effects, the problems of high-gain or chattering can be avoided. This will be confirmed later by the numerical examples.

### *B. Description of the controller*

Motivated by the aforementioned discussions, we design the following local controller $u_{ij}$ for the $j$th generalized coordinate of agent $i$, using only the neighbor information of agent $i$ to ensure the boundedness of the global system signals and to achieve a satisfactory control performance.

$$u_{ij} = -\rho_a l_1^2 r_{sij} - \rho_b \overline{w} r_{sij} - \rho_c \eta_{\max} r_{sij} - \widehat{w}_{ij} - \eta_{ij}\text{sat}(r_{sij}) \tag{25}$$

where $k, \rho_a, \rho_b, \rho_c, l_1, \eta_{ij} > 0$, $\text{sat}(r_{sij}) = r_{sij}/(|r_{sij}| + \delta_{ij})$, $\delta_{ij} > 0$, $\eta_{\max} = \max(\eta_{11}, \cdots, \eta_{1n}, \cdots, \eta_{N1}, \cdots, \eta_{Nn})$.

The controller is explained as follows.

The damping term $-\rho_a l_1^2 r_{sij}$ is adopted to suppress the effects of neglected uncertainties of the global system model summarized as $\xi$ in (8). The constant $l_1$ is chosen such that $l_1 > \| r \|_2 / \sqrt{Nn}$ for all $t > 0$. It will be shown that there exists such an $l_1$. The term $-\widehat{w}_{ij}$ is a compensation term by DOB for each generalized coordinate. Since the DOBs' outputs $\widehat{w}_{ij}(i = 1, \cdots, N,\ j = 1, \cdots, n)$ may disturb mutually, to suppress the interactions due to $\widehat{w}_{ij}$, we employ the damping term term $-\rho_b \overline{w} r_{sij}$. The last term of $u_{ij}$ is a smoothed version of sliding mode control term. The damping term $-\rho_c \eta_{\max} r_{sij}$ is a term to suppress the interactions among the sliding mode control terms.

The global expression is given as below which will be used for analysis of the global system.

$$u = -\rho_a l_1^2 r_s - \rho_b \overline{w} r_s - \rho_c \eta_{\max} r_s - \widehat{w} - \eta\text{Sat}(r_s) \tag{26}$$

where $\widehat{w} = [\widehat{w}_1, \cdots, \widehat{w}_N]^T$, $\widehat{w}_i = [\widehat{w}_{i1}, \cdots, \widehat{w}_{in}]^T$, $\eta = \text{diag}[\eta_1, \cdots, \eta_N]^T$, $\eta_i = \text{diag}[\eta_{i1}, \cdots, \eta_{in}]^T$, $\text{Sat}(r_s) = [\text{sat}(r_{s1}), \cdots, \text{sat}(r_{sN})]^T$, $\text{sat}(r_{si}) = [\text{sat}(r_{si1}), \cdots, \text{sat}(r_{sin})]^T$.

*Remark 1:* The controller (26) is an extension or modification of the decentralized controller for a single EL system [6]. Compared to the controller in [6] where some nonlinear damping terms with signal dependent gains are used, in the present controller, we have to use some linear damping terms with relatively high constant gains, such as $\rho_a l_1^2 r_s$, $\rho_b \overline{w} r_s$ and $\rho_c \eta_{\max} r_s$. This is mainly due to the presence of the matrix $\mathcal{H}$ in (30) given later. Therefore, as the price of multiple EL system control, the controller design is less flexible as the case of a single agent.

### *C. Comments and guidelines of parameter design*

The guidelines of parameter design are aummarized here based on the theoretical analysis given later.

The constant $l_1$ in (25) should meet the requirement that $l_1 > \| r \|_2 / \sqrt{Nn}$ for all $t > 0$. That is, we have to guess the upper bound of $\| r \|_2$. See Theorem 1 later.

The entries of $\phi_i$ that appeared in (2) should not be very large, since large values of them may lead to a very large $\| r \|_2^2$ which may violate the condition imposed on $l_1$.

A small smoothing factor $\delta_{ij}$ for $\text{sat}(r_{sij})$ in (25) leads to a small ultimate tracking error of the corresponding generalized coordinate. However, as well known in the literature, a less smooth switching function may cause the chattering problem. A high sliding mode control gain $\eta_{ij}$ helps to achieve a small control ultimate tracking error, but it may also cause the chattering problem, and may cause a high gain control term $\rho_c \eta_{\max} r_{si}$.

The saturation level $\overline{w}$ of the DOBs should not be very large to avoid causing a high gain control term $\rho_b \overline{w} r_{si}$.

Usually, a smaller $\lambda_{ij}$ of the DOB filter (23) leads to a better disturbance performance. However, too small a $\lambda_{ij}$ may make $\widehat{w}_{ij}$ sensitive to the noise.

When $l_1^2$, $\overline{w}$, $\eta_{\max}$ and $\eta_{ij}$ meet the aforementioned requirements, the choice of $\rho_a$, $\rho_b$ and $\rho_c$ in (25) is trivial. Some moderate values of these parameters are satisfactory.

## IV. Performance analysis

Since all of the agents interact with their neighbors, we cannot easily see if the signals of the agents are all bounded. We should first ensure the boundedness of the global system signals. Then provided the boundedness of the global system signals, we can analyze the control performance of each agent. Therefore, the performance analysis includes two phases.

### A. Analysis of the global system

The results of analysis are given in Theorem 1. The proof is an extension of [8], [6], but with modifications specified by the newly designed controller in this study.

*Theorem 1:* Let Assumptions 1 and 2 hold. For the multiple EL systems (5) controlled by the proposed distributed robust controller (26), there exists a constant $l_1 > 0$, such that $r$ is bounded as $\| r \|_2 < \sqrt{Nn} l_1$ and hence all the internal signals are bounded, provided the following condition.

$$\sqrt{\frac{\mu_{\min}(M)}{\mu_{\max}(M)}}\sqrt{Nn}l_1 > \left[\frac{2Nn(\beta_1+\beta_2 l_1+\beta_3 l_1^2)}{\rho_a \mu_{\min}(\mathcal{H})}\right]^{1/3} \geq \left(\frac{N^2n^2\overline{w}}{2\rho_a\rho_b\mu_{\min}^2(\mathcal{H})}+\frac{N^2n^2\eta_{\max}}{2\rho_a\rho_c\mu_{\min}^2(\mathcal{H})}\right)^{1/4} \tag{27}$$

*Remark 2:* The first inequality of (27) is easily satisfied for sufficiently large $l_1$, $\rho_a$. The second inequality of (27) can be satisfied for sufficiently large $l_1$, $\rho_b$, $\rho_c$.

**Proof.** According to Assumption 2, we have Lemma 1 and hence $\mu_{\min}(\mathcal{H}) > 0$. We are now ready to show that there exists a constant $l_1 > 0$, such that $r$ is bounded as $\| r \|_2 < \sqrt{Nn} l_1$. The conclusion is proved by contradiction. To this end, according to (27) we first let a positive constant $l_1$ satisfy

$$\begin{aligned}\| r(0) \|_2 &< \left[\frac{2Nn(\beta_1+\beta_2 l_1+\beta_3 l_1^2)}{\rho_a \mu_{\min}(\mathcal{H})}\right]^{1/3}\\ &< \sqrt{\frac{\mu_{\min}(M)}{\mu_{\max}(M)}}\sqrt{Nn}l_1 \leq \sqrt{Nn}l_1\end{aligned} \tag{28}$$

Now assume the signal $r(t)$ is not bounded. Thus there always exists a smallest time $T_1$ such that $\| r(T_1) \|_2 = \sqrt{Nn} l_1$. Consider a Lyapunov function candidate with respect to the global tracking error vector.

$$V(t,r) = \frac{1}{2} r^T M(\theta) r \tag{29}$$

Taking the derivative along the trajectory of the closed-loop system, we have

$$\begin{aligned}\dot{V}(t,r) &= r^T\left(u+\xi-C(\theta,\dot{\theta})r+\frac{1}{2}\dot{M}(\theta)r\right)\\ &\leq r^T u + \| r \|_2 \| \xi \|_2\\ &= -\rho_a l_1^2 r^T\mathcal{H}r - \rho_b\overline{w}r^T\mathcal{H}r - r^T\widehat{w}\\ &\quad -\rho_c\eta_{\max}r^T\mathcal{H}r - r^T\eta\mathrm{Sat}(r_s) + \| r \|_2 \| \xi \|_2\\ &\leq -\rho_a l_1^2\mu_{\min}(\mathcal{H}) \| r \|_2^2 + \| r \|_2 \| \xi \|_2\\ &\quad -\rho_b\overline{w}\mu_{\min}(\mathcal{H}) \| r \|_2^2 + \| r \|_2 \| \widehat{w} \|_2\\ &\quad -\rho_c\eta_{\max}\mu_{\min}(\mathcal{H}) \| r \|_2^2 + \sqrt{Nn}\eta_{\max} \| r \|_2\end{aligned} \tag{30}$$

Here, $\eta_{\max}$ is the maximum diagonal element of $\eta$. Completing the squares, we have

$$\begin{aligned}\dot{V}(t,r) &\leq -\rho_a l_1^2\mu_{\min}(\mathcal{H}) \| r \|_2^2 + \| r \|_2 \| \xi \|_2\\ &\quad + \frac{\| \widehat{w} \|_2^2}{4\rho_b\overline{w}\mu_{\min}(\mathcal{H})} + \frac{Nn\eta_{\max}^2}{4\rho_c\eta_{\max}\mu_{\min}(\mathcal{H})}\end{aligned} \tag{31}$$

By (28), and the assumption that there exists a smallest time $T_1$ such that $\| r(T_1) \|_2 = \sqrt{Nn} l_1$, we have $\| r \|_2 < \sqrt{Nn} l_1$ for any $t < T_1$. Then using Lemma 2, we have for $t < T_1$,

$$\begin{aligned}\dot{V}(t,r) &\leq -\frac{\rho_a\mu_{\min}(\mathcal{H})}{Nn} \| r \|_2^4 + \| r \|_2 \left(\beta_1+\beta_2 l_1+\beta_3 l_1^2\right)\\ &\quad + \frac{Nn\overline{w}}{4\rho_b\mu_{\min}(\mathcal{H})} + \frac{Nn\eta_{\max}}{4\rho_c\mu_{\min}(\mathcal{H})}\\ &= -\frac{\rho_a\mu_{\min}(\mathcal{H})}{2Nn} \| r \|_2 \left(\| r \|_2^3 - \frac{2Nn(\beta_1+\beta_2 l_1+\beta_3 l_1^2)}{\rho_a\mu_{\min}(\mathcal{H})}\right)\\ &\quad -\frac{\rho_a\mu_{\min}(\mathcal{H})}{2Nn}\left(\| r \|_2^4 - \frac{N^2n^2\overline{w}}{2\rho_a\rho_b\mu_{\min}^2(\mathcal{H})} - \frac{N^2n^2\eta_{\max}}{2\rho_a\rho_c\mu_{\min}^2(\mathcal{H})}\right)\end{aligned} \tag{32}$$

We then can say that there exists a time instant $t_2 = T_1 - t_1 > 0,\ t_1 > 0$ such that

$$\begin{aligned}\sqrt{Nn}l_1 > \| r(T_1-t_1) \|_2 &= \left[\frac{2Nn(\beta_1+\beta_2 l_1+\beta_3 l_1^2)}{\rho_a \mu_{\min}(\mathcal{H})}\right]^{1/3}\\ &\geq \left(\frac{N^2n^2\overline{w}}{2\rho_a\rho_b\mu_{\min}^2(\mathcal{H})}+\frac{N^2n^2\eta_{\max}}{2\rho_a\rho_c\mu_{\min}^2(\mathcal{H})}\right)^{1/4}\end{aligned} \tag{33}$$

However, according to (32), we have $d/dtV(t) \leq 0$, for all $t \in [T_1 - t_1, T_1]$. Therefore, for all $t \in [T_1 - t_1, T_1]$, we have

$$\begin{aligned}V[T_1, r(T_1)] &\leq V[(T_1-t_1), r(T_1-t_1)]\\ &\leq \frac{1}{2}\mu_{\max}(M)\left[\frac{2Nn(\beta_1+\beta_2 l_1+\beta_3 l_1^2)}{\rho_a \mu_{\min}(\mathcal{H})}\right]^{2/3}\end{aligned} \tag{34}$$

But the definition of $T_1$ leads to

$$V_1[T_1, r(T_1)] \geq \frac{1}{2}\mu_{\min}(M)Nnl_1^2 \tag{35}$$

Clearly, the last two inequalities are in contradiction, according to (28). This implies that the assumption of $\| r(T_1) \|_2 = \sqrt{Nn} l_1$ is false. Thus the error signal vector $r$ is bounded and satisfies $\| r(t) \|_2 < \sqrt{Nn} l_1$ for all $t \geq 0$.

Furthermore, according to Assumption 1, and (2), (15) and (14), we conclude that $e$, $\dot{e}$, $\theta$, $\dot{\theta}$, $\xi$ and $r_s$ are bounded. And hence each local controller $u_{ij}$ is bounded. Therefore, all the internal signals are bounded. □

*Remark 3:* The condition (27) is always satisfied for a sufficiently large bound $l_1$. The results of Theorem 1 only tell us that $r$ and hence all the internal signals can be made to be bounded. It should be emphasized here that at the present stage our purpose is only to ensure the boundedness of the signals. And hence a conservative bound of the signals is acceptable. Later, we will show that the individual synchronization error $r_{sij}$ can be made sufficiently small by virtue of the corresponding local controller.

### B. Analysis of each agent

We are now ready to analyze how the DOBs and sliding mode control techniques bring improvement in each generalized coordinate. Substituting the local controller (25) into the subsystem (21), the resultant subsystem of the $j$th generalized coordinate of the $i$th agent becomes

$$\begin{aligned} m_{s0ij}\dot{r}_{sij} = &-\rho_a l_1^2 r_{sij} - \rho_b \overline{w} r_{sij} - \rho_c \eta_{\max} r_{sij} \\ &-\widehat{w}_{ij} - \eta_{ij}\frac{r_{sij}}{|r_{sij}| + \delta_{ij}} + w_{ij} \end{aligned} \tag{36}$$

Owing to the results of Theorem 1, $w_{ij}(t)$ and $\widehat{w}_{ij}(t)$ are bounded. Define

$$\eta^*_{ij,0} = \sup_{0\le\tau\le t} |w_{ij}(\tau) - \widehat{w}_{ij}(\tau)| \tag{37}$$

*Theorem 2:* Let the assumptions and results of Theorem 1 hold. The synchronization error of the $j$th generalized coordinate of the $i$th agent satisfies

$$|r_{sij}(t)| \le |r_{sij}(0)| e^{-\frac{c}{m_{s0ij}}t} + \sqrt{\frac{\delta_{ij}\eta^*_{ij,0}}{c}} \tag{38}$$

if $\eta_{ij} \ge \eta^*_{ij,0}$, or

$$|r_{sij}(t)| \le |r_{sij}(0)| e^{-\frac{c}{2m_{s0ij}}t} + \frac{(\eta^*_{ij,0} - \eta_{ij})}{c} + \sqrt{\frac{2\delta_{ij}\eta_{ij}}{c}} \tag{39}$$

if $0 \le \eta_{ij} < \eta^*_{ij,0}$, where $c = \rho_a l_1^2 + \rho_b \overline{w} + \rho_c \eta_{\max}$.

**Proof.** We first consider the case of $\eta_{ij} \ge \eta^*_{ij,0}$, i.e., the sliding mode control gain exceeds the maximum amplitude of $(w_{ij} - \widehat{w}_{ij})$. From (36), we have

$$\frac{d}{dt}\left(\frac{m_{s0ij} r_{sij}^2}{2}\right) \le -c r_{sij}^2 + \delta_{ij}\eta^*_{ij,0} \tag{40}$$

and hence

$$r_{sij}^2(t) \le e^{-\frac{2c}{m_{s0ij}}t} r_{sij}^2(0) + \frac{\delta_{ij}\eta^*_{ij,0}}{c} \tag{41}$$

This leads to (38).

In the case of $0 \le \eta_{ij} < \eta^*_{ij,0}$, we have

$$\begin{aligned} \frac{d}{dt}\left(\frac{m_{s0ij} r_{sij}^2}{2}\right) \le &-\frac{c}{2} r_{sij}^2 - \left(\sqrt{\frac{c}{2}}|r_{sij}| - \frac{\eta^*_{ij,0} - \eta_{ij}}{\sqrt{2c}}\right)^2 \\ &+ \frac{(\eta^*_{ij,0} - \eta_{ij})^2}{2c} + \delta_{ij}\eta_{ij} \end{aligned} \tag{42}$$

and hence

$$r_{sij}^2(t) \le e^{-\frac{c}{m_{s0ij}}t} r_{sij}^2(0) + \left(\frac{(\eta^*_{ij,0} - \eta_{ij})^2}{c_i^2} + \frac{2\delta_{ij}\eta_{ij}}{c_i}\right) \tag{43}$$

This leads to (39). □

However, $\eta^*_{ij,0}$ may not be small since the initial value $\widehat{w}_{ij}(0)$ is often set to be zero. To investigate the performance after a short transient phase of DOB. Let $t_{ij}(\lambda_{ij})$ be an effective time-constant of the DOB depending on $\lambda_{ij}$, until which the initial value of $(w_{ij} - \widehat{w}_{ij})$ has decayed out sufficiently such that for a relatively small constant $\eta^*_{ij,t_{ij}}$ we have

$$\eta^*_{ij,t_{ij}} = \sup_{t_{ij}(\lambda_{ij})\le\tau\le t} |w_{ij}(\tau) - \widehat{w}_{ij}(\tau)| \tag{44}$$

Comparing (37) and (44), it is expected that $\eta^*_{ij,t_{ij}}$ can be much smaller than $\eta^*_{ij,0}$. Then we have

*Corollary 1:* For $t \ge t_{ij}(\lambda_{ij})$, the synchronization error of the $j$th generalized coordinate of the $i$th agent satisfies

$$|r_{sij}(t)| \le |r_{sij}(t_{ij})| e^{-\frac{c}{m_{s0ij}}(t-t_{ij})} + \sqrt{\frac{\delta_{ij}\eta^*_{ij,t_{ij}}}{c}} \tag{45}$$

if $\eta_{ij} \ge \eta^*_{ij,t_{ij}}$, or

$$\begin{aligned} |r_{sij}(t)| \le &|r_{sij}(t_{ij})| e^{-\frac{c}{2m_{s0ij}}(t-t_{ij})} \\ &+ \frac{(\eta^*_{ij,t_{ij}} - \eta_{ij})}{c} + \sqrt{\frac{2\delta_{ij}\eta_{ij}}{c}} \end{aligned} \tag{46}$$

if $0 \le \eta_{ij} < \eta^*_{ij,t_{ij}}$, where $c = \rho_a l_1^2 + \rho_b \overline{w} + \rho_c \eta_{\max}$.

Theorem 2 and Corollary 1 imply that the auxiliary synchronization error $r_s$ is uniformly ultimately bounded (UUB), and hence the auxiliary tracking error $r = \mathcal{H}^{-1} r_s$ is UUB.

## V. Simulation studies

For the sake of comparison, we borrow the example in [3] and carry out the numerical simulations under the same conditions as possible. Consider a group of 6 two-DOF planar robot arms:

$$\begin{aligned} &\begin{bmatrix} m_{i11}(\theta_i) & m_{i12}(\theta_i) \\ m_{i21}(\theta_i) & m_{i22}(\theta_i) \end{bmatrix} \begin{bmatrix} \ddot{\theta}_{i1} \\ \ddot{\theta}_{i2} \end{bmatrix} + \begin{bmatrix} h_{i1}(\theta_i, \dot{\theta}_i) \\ h_{i2}(\theta_i, \dot{\theta}_i) \end{bmatrix} \\ &\qquad + \begin{bmatrix} g_{i1}(\theta_i) \\ g_{i2}(\theta_i) \end{bmatrix} = \begin{bmatrix} u_{i1} \\ u_{i2} \end{bmatrix} \end{aligned} \tag{47}$$

where $i = 1, \cdots, 6$, and

$$\begin{aligned} m_{i11}(\theta_i) &= m_{i1} l_{ci1}^2 + m_{i2}(l_{i1}^2 + l_{ci2}^2) + I_{i1} + I_{i2} \\ &\quad + 2 m_{i2} l_{i1} l_{ci2} \cos(\theta_{i2}) \\ m_{i12}(\theta_i) &= m_{i2} l_{ci2}^2 + I_{i2} + m_{i2} l_{i1} l_{ci2} \cos(\theta_{i2}) \\ m_{i21}(\theta_i) &= m_{i12} \\ m_{i22}(\theta_i) &= m_{i2} l_{ci2}^2 + I_{i2} \end{aligned} \tag{48}$$

$$\begin{aligned} h_{i1}(\theta_i, \dot{\theta}_i) &= -m_{i2} l_{i1} l_{ci2} (2\dot{\theta}_{i1}\dot{\theta}_{i2} + \dot{\theta}_{i2}^2) \sin(\theta_{i2}) \\ h_{i2}(\theta_i, \dot{\theta}_i) &= m_{i2} l_{i1} l_{ci2} \dot{\theta}_{i1}^2 \sin(\theta_{i2}) \\ g_{i1}(\theta_i) &= g(m_{i1} l_{ci1} + m_{i2} l_{i1}) \cos(\theta_{i1}) \\ &\quad + m_{i2} g l_{ci2} \cos(\theta_{i1} + \theta_{i2}) \\ g_{i2}(\theta_i) &= m_{i2} g l_{ci2} \cos(\theta_{i1} + \theta_{i2}) \end{aligned} \tag{49}$$

where $g = 9.807$[m/s$^2$], and the physical parameters are given in Table I.

The network topology for communication among the agents is shown in Fig. 1. It can be verified that agents 1∼6 are undirectionally connected, and only agents 3 and 6 have access

TABLE I
PHYSICAL PARAMETERS OF THE ROBOT MANIPULATORS ($i = 1, \cdots, 6$)

| | |
|---|---|
| link mass [kg] | $m_{i1} = 1.0 + 0.3i$, $m_{i2} = 1.5 + 0.3i$ |
| link length [m] | $l_{i1} = 0.2 + 0.06i$, $l_{i2} = 0.3 + 0.06i$ |
| mass center [m] | $l_{ic1} = 0.1 + 0.03i$, $l_{ic2} = 0.15 + 0.03i$ |
| inertial tensor [kg·m$^2$] | $I_{i1} = 0.0073, 0.0137, 0.0229,$ $0.0355, 0.0521, 0.0732$<br>$I_{i2} = 0.0194, 0.0309, 0.0461,$ $0.0656, 0.0900, 0.1198$ |

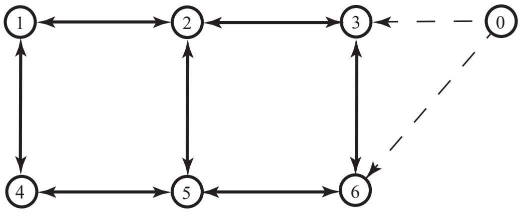


Fig. 1. Information exchange graph of the leader and followers

to the leader (agent 0). The corresponding adjacency matrix and pinning vector are given as follows.

$$\mathcal{A} = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}, \quad b = [0, 0, 1, 0, 0, 1]^T \tag{50}$$

We investigate the synchronized tracking performance for the following reference trajectory vector and its derivative generated by the leader.

$$\begin{aligned} \theta_d &= \left[\cos\left(\frac{2\pi}{60}t\right),\ \sin\left(\frac{2\pi}{60}t\right)\right]^T \quad \text{[rad]} \\ \dot{\theta}_d &= \left(\frac{2\pi}{60}\right)\left[-\sin\left(\frac{2\pi}{60}t\right),\ \cos\left(\frac{2\pi}{60}t\right)\right]^T \quad \text{[rad/s]} \end{aligned} \tag{51}$$

And to show that the controllers are robust against nonzero initial tracking errors, the initial conditions are given as

$$\begin{aligned} \theta_i(0) &= \left[\frac{\pi}{7}i,\ \frac{\pi}{8}i\right]^T \quad \text{[rad]} \\ \dot{\theta}_i(0) &= [0.05i - 0.2,\ -0.05i + 0.2]^T \quad \text{[rad/s]} \end{aligned} \tag{52}$$

According to the design guidelines, we choose the design parameters of the local controllers (25) as follows.

$$\begin{aligned} &\phi_{i1} = \phi_{i2} = 0.5 \\ &\rho_a = \rho_b = \rho_c = 1,\ \eta_{i1} = \eta_{i2} = 3 \\ &l_1 = 2,\ \delta_{i1} = \delta_{i2} = 0.05 \end{aligned} \tag{53}$$

where $i = 1, \cdots, 6$. The nominal values in (21) are given as $m_{s0i1} = m_{s0i2} = 1$. The time-constants of the DOB filters are given as $\lambda_{i1} = \lambda_{i2} = 0.02$. And the saturation level of the DOBs is chosen as $\overline{\widehat{w}} = 30$ (see (24)).

The simulation results are shown in Fig. 2, where from the top to the bottom are respectively the position-tracking errors $e_{i1}$, $e_{i2}$, auxiliary errors $r_{i1}$, $r_{i2}$, control signals $u_{i1}$, $u_{i2}$ and DOBs' outputs $\widehat{w}_{i1}$, $\widehat{w}_{i2}$, where the lines of magenta, cyan, red, green, black and blue represents the signals of agents 1∼6 respectively. It can be found in Fig. 1 that the proposed distributed controllers deliver a very excellent synchronized tracking performance, owing to the cooperative effects by DOBs and sliding mode control terms.

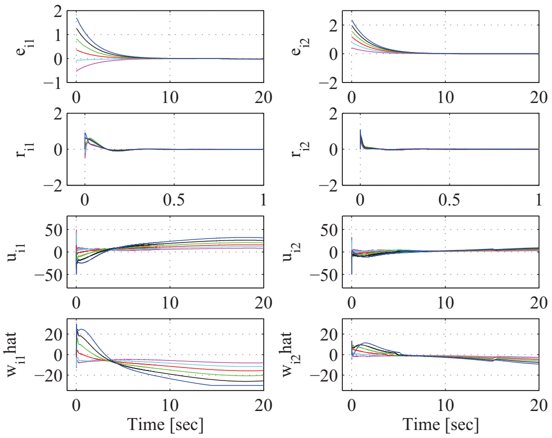


Fig. 2. Synchronized tracking results of 6 two-DOF planar robot arms.

## VI. CONCLUSIONS

In this paper, a distributed robust control method for synchronized tracking of multiple EL systems has been proposed. The problem setting is similar to the works of [3], [4], where the time-varying reference trajectory is sent to only a subset of the agents and the network graph is assumed to be undirectionally connected. The proposed distributed controllers while delivering a very excellent control performance, are model-free and require only the neighbors' information. Therefore the proposed method is considered to be simple and requires moderate computational burden.

## REFERENCES

[1] W. Ren and Y. Cao, Distributed Coordination of Multi-agent Networks, Emergent Problems, Models, and Issues, Springer, 2011

[2] S. Khoo, L. Xie and Z. Man, Robust finite-time consensus tracking algorithm for multirobot systems, IEEE Transactions on Mechantronics, Vol. 14, No. 2, 219/228, 2009.

[3] J. Mei, W. Ren and G. Ma, Distributed coordinated tracking with a dynamic leader for multiple Euler-Lagrange Systems, IEEE Transactions on Automatic Control, Vol. 56, No. 6, 1415/1421, 2011.

[4] W. Dong, On consensus algorithms of multiple uncertain mechanical systems with a reference trajectory, Automatica, Vol. 47, No. 9, 2023/2028, 2011.

[5] P. F. Hokayem, D. M. Stipanovic and M. W. Spong, Semiautonomous control of multiple networked Lagrangian systems, International Journal of Robust and Nonlinear Control, Vol. 19, No. 18, 2040/2055, 2009.

[6] Z. J. Yang, Y. Fukushima and P. Qin, Decentralized adaptive robust control of robot manipulators using disturbance observers, IEEE Transactions on Control Systems Technology, Vol. 20 (DOI: 10.1109/TCST.2011.2164076).

[7] M. W. Spong and M. Vidysagar, Robot Dynamics and Control, Wiley, 1989.

[8] S. H. Hsu and L. C. Fu, A fully adaptive decentralized control of robot manipulators, Automatica, Vol. 42, No. 10, 1761/1767, 2006.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Robust hybrid fuzzy logic control of a novel two-wheeled robotic vehicle with a movable payload under various operating conditions

A M Almeshal and M O Tokhi
Department of Automatic Control and Systems Engineering
University of Sheffield
Sheffield, United Kingdom
a.m.almeshal@sheffield.ac.uk

K M Goher
Department of Mechanical and Industrial Engineering
Sultan Qaboos University
Muscat, Oman
kgoher@squ.edu.om

***Abstract*—A novel design of two-wheeled double inverted pendulum-like vehicle with a movable payload is presented in this paper. The developed design extends the abilities of the vehicle with five degrees of freedom. The increase of degrees of freedom provides the vehicle with more flexibility in maneuvering in narrow spaces and limited travel distances. The dynamic model of the system is derived using Euler-Lagrange approach and simulated in Matlab Simulink environment. A hybrid fuzzy logic control approach is adopted to control and stabilise the vehicle. Various external disturbances are applied to the system to test the robustness of the control approach. It is demonstrated that the proposed controller successfully stabilises the vehicle in the upright position.**

***Keywords; modelling ; Lagrangian dynamics ; double inverted pendulum ; hybrid control; fuzzy logic control;***

## I. Introduction

An extensive amount of research on inverted pendulum (IP) system is found in the literature due its high nonlinearity and under-actuated nature [1]. It has been used as a platform to test various control algorithms as well as a basis to develop various applications. The applications include, but not limited to, gait of humanoid robots, personal transporters, and self-balancing wheelchairs [2]. The IP system exists in many variations such as mobile inverted pendulum on cart and rotational IP systems. In addition, IP systems have been extended to have multiple links to increase the degrees of freedom (DOF); such as double inverted and triple inverted pendulums [3][4][5]. The increased DOF increase the complexity of the system adding more space for new applications to be developed.

A novel configuration of balancing two-wheeled double inverted pendulum-like vehicle with a movable payload has been presented in [1]. The novelty in the configuration relies in the increased degrees of freedom with the ability to lift a payload to a higher extent. Mathematical model of system dynamics has been derived and presented in [2]. A hybrid fuzzy logic control (FLC) approach is developed to control the vehicle. The vehicle model is used as a basis for an ongoing research of new type of self-balancing wheelchairs with an extended height and ability to maneuver on irregular and uneven surfaces.

## II. System Description and Mathematical Model

A schematic diagram of the designed vehicle is shown in Figure 1. The vehicle is designed based on the double inverted pendulum system model with novel modifications [1]. The vehicle consists of two links and a cart driven by two DC motors that in turn drive the entire system. A third DC motor is used to drive the second link. These motors will help stabilizing the system in the upright position by applying an appropriate control signal. The second link consists of two co-axial rods connected by a linear actuator to enable lifting up the payload to a demanded height. Therefore, the system has five degrees of freedom; translational motion with the right and left wheels, first and second links and linear actuator on the second link. The tilt angles of the first and second links are $\theta_1$ and $\theta_2$ respectively. The linear displacement of the payload designated as Q, while the angular displacements of the left and right wheels designated as $\delta_L$ and $\delta_R$ respectively.

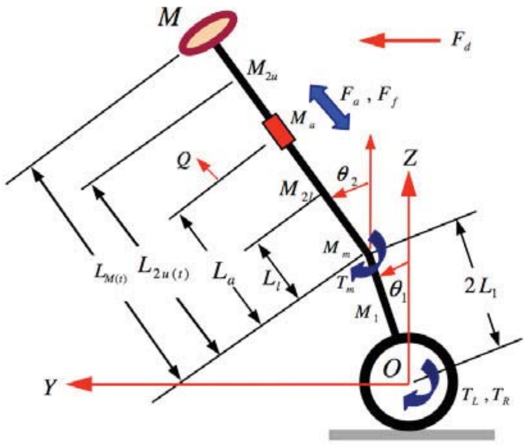


Figure 1. Schematic description of the vehicle

Table 1 describes system model terms and parameters. The system mathematical model consists of five differential equations. The system was modeled using Euler-Lagrange approach because of the system complexity and high coupling nature. System equations of motion were derived in detail in [2] and are presented as in equations (1-5).

Table 1. Nomenclature

| *Variable* | *Description* | *Unit* |
|---|---|---|
| $L_{M(t)}$ | Distance to the COM of the payload | m |
| $L_{2u(t)}$ | Distance to the COM of the upper part link 2 | m |
| $L_a$ | Position of the linear actuator | m |
| $L_l$ | Half length of link 1 | m |
| Q | Displacement of the linear actuator | m |
| $M_1$ | Mass of link 1 | kg |
| $M_m$ | Mass of motor driving link 2 | kg |
| $M_{2l}$ | Mass of the lower part of link 2 | kg |
| $M_a$ | Mass of the linear actuator | kg |
| $M_{2u}$ | Mass of the upper part of link 2 | kg |
| M | Payload mass | kg |
| $T_R$,$T_L$ | Right and left wheels driving torques | N.m |
| $T_m$ | Motor torque | N.m |
| $F_f$ | Frictional force in the linear actuator | N |
| $F_d$ | External disturbance force | N |
| $\theta_1$ | Angular position of link 1 to the positive Z axis | rad |
| $\theta_2$ | Angular position of link 2 to the positive Z axis | rad |
| $\phi$ | Yaw angle of the vehicle around the Z axis | rad |
| $\delta_R$ , $\delta_L$ | Angular position of the right and left wheels | rad |
| $J_l$ | Mass moment of inertia of link 1 | kg.m$^2$ |
| $J_{2u}$ | Mass moment of inertia of the upper rod of link 2 | kg.m$^2$ |
| $J_a$ | Mass moment of inertia of the actuator | kg.m$^2$ |
| $J_M$ | Mass moment of inertia of the payload | kg.m$^2$ |
| $J_w$ | Mass moment of inertia of the wheels | kg.m$^2$ |
| $J_{IB}$ | Mass moment of inertia of the intermediate body | kg.m$^2$ |
| $J_{2L}$ | Mass moment of inertia of the lower rod of link 2 | kg.m$^2$ |

$$\begin{aligned}
&2C_{21}\ddot{\delta}_L + C_{22}\ddot{\delta}_R + C_9\frac{R_w}{2}L_1\ddot{\theta}_1\cos\theta_1 - C_9\frac{R_w}{2}L_1\dot{\theta}_1^2\sin\theta_1\\
&+\frac{R_w}{2}(C_{10}+C_8Q)\ddot{\theta}_2\cos\theta_2 - \frac{R_w}{2}(C_{10}+C_8Q)\dot{\theta}_2^2\sin\theta_2\\
&+\frac{R_w}{2}C_8\dot{Q}\dot{\theta}_2\cos\theta_2 = T_L - T_{fL}
\end{aligned}\tag{1}$$

$$\begin{aligned}
&2C_{21}\ddot{\delta}_R + C_{22}\ddot{\delta}_L + C_9\frac{R_w}{2}L_1\ddot{\theta}_1\cos\theta_1 - C_9\frac{R_w}{2}L_1\dot{\theta}_1^2\sin\theta_1\\
&+\frac{R_w}{2}(C_{10}+C_8Q)\ddot{\theta}_2\cos\theta_2 - \frac{R_w}{2}(C_{10}+C_8Q)\dot{\theta}_2^2\sin\theta_2\\
&+\frac{R_w}{2}C_8\dot{Q}\dot{\theta}_2\cos\theta_2 = T_R - T_{fR}
\end{aligned}\tag{2}$$

$$\begin{aligned}
&2C_{18}\ddot{\theta}_1 + C_9\frac{R_w}{2}L_1(\ddot{\delta}_L+\ddot{\delta}_R)\cos\theta_1 - C_9\frac{R_w}{2}L_1(\dot{\delta}_L+\dot{\delta}_R)\dot{\theta}_1\sin\theta_1\\
&+2L_1(C_{10}+C_8Q)\ddot{\theta}_2\cos(\theta_1-\theta_2) - 2L_1(C_{10}+C_8Q)\dot{\theta}_1\dot{\theta}_2\sin(\theta_1-\theta_2)\\
&+2L_1(C_{10}+C_8Q)\dot{\theta}_2^2\sin(\theta_1-\theta_2) + 2L_1C_8\dot{Q}\dot{\theta}_2\cos(\theta_1-\theta_2)\\
&+C_9\frac{R_w}{2}L_1\dot{\theta}_1^2(\dot{\delta}_L+\dot{\delta}_R)\sin\theta_1 + 2L_1(C_{10}+C_8Q)\dot{\theta}_1^2\dot{\theta}_2\sin(\theta_1-\theta_2)\\
&-C_3g\dot{\theta}_1\sin\theta_1 = \frac{1}{2}(T_{LT}+T_{RT})
\end{aligned}\tag{3}$$

$$\begin{aligned}
&C_{20}\ddot{\theta}_2 + (C_{12}\dot{Q} + 2C_8Q)\dot{\theta}_2 + (C_{12}Q + C_8Q^2)\ddot{\theta}_2\\
&+\frac{R_w}{2}(C_{10}+C_8Q)(\ddot{\delta}_L+\ddot{\delta}_R)\cos\theta_2\\
&-\frac{R_w}{2}(C_{10}+C_8Q)(\dot{\delta}_L+\dot{\delta}_R)\dot{\theta}_2\sin(\theta_2)\\
&+C_8\frac{R_w}{2}\dot{Q}(\dot{\delta}_L+\dot{\delta}_R)\cos\theta_2 + 2L_1(C_{10}+C_8Q)\ddot{\theta}_1\cos(\theta_1-\theta_2)\\
&-2L_1(C_{10}+C_8Q)\dot{\theta}_1^2\sin(\theta_1-\theta_2) + 2L_1(C_{10}+C_8Q)\dot{\theta}_1\dot{\theta}_2\sin(\theta_1-\theta_2)\\
&+2C_8L_1\dot{\theta}_1\dot{\theta}_2\cos(\theta_1-\theta_2) + \frac{R_w}{2}(C_{10}+C_8Q)(\dot{\delta}_L+\dot{\delta}_R)\dot{\theta}_2^2\sin\theta_2\\
&-(C_{15}+C_8Q)g\dot{\theta}_2\sin\theta_2\\
&-2L_1(C_{10}+C_8Q)\dot{\theta}_1\dot{\theta}_2^2\sin(\theta_1-\theta_2) = T_M - T_{FM} - L_dF_d
\end{aligned}\tag{4}$$

$$\begin{aligned}
&C_8\ddot{Q} - \frac{1}{2}(C_{12}+2C_8Q)\dot{\theta}_2^2 - C_8\frac{R_w}{2}\dot{\theta}_2(\dot{\delta}_L+\dot{\delta}_R)\cos\theta_2\\
&-2L_1C_8\dot{\theta}_1\dot{\theta}_2\cos(\theta_1-\theta_2) + C_8g\cos\theta_2 = F_a - F_{fa}
\end{aligned}\tag{5}$$

## III. Hybrid Fuzzy Logic Control Strategy

Two types of hybrid FLC are designed and implemented to control the novel non-linear two-wheeled vehicle with five degrees of freedom and movable payload. Referring to Figure 2, the block diagram of the system is presented. Proportional-Derivative-like fuzzy logic controller (PD-like FLC) is used to control the angular displacement of the wheels, the first link tilt angle and the payload linear actuator displacement. A PD plus integral fuzzy logic controllers (PD+I FLC) are used to control the links tilt angles to overcome the steady-state error.

The inputs to the PD-like FLC are the error signal and the change of error. While the inputs for the PD+I FLC are the error signal, change of error and the sum of previous errors. Figure 2 presents the Simulink block diagram of the controlled system.

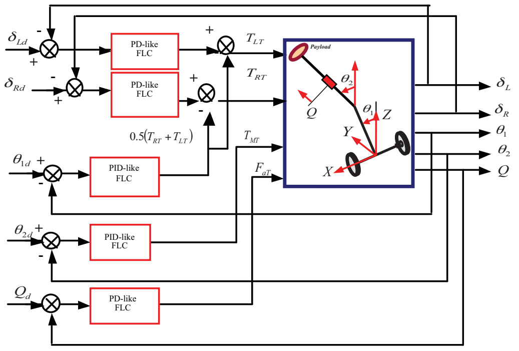


Figure 2. System simulink block diagram

The design of the FLC involves choosing a suitable fuzzy inference engine, defining the fuzzy rules and choosing the membership function type. The FLC developed here is based on Mamdani-type fuzzy inference engine with 25 fuzzy rules presented in Table 2 and Figure 3. The generation of the fuzzy rules-base is based on the required system performance to minimize the system error between the output signal and the desired signal.

Table 2. Fuzzy rule base

| **e\|ê** | **NB** | **NS** | **Z** | **PS** | **PB** |
|---|---|---|---|---|---|
| **NB** | NB | NB | NB | NS | Z |
| **NS** | NB | NB | NS | Z | PS |
| **Z** | NB | NS | Z | PS | PB |
| **PS** | NS | Z | PS | PB | PB |
| **PB** | Z | PS | PB | PB | PB |

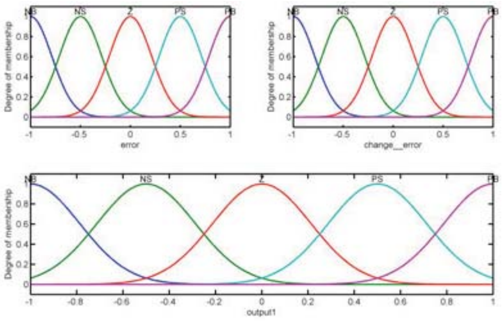


Figure 3. Gaussian fuzzy membership functions

## IV. Simulation And Results

In order to test the robustness of the developed control approach, two types of external disturbances were applied to the system. The first simulation scenario incorporates applying disturbances with different amplitudes to the system. The second scenario is to apply disturbances with different durations to the system.

### *A. Variation of disturbance force amplitudes*

Different levels of the applied disturbance force; 0 N, 40 N, 80 N, 160 N, 300 N, were used to examine the robustness of the developed hybrid FLC. The response of the system to the disturbances applied to the centre of the first link and second link are presented in Figures 4 and 5 respectively.

As noted in Figure 4, overshoots resulted in the tilt angle of the first link with a maximum peak amplitude increase of 30% at the maximum disturbance. The average settling time is approximately 5 seconds. The settling time increased by 3% of the initial value with each increase in the disturbance amplitude. Negative peaks at displacements of both wheel motors. Thus explains the opposing movement of the cart to overcome the disturbances and to stabilise the first link at the upright position. The tilt angle of the second link and the displacement of the payload actuator were not affected by this disturbance.

Referring to Figure 5, the effect of applying disturbances to the centre of the second link can be noted. Disturbances resulted in oscillations in the tilt angle of the second link. The tilt angle converges back to the set point within 5 seconds at the maximum applied disturbance force while the rise-time remained unchanged. With every increment of disturbance force amplitude, the peak value increased by an average of 50%. The angular displacements of both wheels, the tilt angle of the first link and the displacement of the payload actuator remained almost unaffected by this disturbance. This is due to the high damping effects caused by the joints of the vehicle and the motor linking the pendulum links.

### *B. Variation of disturbance force duration*

Different durations of the applied disturbance force; 0 sec, 1.25 sec, 2.5 sec, 7.5 sec, 12.5 sec, with a fixed force amplitude of 40 N were used. The response of the system to the disturbances applied to the centre of the first and second links are presented in Figures 6 and 7 respectively.

As noted in Figure 6, oscillations were observed at the tilt angle of the first link with varied amplitudes. The longer the applied disturbance duration the larger the amplitudes of oscillations were noted. The settling-time increased by an average of 6% while the rise-time varied by increasing and decreasing depending on the duration of the applied disturbance. The right and left wheel angular displacements were affected by this disturbance and fluctuations can be noted. The tilt angle of the second link and the payload actuator displacement were not affected by this disturbance.

Figure 7 shows the response of the system for disturbances applied at the centre of the second link. The disturbances resulted in an oscillatory response at the tilt angle of the second link; at disturbance durations up to 7.5 seconds, the controller was able to control the tilt angle of the second pendulum and stabilise the system within an approximate average times of 7 seconds. Settling-time increased by an average of 7% at a time. Finally, The right and left wheel angular displacements, tilt angle of the first link and the payload actuator displacement were not affected by this disturbance.

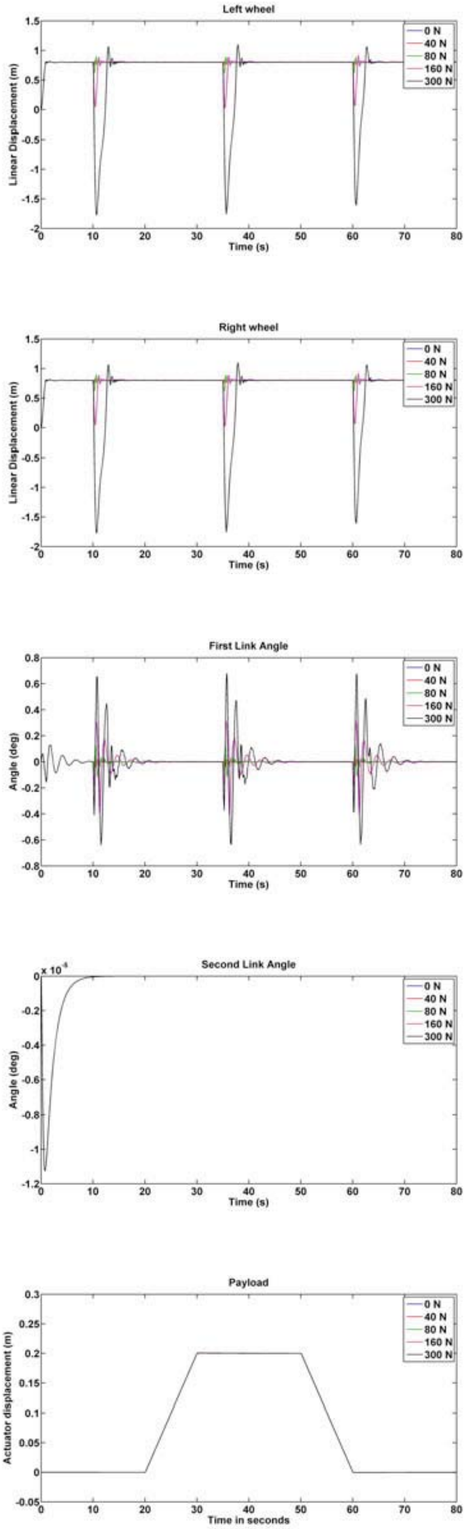


Figure 4. System performance with disturbances
Of different amplitudes applied at the centre of 1st link

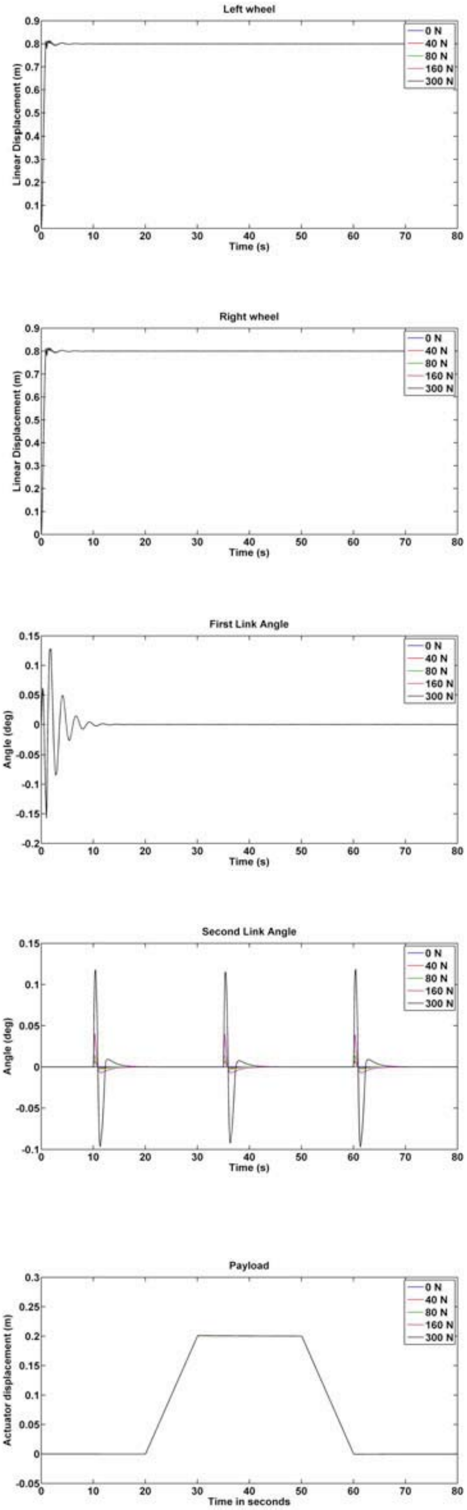


Figure 5. System performance with disturbances
Of different amplitudes applied at the centre of 2nd link

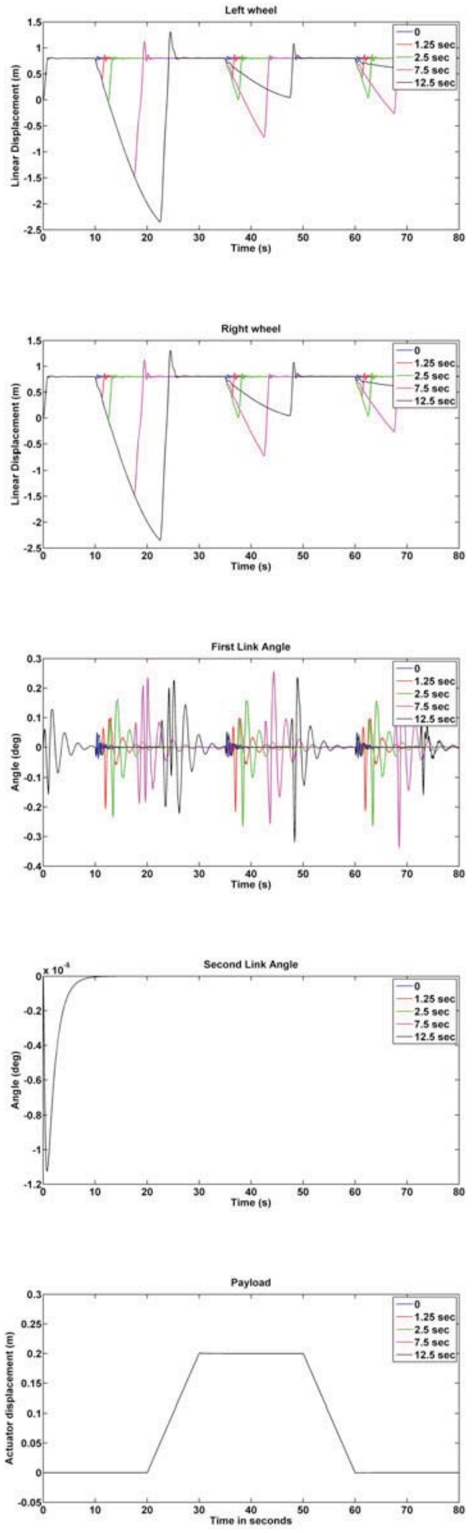


Figure 6. System performance with disturbances Of different durations applied at the centre of 1st link

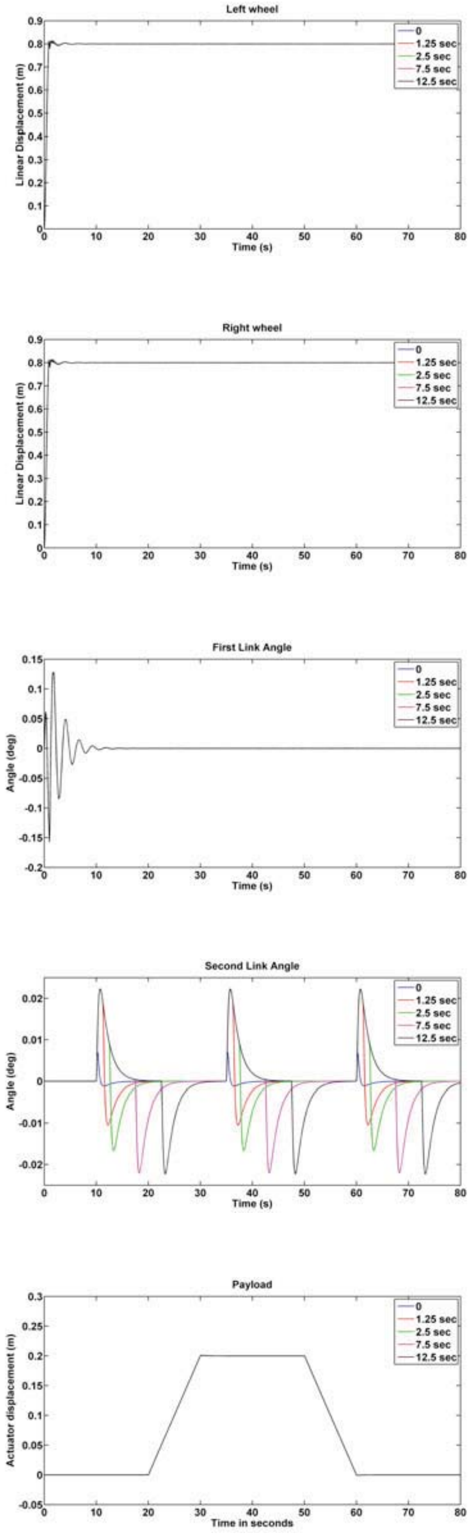


Figure 7. System performance with disturbances Of different durations applied at the centre of 2nd link

## V. CONCLUSION

A hybrid fuzzy logic control approach has been designed and applied on a novel configuration of two-wheeled robotic vehicle. The mathematical model of the vehicle with five degrees of freedom has been derived and equations of motion presented. The model was simulated in Matlab Simulink environment and successfully stabilised. Simulations of the system undergoing different disturbance scenarios have been demonstrated and it has been proved that the controller is able to balance the vehicle in the upright position with various disturbances. The developed hybrid FLC mechanism has been shown to be robust and overcoming external impacts of disturbances.

## REFERENCES

[1] K. Goher, S. Ahmad, M. O. Tokhi, A new configuration of two wheeled vehicles: Towards a more workspace and motion flexibility, Proceedings of 4th Annual IEEE conference on Intelligent Systems, San Diego, CA, USA, 5-8 April 2010.

[2] A. M. Al-Meshal, K. M. Goher, M.O. Tokhi, (2011, 8-10 September), Modelling of two-wheeled robotic wheelchair with moving payload. Proceedings of the 14th International Conference on Climbing and Walking Robots and the Support Technologies for Mobile Machines (CLAWAR 2011), Paris, France.

[3] K.G. Eltohamy, and C.Y. Kuo, Real time stabilisation of a triple link inverted pendulum using single control input , IEE Proceedings on Control Theory and Applications , vol.144, no.5, pp.498-504, Sep 1997.

[4] S. Ahmad S., M.O. Tokhi, Modeling and Control of a Wheelchair on Two Wheels, Second Asia International Conference on Modeling & Simulation, 2008. AICMS 08. ,pp.579-584, 13-15 May 2008.

[5] L. Yang, Y. Zhang, Bilinear control for the triple inverted pendulum based on model bias separation, Computer and Automation. 5 (2010) 436-440.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# A Novel Hybrid Spiral Dynamics Bacterial Chemotaxis Algorithm for Global Optimization with Application to Controller Design

A. N. K. Nasir, M. O. Tokhi and N. M. Abd Ghani
Department of Automatic Control & Systems Engineering, University of Sheffield, Sheffield, UK.
cop11ann@sheffield.ac.uk

M. A. Ahmad
Faculty of Electrical & Electronics Engineering, Universiti Malaysia Pahang (UMP), 26600 Pekan, Pahang, Malaysia.

*Abstract*—**This paper presents a hybrid optimization algorithm, referred to as hybrid spiral dynamics bacterial chemotaxis (HSDBC) algorithm. HSDBC synergizes bacterial foraging algorithm (BFA) chemotaxis strategy and spiral dynamics algorithm (SDA). The original BFA has higher convergence speed while SDA has better accuracy and stable convergence when approaching the optimum value. This hybrid approach preserves the strengths of BFA and SDA and thus has the capability of producing better results. Moreover, it has simple structure, hence reduced computational cost. Several unimodal and multimodal benchmark functions are employed to test the algorithm in finding the global optimum point. Furthermore, the proposed algorithm is tested in the design of PD controller for a flexible manipulator system. The results show that the HSDBC outperforms SDA and BFA in all test functions and successfully optimizes the PD controller.**

***Keywords-Spiral dynamics; bacterial chemotaxis; optimization algorithm; PD control; flexible manipulator.***

## I. Introduction

Metaheuristic optimization algorithms have gained a lot of interest by many researchers worldwide. These algorithms are inspired by biological phenomena or natural phenomena. Some of the newly introduced algorithms include biogeography-based optimization (BBO) [1], firefly optimization algorithm [2], cuckoo search optimization [3], galaxy-based search algorithm [4], and spiral dynamics inspired optimization (SDA) [5]. All these algorithms have gained attention due to their simplicity to program, fast computing time, easy to implement, and possibility to apply to various applications. Each of these algorithms has its own unique features, advantages and also disadvantages. Therefore, there are a lot of possibilities to improve the algorithms from various aspects. Many attempts have been made to improve performances of the algorithms such as developing adaptive approaches or incorporating powerful mathematical functions into the algorithms and mostly hybridizing two or more algorithms.

Hybridisation is a common approach used in metaheuristic to enhance capability of optimization algorithms. It may reduce computational cost by making a simple and better structure to lead to higher performance. Moreover, with the rapidly emerging computing tools and efficiency in current technology, hybrid approaches have become increasingly popular to explore. Various combinations of optimization algorithms have been considered by researchers with the aim to increase system performance. [6] developed a hybrid optimization algorithm combining bacterial foraging optimisation algorithm (BFA) with BBO, and referred to it as intelligent biogeography-based optimization. In the algorithm, chemotaxis behaviour of bacteria is adopted into BBO migration process to determine a valid emigration of an individual from one place to another. This ensures the island that receives the emigrated solution preserves its fitness level by only accepting individuals that contribute to a better fitness value. [7] introduced hybrid version of BFA with differential evolution (DE) algorithm called chemotaxis differential evolution. In the algorithm, chemotaxis strategy of bacteria is combined with the mutation process in DE. [8], [9] and [10] introduced hybrid GA-BF algorithm employing modified mutation and crossover operation in GA while applying variation bacterial chemotaxis step size in BFA. [11] developed cooperative (BF-TS) by combining adaptive bacterial foraging optimization algorithm (ABFA) and adaptive tabu search (ATS). With limited exploration capability of ATS in the search space and complexity of ABFA, the chemotaxis strategy of ABFA is incorporated into ATS to provide suitable exploration at the early stage. On the other hand, [12] used hybrid ABFA and ATS called BTSO, to analyze Lyapunov's stability of linear and nonlinear systems. [13] introduced a hybrid algorithm namely BPSO-DE synergizing BFA, particle swarm optimization (PSO), and DE to solve dynamic economic dispatch problem with valve-points effect. Bacterial chemotaxis strategy with adaptive step-size in BFA is used to perform local search to enhance exploitation while PSO-DE features containing evolutionary operators and velocity update equation are used to perform exploration search over the entire search space. Hybrid BFA and PSO on the other hand, has received the most attention. [14], [15], [16], [17] and [18] employed velocity and position update equation in PSO to act as global search method while utilizing chemotaxis strategy in BFA to serve as local search method. [19] introduced simplified version of BFA employing bacterial chemotaxis strategy and PSO velocity update equation to solve parameter identification problem of heavy oil thermal cracking model. Reproduction and elimination stages were omitted to reduce computational time.

This paper presents hybrid version of bacterial foraging algorithm (BFA) chemotaxis strategy and spiral dynamics algorithm (SDA). The rest of the paper is organized as follows. Section II provides a brief literature review of the original BFA and spiral dynamics inspired optimization. The proposed HSDBC is described in section III. Validation of the proposed HSDBC in comparison to SDA and BFA with uni-modal and multi-modal test functions as well as application of the algorithm in optimizing a PD controller is presented in section IV. Section V presents concluding remarks.

## II. BFA AND SDA

The original versions of BFA and SDA are briefly described in this section.

### A. *Bacterial foraging optimization algorithm*

The BFA is a biologically inspired algorithm introduced in [20]. It is based on adaptation technique of Escherichia Coli (E. Coli) bacteria to find nutrient or food source during their lifetime or alternatively the technique might be called bacterial foraging strategy. Furthermore, E. Coli bacteria use saltatory search technique, which is the combination of cruise and ambush movement. One of the exceptional features of E. Coli is that it has very high growth rate, which is normally exponential. This extraordinary capability of E. Coli has motivated researchers to adopt the strategy as optimization technique. Bacterial foraging strategy consists of three basic cycles namely chemotaxis, reproduction and elimination & dispersal. These cycles are continuing processes and very effective for optimization purposes [21]. Moreover, it offers flexibility for researchers to manipulate the strategy according to a specific application area. When searching for food or nutrient, tumbling and swimming will take place. Tumbling is similar to cruise and it happens when the E. Coli navigates in the search area and once the food source is found, it swims like ambushing a target area with great speed, up to 20μm/s or faster in a rich nutrient medium. This unique movement is called chemotaxis. Reproduction, elimination and dispersal events then happen to bacteria with high fitness or healthier that has capability to reach food source accurately and quickly. The strength of BFA lies in the bacterial chemotaxis strategy adopted by many researchers to improve the optimization algorithm. The details of the original algorithm and pseudocode of BFA can be found in [20]. In this paper, number of bacteria, number of chemotaxis, chemotactic step size, number of swims, number of reproduction, number of elimination & dispersal are represented as *S, Nc, C, Ns, Nre* and *Ned* respectively. The probability that each bacterium will be eliminated and dispersed is defined as *ped* = 0.25 for the problems considered.

### B. *Spiral dynamics inspired optimization algorithm*

The SDA is another metaheuristic algorithm adopted from spiral phenomena in nature [5]. 2-dimensional and N-dimensional [5] logarithmic spiral discrete models have been tested on several benchmark functions. Moreover, comparisons with other optimization algorithms such as PSO and DE have shown that SDA performance is either better or the same as those [5]. This simple and effective strategy retains the diversification and intensification at the early phase and later phase of the trajectory as diversification and intensification are important features of the optimization algorithm. At the early stage, the spiral trajectory explores a wider search space and it continuously converges with a smaller radius providing dynamics step size when approaching the final point, which is the best solution, located at the centre. The distance between a point in a path trajectory and the centre point is varied constantly if the radius of the trajectory is changing at constant rate thus making the radius an important converging parameter for the algorithm. The strength of SDA lies in its spiral dynamics model. An n-dimensional spiral mathematical model that is derived using composition of rotational matrix based on combination of all 2 axes is given as:

$$x(k+1) = S_n(r,\theta)x(k) - (S_n(r,\theta) - I_n)x^* \qquad (1)$$

where

$$S_n(r,\theta)x(k) = rR^n(\theta_{1,2},\theta_{1,3},\ldots,\theta_{n-1,n})x(k).$$

or

$$S_n(r,\theta)x(k) = \prod_{i-1}^{n-1}\left(\prod_{j=1}^{i}(R^n_{n-i,n+1-j}(\theta_{n-i,n+1-j}))\right)$$

and $R^n_{i,j}(\theta_{i,j}) :=$

$$\begin{array}{c|ccccccccc|}
 & & & & i & & j & & & \\
 & 1 & & & & & & & & \\
 & & \ddots & & & & & & & \\
 & & & 1 & & & & & & \\
i & & & & \cos\theta_{i,j} & \cdots & -\sin\theta_{i,j} & & & \\
 & & & & & 1 & & & & \\
 & & & & \vdots & \ddots & \vdots & & & \\
 & & & & & 1 & & & & \\
j & & & & \sin\theta_{i,j} & \cdots & \cos\theta_{i,j} & & & \\
 & & & & & & & 1 & & \\
 & & & & & & & & \ddots & \\
 & & & & & & & & & 1
\end{array}$$

Parameters and descriptions used in equation (1) are similar to those used in HSDBC optimization algorithm, which are shown in Table 1. Since SDA is relatively new, not much work in the literature involving the algorithm has been reported. The details of the original SDA algorithm for 2-dimension and n-dimension can be found in [5]. The hybrid approach of this algorithm and its details are provided in the next section.

## III. HYBRID SPIRAL DYNAMICS BACTERIAL CHEMOTAXIS ALGORITHM

The HSDBC is a combination of bacterial chemotaxis strategy used in BFA and SDA. BFA has faster convergence speed due to the chemotaxis approach but suffers from oscillation problem towards the end of its search process. On the other hand, SDA provides better stability when approaching optimum point due to dynamic spiral step in its trajectory motion but has slower convergence speed. HSDBC algorithm preserves the strengths possessed by BFA and SDA. Moreover, by incorporating only chemotaxis part of BFA simple structure of SDA can be retained, thus reducing computational time and enhancing performance of the algorithm. The parameters and description used in n-dimensional HSDBC optimization algorithm are presented in Table 1 and the algorithm is shown in Fig. 1.

TABLE I. PARAMETERS FOR HSDBC OPTIMIZATION ALGORITHM

| Symbols | Description |
|---|---|
| $\theta_{i,j}$ | Bacteria angular displacement on $x_i - x_j$ plane around the origin |
| $r$ | Spiral radius |
| $m$ | Number of search points |
| $k_{\max}$ | Maximum iteration number |
| $N_s$ | Maximum number of swim |
| $x_i(k)$ | Bacteria position |
| $R^n$ | n x n matrix |

An n-dimensional hybrid spiral dynamics bacteria chemotaxis optimization algorithm.

```
Step 0: Preparation
    Select the number of search points (bacteria) $m \geq 2$ , parameters
    $0 \leq \theta < 2\pi, 0 < r < 1$ of $S_n(r,\theta)$ , maximum iteration number,
    $k_{\max}$ and maximum number of swim, $N_s$ for bacteria chemotaxis.
    Set $k = 0, s = 0$ .
Step 1: Initialization
    Set initial points $x_i(0) \in R^n$, $i = 1,2,...m$ in the feasible region
    at random and center $x^*$ as $x^* = x_{i_g}(0)$ ,
    $i_g = \arg\min_i \ f(x_i(0)), i = 1,2,...,m$ .
Step 2: Applying bacteria chemotaxis
    (i) Update $x_i$
        $x_i(k+1) = S_n(r,\theta)x_i(k) - (S_n(r,\theta) - I_n)x^*$
        $i = 1,2,...,m$ .
    (ii) Bacteria swim
        (a) Check number swim for bacteria i.
            If $s < N_s$ , then check fitness,
            Otherwise set $i = i+1$ , and return to step (i).
        (b) Check fitness
            If $f(x_i(k+1)) < f(x_i(k)$ , then update $x_i$ ,
            Otherwise set $s = N_s$ , and return to step (i).
        (c) Update $x_i$
            $x_i(k+1) = S_n(r,\theta)x_i(k) - (S_n(r,\theta) - I_n)x^*$
            $i = 1,2,...,m$ .
Step 3: Updating $x^*$
    $x^* = x_{i_g}(k+1)$ ,
    $i_g = \arg\min_i \ f(x_i(k+1)), i = 1,2,...,m$ .
Step 4: Checking termination criterion
    If $k = k_{\max}$ then terminate. Otherwise set $k = k+1$ ,
    and return to step 2.
```

Figure 1. HSDBC optimization algorithm.

In the proposed hybrid approach, bacterial chemotaxis strategy is employed in step 2 to balance and enhance exploration and exploitation of the search space. The bacteria move from low nutrient location towards higher nutrient location, placed at the centre of a spiral. The most important factor of HSDBC algorithm is the respective diversification and intensification at the early phase and later phase of the spiral motion. In the diversification phase, bacteria are located at low nutrient location and move with larger step size thus producing faster convergence. On the other hand, in the intensification phase, bacteria are approaching rich nutrient location and move with smaller step size hence avoiding oscillation around the optimum point. Another factor contributing to better performance of the algorithm is the swimming action in bacterial chemotaxis. Bacteria continuously swim towards optimum point if the next location has higher nutrient value compared to previous location until the maximum number of swim is reached.

## IV. VALIDATION TEST AND RESULTS

In this section, the proposed algorithm is validated through simulation tests on two 3-dimensional uni-modal and two 2-dimensional multi-modal benchmark functions. Moreover, the HSDBC algorithm is tested in optimizing PD controller of a flexible manipulator system. Comparison with the original version of SDA and BFA tested on the four benchmark functions is also given to show the improved performance of HSDBC. The parameters used in the simulation are chosen heuristically for all test functions.

### A. *Uni-modal sphere function*

The sphere function is defined as:

$$f(x) = \sum_{i=1}^{n} x_i^2 \qquad (2)$$

The function has a global minimum at $x_i$ = [0, 0, 0] with fitness *f(x)* = 0. In this simulation, the sphere function is considered to have dimension n = 3 and variable $x_i$ is in the range [–5.12, 5.12]. Number of search points, m = 30, iteration number, 80, angular displacement, θ = π/4, and spiral radius, *r* = 0.96 were used for both algorithms. Number of swims with for HSDBC was defined as $N_s$ = 5. BFA parameters for this function were *S* = 30, *Nc* = 30, *C*=0.01, *Ns* = 4, *Nre* = 4 and *Ned* = 2. The convergence plot for 3 dimensional sphere function thus achieved is shown in Fig 2.

### B. *Uni-modal Ackley function*

The Ackley function is mathematically defined as:

$$\begin{aligned} f(x) = -20\exp(-0.2\sqrt{(\frac{1}{n}\sum_{i=1}^{n} x_i^2)} \\ -\exp(\frac{1}{n}\sum_{i=1}^{n}\cos(2\pi x_i) + 20 + e \end{aligned} \qquad (3)$$

The function has a global minimum at $x_i$ = [0, 0, 0] with fitness *f(x)* = 0. The Ackley function is considered with dimension n = 3 and variable $x_i$ in the range [–32.768, 32.768]. Number of search points, m = 30, iteration number

200, angular displacement, θ = π/4, and spiral radius, $r$ = 0.96 were used in both algorithms. Number of swims for HSDBC with swim radius, $r$ = 0.6 was defined as $N_s$ = 1. BFA parameters for this function were $S$ = 20, $Nc$ = 20, $C$ = 0.02, $Ns$ = 4, $Nre$ = 4 and $Ned$ = 2.The resulting convergence plot for 3-dimension Ackley function is shown in Fig 3.

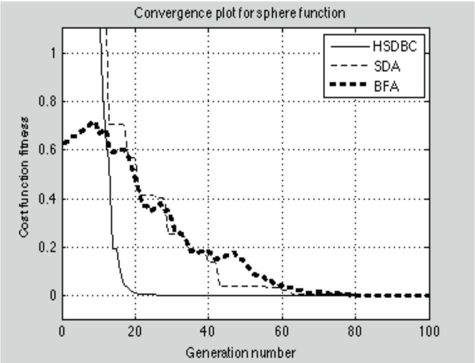


Figure 2. Convergence plot for 3D sphere function.

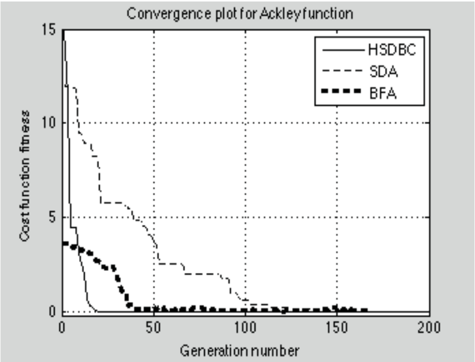


Figure 3. Convergence plot for 3D Ackley function.

### C. Multi-modal Rastrigin function

The Rastrigin function is defined as:

$$f(x) = \sum_{i=1}^{n} [x_i^2 - 10\cos(2\pi x_i) + 10] \qquad (4)$$

The function has a global minimum at $x_i$ = [0, 0] with fitness $f(x)$ = 0. The Rastrigin function is considered with dimension n = 2 and variable $x_i$ in the range [–5.12, 5.12]. The number of search points, m = 50, iteration number 120, angular displacement, θ = π/4, and spiral radius, $r$ = 0.96 were used in both algorithms. Number of swims for HSDBC with swim radius, $r$ = 0.65 was defined as $N_s$ = 2. BFA parameters for this function were $S$ = 30, $Nc$ = 20, $C$ = 0.01, $Ns$ = 4, $Nre$ = 4 and $Ned$ = 2. The resulting convergence plot for the 2-dimensional Rastrigin function is shown in Fig 4.

### D. Multi-modal Griewank function

The Griewank function is defined as:

$$f(x) = \frac{1}{4000}\sum_{i=1}^{n} x_i^2 - \prod_{i=1}^{n}\cos\left(\frac{x_i}{\sqrt{i}}\right) + 1 \qquad (5)$$

The function has a global minimum at $x_i$ = [0, 0] with fitness $f(x)$ = 0. The Griewank function was considered with dimension n = 2 and variable $x_i$ in the range [–600, 600]. The number of search points, m = 50, iteration number 200, angular displacement, θ = π/4, and spiral radius, $r$ = 0.96 were used for both algorithms. Number of swims for HSDBC with swim radius, $r$ = 0.55 was defined as $N_s$ = 1. BFA parameters for Griewank function were $S$ = 30, $Nc$ = 10, $C$ = 0.1, $Ns$ = 4, $Nre$ = 4 and $Ned$ = 2. The resulting convergence plot for the 2-dimensional Griewank function is shown in Fig 5.

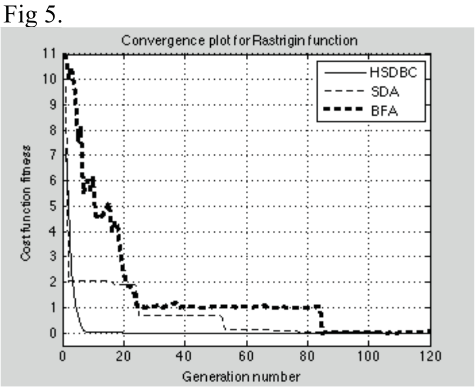


Figure 4. Convergence plot for 2D Rastrigin function.

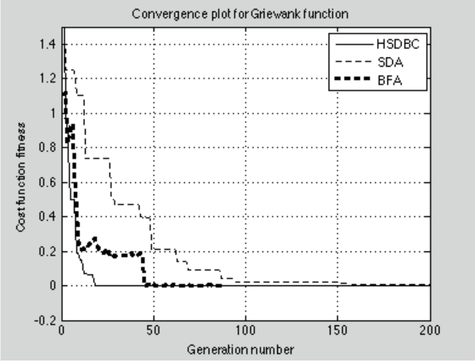


Figure 5. Convergence plot for 2D Griewank function.

It can be clearly seen in the plots, in Figures 2-5 that the HSDBC outperformed SDA and BFA in terms of convergence speed and improved accuracy. Numerical results of HSDBC, SDA and BFA performance tests with the benchmark functions are shown in Tables II, III and IV respectively. It is noted that HSDBC has achieved better performance than SDA and BFA with the test functions in terms of convergence speed and accuracy.

TABLE II. HSDBC PERFORMANCE ON BENCHMARK FUNCTIONS

| **Cost Function Name** | **Performance** | | | | |
|---|---|---|---|---|---|
| | ***Best fitness*** | ***Converge time (iter)*** | $X_1$ | $X_2$ | $X_3$ |
| Sphere | $6\text{x}10^{-7}$ | 26 | $2\text{x}10^{-4}$ | $6\text{x}10^{-7}$ | $-7\text{x}10^{-4}$ |
| Ackley | $3\text{x}10^{-7}$ | 20 | $1\text{x}10^{-7}$ | $2\text{x}10^{-8}$ | $-3\text{x}10^{-8}$ |
| Rastrigin | 0 | 15 | $-2\text{x}10^{-9}$ | $4\text{x}10^{-10}$ | - |
| Griewank | $2\text{x}10^{-11}$ | 18 | $-3\text{x}10^{-6}$ | $6\text{x}10^{-6}$ | - |

TABLE III. SDA PERFORMANCE ON BENCHMARK FUNCTIONS

| Cost Function Name | Performance | | | | |
|---|---|---|---|---|---|
| | *Best fitness* | *Converge time (iter)* | $X_1$ | $X_2$ | $X_3$ |
| Sphere | 5x10$^{-3}$ | 63 | -4x10$^{-2}$ | -5x10$^{-2}$ | -7x10$^{-3}$ |
| Ackley | 6x10$^{-3}$ | 159 | 9x10$^{-4}$ | 2x10$^{-5}$ | -2x10$^{-3}$ |
| Rastrigin | 1x10$^{-6}$ | 84 | -8x10$^{-5}$ | -2x10$^{-5}$ | - |
| Griewank | 7x10$^{-5}$ | 91 | -6x10$^{-3}$ | -1x10$^{-2}$ | - |

TABLE IV. BFA PERFORMANCE ON BENCHMARK FUNCTIONS

| Cost Function Name | Performance | | | | |
|---|---|---|---|---|---|
| | *Best fitness* | *Converge time (iter)* | $X_1$ | $X_2$ | $X_3$ |
| Sphere | 5x10$^{-5}$ | 84 | 4x10$^{-3}$ | -4x10$^{-3}$ | 3x10$^{-3}$ |
| Ackley | 2x10$^{-2}$ | 40 | -5x10$^{-3}$ | -9x10$^{-3}$ | -1x10$^{-3}$ |
| Rastrigin | 5x10$^{-4}$ | 85 | -8x10$^{-4}$ | -1x10$^{-3}$ | - |
| Griewank | 7x10$^{-4}$ | 45 | -1x10$^{-2}$ | 4x10$^{-2}$ | - |

### *E. Controller design optimization*

The HSDBC algorithm is employed here to optimize PD controller of a flexible manipulator system (FMS). Schematic diagram of the flexible manipulator system is shown in Fig. 6. $X_oOY_o$ and XOY represent the stationary and moving coordinate frames respectively. $\tau$ represents the applied torque at the hub. Young modulus, area moment of inertia, mass density per unit volume, cross-sectional area, hub inertia, displacement and hub angle of the manipulator are represented by $E, I, \rho, A, I_h$, $v(x,t)$ and $\theta(t)$ respectively [23].

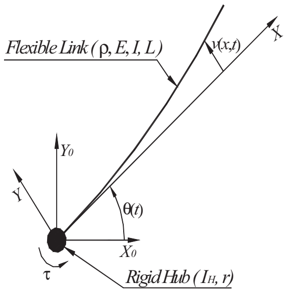


Figure 6. Schematic diagram of flexible manipulator system.

Mathematical model of FMS adopted here is that derived using Lagrange method in [22]. The FMS model has been used by many researchers in testing various types of controller for flexible systems [23], [24]. The dynamic equation of motion of FMS can be represented as:

$$M\ddot{Q}(t)+D\dot{Q}(t)+KQ(t)=F(t) \qquad (6)$$

where $M$, $D$ and $K$ are mass, damping and stiffness matrices respectively. $F(t)$ and $Q(t)$ are vectors of external forces and modal displacement respectively;

$$F(t)=[\ \tau \quad 0 \quad 0 \quad \cdots \quad 0\ ]^T \qquad (7)$$

$$Q(t)=[\ \theta \quad q_1 \quad q_2 \quad \cdots \quad q_n\ ]^T=[\ \theta \quad q^T\ ]^T \qquad (8)$$

More details of the derivation and parameters of FMS can be found in [22], [23]. A state-space model of FMS is obtained by linearizing (6) and it is used to design PD controller through HSDBC. The control strategy of FMS is adopted from [23] and [24] where PD feedback of collocated sensor signals is employed. A block diagram of the control structure is shown in Fig. 7, where $K_p$, $K_v$ and $A_c$ are the proportional, derivative and motor amplifier gains respectively. The input of the system is reference hub angle, $R_f$ and the outputs of the system are hub angle, θ and hub angle velocity, $\dot{\theta}$. In this simulation, number of search points, m = 30, iteration number 100, angular displacement, θ = π/4, and spiral radius, $r$ = 0.96, and number of swim $N_s$ = 3 were used to optimize the PD controller.

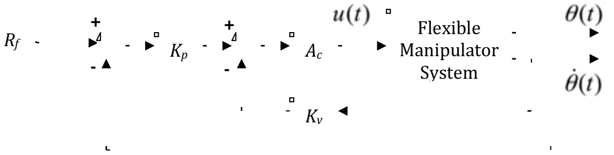


Figure 7. Collocated PD control structure of FMS.

Integral square error (ISE) of hub angle was chosen as cost function for the optimization algorithm. As a means of examining the proposed algorithm, this paper is only dealing with step input tracking capability of FMS. Step input was defined to have final value at 0.8 radians, which is the final location of hub angle. Graphical plot of the hub angle achieved is shown in Fig. 8.

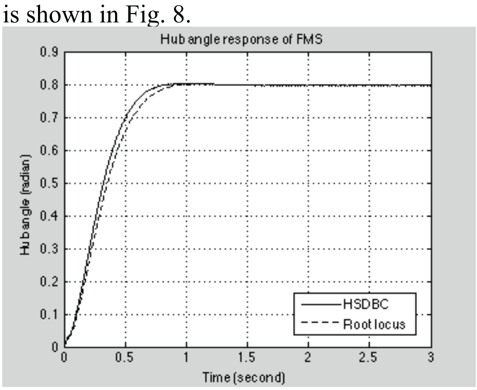


Figure 8. Hub angle response of FMS.

In this plot, for the purpose of comparison, the hub angle response with PD controller designed using root locus approach from [24] is also shown. Simulation with HSDBC optimization algorithm on the FMS gave $K_p$ =72.3459 and $K_v$ =20.6227 while $K_p$ =60 and $K_v$ =19 using root locus technique [24]. It is clear from Fig. 8, that hub angle response of FMS using HSDBC was better than hub angle response using root locus technique in terms of speed of response. Numerical results of the hub angle response are shown in Table V. It is noted that the HSDBC approach resulted slightly larger overshoot within acceptable range. However, the response rise time with HSDBC was better, which indicates that the algorithm can perform faster with satisfactory response overshoot and no error at steady state.

TABLE V. PERFORMANCE SPECIFICATION OF HUB ANGLE RESPONSE

| Tuning Method | Performance Specification | | | |
|---|---|---|---|---|
| | *Overshoot, %os (%)* | *Settling time, ts (s)* | *Rise time, tr (s)* | *Steady state error, ess* |
| HSDBC | 0.84 | 1.47 | 0.44 | 0 |
| Root locus | 0.53 | 1.47 | 0.50 | 0 |

## V. CONCLUSION

A novel hybrid spiral dynamics bacterial chemotaxis optimization algorithm has been proposed. Validation with uni-modal and multi-modal benchmark functions and comparison with standard SDA and BFA have been carried out. Moreover, the HSDBC has been used in controller design of a flexible manipulator in comparison with root locus design approach. Simulation results have shown that the proposed algorithm outperformed its counterpart in all test functions and it successfully optimized PD controller of flexible manipulator system in terms of convergence speed and accuracy.

## ACKNOWLEDGMENT

Ahmad Nor Kasruddin Nasir is on study leave from Faculty of Electrical & Electronics Engineering, Universiti Malaysia Pahang (UMP) and is sponsored by Ministry of Higher Education Malaysia.

## REFERENCES

[1] D. Simon, Biogeography-based optimization, IEEE Transaction on Evolutionary Computation, 12(16), December 2008, pp. 702-713.

[2] X. S. Yang, "Firefly algorithm, levy flight and global optimization," Research and Development in Intelligent systems XXVI, Springer London, 2010, pp. 209-218.

[3] X. S. Yang, and S. Deb, “Cuckoo search via levy flights,” Proceeding of World Congress on Nature & Biologically Inspired Computing (Nabic 2009), December 2009, India. IEEE Publication, USA, pp. 210 - 214.

[4] H. S. Hosseini, “Principal components analysis by the galaxy-based search algorithm, a novel metaheuristic for continuous optimization,” International Journal of Computational Science and Engineering, **6**(1/2), 2011, pp. 132 - 140.

[5] K. Tamura, and K. Yasuda, “Primary study of spiral dynamics inspired optimization,” IEEJ Transactions on Electrical and Electronic Engineering, **6**(S1), 2011, pp. 98 – 100.

[6] M. R. Lohokare, S. S. Pattnaik, S. Devi, B. K. Panigrahi, S. Das, K. M. Bakwad, “Intelligent biogeography-based optimization for discrete variables,” Proceeding of World Congress on Nature & Biologically Inspired Computing (NaBIC 2009), Coimbatore, India, 9 – 11 December, 2009, pp. 1088 – 1093.

[7] A. Biswas, S. Dasgupta, S. Das, A. Abraham, “A synergy of differential evolution and bacterial foraging optimization for global optimization,” Journal of Neural Network World, **17**(6), 2007, pp. 607 – 626.

[8] D. H. Kim, A. Abraham, J. H. Cho, “A hybrid genetic algorithm and bacterial foraging approach for global optimization,” Journal of Information Sciences, **177**, 2007, pp. 3918 – 3937.

[9] T. C. Chen, P. W. Tsai, S. C. Chu, J. S. Pan, “A novel optimization approach: bacterial-ga foraging,” Proceedings of the Second International Conference on Innovative Computing, Information and Control ICICIC '07, Kumamoto, Japan, September 5 - 7, 2007, pp. 391 – 394.

[10] D. G. Jadhav, S. S. Pattnaik, S. Devi, M. R. Lohokare, K. M. Bakwad, “Approximate memetic algorithm for consistent convergence,” Proceeding of National Conference on Computational Instrumentation (NCCI 2010), Chandigarh, India, 19-20 March 2010, pp. 118 – 122.

[11] N. Sarasiri, S. Sujitjorn, “Bacterial foraging optimization and tabu search: performance issues and cooperative algorithms,” Proceedings of the 10th WSEAS International Conference on Systems Theory and Scientific Computation, Taipei Taiwan, August 20-22, 2010, pp. 186 – 191.

[12] S. Panikhom, N. Sarasiri, S. Sujitjorn, “Hybrid bacterial foraging and tabu search optimization (BTSO) algorithms for Lyapunov’s stability analysis of nonlinear systems,” International Journal of Mathematics and Computers in Simulation, **3**(4), 2010, pp. 81 – 89.

[13] P. Praveena, K. Vaisakh, S.R.M. Rao, “A bacterial foraging and pso-de algorithm for solving dynamic economic dispatch problem with valve-point effects,” First International Conference on Integrated Intelligent Computing (ICIIC), Bangalore, India, 5-7 Aug. 2010, pp. 227 – 232.

[14] A. Biswas, S. Dasgupta, S. Das, A. Abraham, “Synergy of pso and bacterial foraging optimization – A comparative study on numerical benchmarks. Springer-Verlag Berlin Heidelberg, 2007, pp. 255 – 263.

[15] Y. Chu, H. Mi, H. Liao, Z. Ji, Q. H. Wu, “A fast bacterial swarming algorithm for high-dimensional function optimization,” Proceeding of 2008 IEEE Congress on Evolutionary Computation (CEC 2008), 1-6 June 2008, Hong Kong, pp. 3135 – 3140.

[16] W. M. Korani, “Bacterial foraging oriented by particle swarm optimization strategy for PID tuning,” Proceedings of the 2008 GECCO conference companion on Genetic and evolutionary computation, Atlanta, GA, USA, July 12-16, 2008, pp. 1823 – 1826.

[17] H. Shen, Y. Zhu, X. Zhou, H. Guo, C. Chang, “Bacterial foraging optimization algorithm with particle swarm optimization strategy for global numerical optimization,” Proceedings of 2009 World Summit on Genetic and Evolutionary Computation, Shanghai, China, June 12 – 14, 2009, pp 497 – 504.

[18] R. A. Hooshmand, M. E. Pour, “Corrective action planning considering FACTS allocation and optimal load shedding using bacterial foraging oriented by particle swarm optimization algorithm,” Turkish Journal of Electrical Engineering and Computer Sciences, **18**(4), 2010, pp. 597 – 612.

[19] Y. Fang, Y. Liu, J. Liu, “A novel simplified foraging optimization algorithm for parameter identification of nonlinear system model,” Proceedings of the IEEE International Conference on Automation and Logistics, Jinan, China, August 18 - 21, 2007, pp. 798 – 802.

[20] K. M. Passino, “Biomimicry of bacterial foraging for distributed optimization and control,” IEEE Control System Magazine, June 2002, pp. 52 - 67.

[21] B. Niu, H. Wang, L. J. Tan, J. Xu, “Multi-objective optimization using BFO algorithm,” Springer-Verlag Berlin Heidelberg, 2012, pp. 582 - 587.

[22] J. M. Martins, Z. Mohamed, M. O. Tokhi, J. Sa da Costa and M. A. Botto, “Approaches for dynamic modeling of flexible manipulator system,” IEE Proceedings-Control theory and Application, vol. **150**(4), 2003, pp. 401 - 411.

[23] M.A. Ahmad, A. N. K. Nasir, R.M.T Raja Ismail, and M. S. Ramli, “Comparison of hybrid control schemes for vibration suppression of flexible robot manipulator,” International Conference on Computer Modeling and Simulation 2009 (ICCMS ‘09), Macau, China, February 20- 22, 2009, pp. 356 – 360.

[24] Z. Mohamed and M. A. Ahmad, “Hybrid input shaping and feedback control scheme of a flexible robot manipulator,” Proceeding of the 17th World Congress The International Federation of Automatic Control, Seoul, Korea, July 6-11, 2008, pp. 11714 – 11719.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# State–dependent control of a hydraulically–actuated nuclear decommissioning robot

David Robertson
Engineering Department
Lancaster University, UK
Email: d.robertson@lancaster.ac.uk

C. James Taylor
Engineering Department
Lancaster University, UK
Email: c.taylor@lancaster.ac.uk

***Abstract*—This article develops and evaluates state–dependent parameter (SDP) control systems for the hydraulically actuated dual–manipulators of a mobile nuclear decommissioning robot. A unified framework for calibration, data collection and SDP model identification is proposed, in which the state–dependent variable is a delayed voltage signal associated with the time–varying gain of the system. The latter can cause undesirable joint movements when the device is regulated using linear control algorithms. By contrast, the present article develops a novel nonlinear pole assignment algorithm based on the SDP model. Closed–loop experimental data shows that the SDP design more closely follows the joint angle commands than the equivalent linear algorithm, offering improved resolved motion.**

## I. Introduction

The UK nuclear legacy comprises a number of facilities that are significantly contaminated by radioactivity and non–radiological toxins, are sometimes in a relatively poor state of repair and for which knowledge of their use can be incomplete. In fact, the regulatory requirements were very different when first constructed, with the result that many facilities were not designed with decommissioning strategies in mind. In areas of significant contamination, it is necessary to resort to the use of remote and teleoperated mobile robots. These provide an invaluable option for the safe retrieval and disposal of contaminated materials in high–hazard legacy facilities [1].

Mobile robots are used in many hazardous environments, including explosive ordnance disposal, military reconnaissance, natural disaster search and rescue, and in the nuclear decommissioning sector. In the early stages of nuclear clean–up, expensive, bespoke machines were designed, built and commissioned. However, these have suffered from reliability problems and are usually restricted to specific tasks. More recently, off–the–shelf remote solutions are striven for but these lack the ease of control afforded by high–specification bespoke solutions. The research described in this article aims to alleviate this unsatisfactory situation by developing optimized, widely applicable control architectures, that are being tested on an off–the–shelf robotic platform.

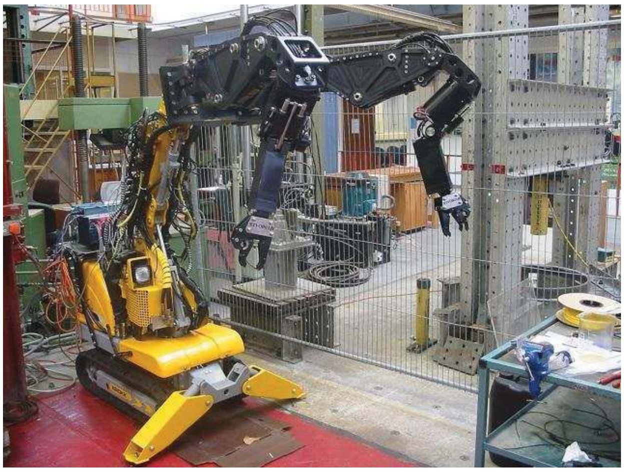


Fig. 1. Brokk–40 and dual HydroLek–7W manipulators.

The research utilizes a Brokk–40 demolition robot, consisting of a moving vehicle with a single manipulator. Two seven–function HydroLek–HLK–7W robotic arms have been attached to the Brokk, as shown by Fig. 1. Such dual–arm mobile robots now offer a powerful and versatile tool for various types of decommissioning activity [2]. Unfortunately, devices initially developed for heavy lifting are not necessarily suitable for 'soft touch' duties such as picking up relatively fragile objects or accurately aligning the end effectors. Indeed, the manipulator can suffer from a relatively slow control action because of limitations in existing linear methods.

Since the behaviour of hydraulically–driven manipulators is dominated by the nonlinear, lightly–damped dynamics of the actuators, high performance control depends on the introduction of some type of nonlinear model structure. Research embraces approaches such as sliding mode [3], adaptive [4], quasi–linear parameter varying [5] and state–dependent parameter (SDP) design [6], among others. An earlier article considered SDP control of the HydroLek but was limited to simulation and did not consider resolved motion [7].

By contrast, the present article utilises open and closed–loop experimental data to investigate potential state dependencies and resolved motion. Here, the nonlinear system is modelled using the quasi–linear SDP structure [8]. For the HydroLek, a novel non–minimal state variable feedback 'regulator' is adapted from the nonlinear pole assignment algorithm of [9]. The data collection and kinematics are described in sections II and III, followed in section IV by an overview of the control design method. Finally, the experimental results and conclusions are discussed in sections V and VI.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

## II. Mobile Robot Platform

The hardware arrangement was developed at Lancaster University from components supplied by Brokk UK Ltd and HydroLek Ltd [2]. The Brokk–40 base machine is 650mm wide, allowing for access through narrow doorways. The five Degrees–Of–Freedom (DOF) manipulator is usually equipped with a variety of tools, including percussive breakers, hydraulic crushing jaws, excavating buckets and concrete milling heads. The unit is electrically powered to facilitate internal use, with an onboard hydraulic pump to power the caterpillar tracks and, by means of several hydraulic pistons, the manipulator.

For decommissioning tasks, accessing the robot on–site to change tools could be a slow, laborious and potentially hazardous task. A more flexible system with the ability to perform multiple tasks without a tool change has been achieved with the mounting of two HydroLek–HLK–7W, seven DOF manipulators, each consisting of six rotational joints and a gripper. The combined system is shown in Fig. 1, whilst Fig. 2 illustrates the azimuth yaw, shoulder pitch, elbow pitch, forearm roll and wrist pitch joints. These are fitted with potentiometer feedback sensors, allowing the position of the end–effector to be determined during operation. The joints are actuated via hydraulic pistons, which are powered via an auxiliary output from the hydraulic pump of the Brokk unit.

A standard input device, such as a joystick, is connected to a PC running a graphical user interface developed for the NI Labview software environment. The PC transmits information to a NI Compact Fieldpoint Real–Time controller (cFP) via an Ethernet networking connection. The cFP is a stand alone device running a real–time operating system, allowing for the precise sampling rates needed for discrete–time control. The hydraulic pistons are controlled by seven pairs of control valves, where each pair has an input for both positive and negative flow. Output modules convert the digital cFP signal to a varied voltage fed to the control valves.

The present authors have developed a semi–automated system for calibrating and initialising the robot for open–loop data collection [10]. Here, the robot is manipulated into a suitable configuration using ‘de–tuned’ proportional control systems. The operator selects from classical step experiments or pseudo–random signals for the estimation of nonlinear models. Figure 3 summarises the basic control system for a single joint. The system input $u_k$ is scaled to lie in the range -100, representing the maximum power in a negative ‘closing’ direction, through to +100, representing the maximum power in a positive ‘opening’ direction. The dead–band of the system is eliminated by the input calibration step, hence an input of zero represents no movement. The output $y_k$ is the potentiometer voltage, representing a scaled joint angle.

## III. Inverse Kinematics

Since it has no bearing on the end–effector location, the gripper is neglected from the kinematic solution illustrated in Figure 4. Gripper control will be considered separately on the basis of pressure feedback and is beyond the scope of the present article. Hence, each manipulator is described as a kinematic model with six solid links and rotational joints. In fact, the manipulator is over–specified in terms of end–effector positioning, with the additional degrees–of–freedom potentially utilised to allow the end–effectors orientation to be set and to reach past obstacles (for example, to pick up a contaminated object from a container/skip). The system is non–trivial to solve for a closed–form, hence incremental steps in complexity are utilised for testing control performance.

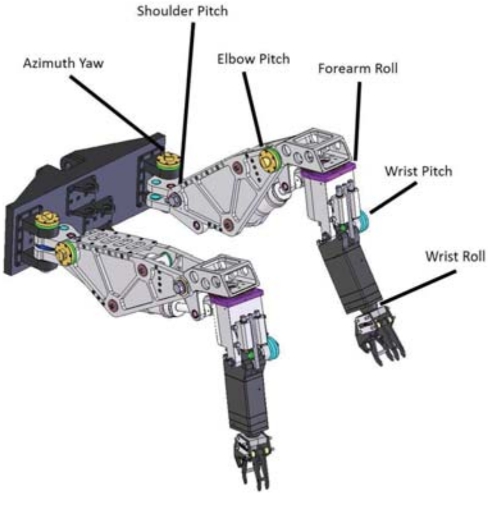


Fig. 2. HydroLek–7W rotational joints.

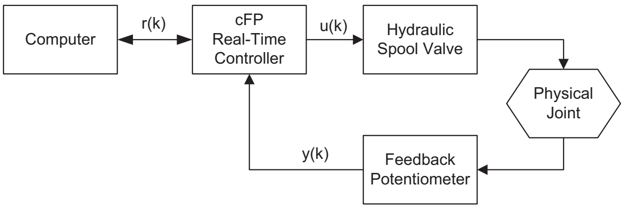


Fig. 3. Control system overview for a single joint.

The present research concentrates on the development of novel joint control systems and initially evaluates these using a reduced 2–DoF system that allows for straightforward movement of the end–effector in a plane. Figure 5 shows how the manipulator is limited to a straightforward 2–DoF system for Joints 2 and 3, with the remaining joints locked off. Equations (1) and (2) describe the forward kinematic relationship of $P_x$ and $P_z$ to the joint angles $\theta_1$ and $\theta_2$, for a generic 2–DoF planer robot, where $P_x$ and $P_z$ represent the horizontal and vertical positions of the end–effector respectively.

$$P_x = l_2 c_{12} + l_1 c_1 \tag{1}$$

$$P_z = l_2 s_{12} + l_1 s_1 \tag{2}$$

Here, $c_1$ and $s_1$ denote $\cos(\theta_1)$ and $\sin(\theta_1)$ respectively, $c_{12}$ and $s_{12}$ represent $\cos(\theta_1 + \theta_2)$ and $\sin(\theta_1 + \theta_2)$, while $l_1$ and $l_2$ are link lengths. Equations (3) and (4) show how this description is applied to the manipulator geometry in Figure 5,

$$P_x = Rc_{23-r} + a_2 c_2 + a_1 \tag{3}$$

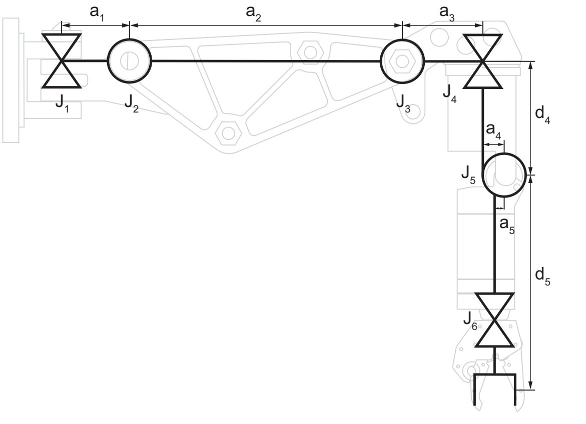


Fig. 4. 6–DOF kinematic description of HydroLek–7W Manipulator.

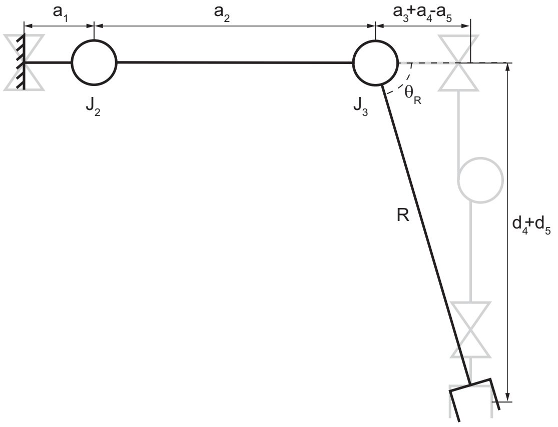


Fig. 5. 2–DOF description of HydroLek–7W Manipulator.

$$P_z = Rs_{23-r} + a_2 s_2 \tag{4}$$

in which $\theta_2$ and $\theta_3$ represent the shoulder pitch and elbow pitch respectively (Figure 2), $\theta_R = \text{atan2}(a_3+a_4-a_5, d_4+d_5)$, $R = \sqrt{(a_3+a_4-a_5)^2+(d_4+d_5)^2}$ and $c_{23-r}$ represents $\cos(\theta_2+\theta_3-\theta_R)$. The inverse kinematics are subsequently derived as follows [11],

$$c_3 = \frac{(P_x-a_1)^2+P_z^2-R^2-a_2^2}{2a_2R} \tag{5}$$

$$s_3 = \pm\sqrt{1-c_3^2} \tag{6}$$

$$\theta_3 = \text{atan2}\,(s_3, c_3) + \theta_R \tag{7}$$

Similarly,

$$c_2 = \frac{(P_x-a_1)(Rc_3+a_2)+P_zRs_3}{d} \tag{8}$$

$$s_2 = \frac{P_z(Rc_3+a_2)-(P_x-a_1)Rs_3}{d} \tag{9}$$

where $d = (Rc_3+a_2)^2 + (Rs_3)^2$. Hence, $\theta_2$ is solved by,

$$\theta_2 = \text{atan2}\,(s_2, c_2) \tag{10}$$

## IV. Nonlinear pole assignment

Consider the deterministic form of the SDP model:

$$y_k = \mathbf{w}_k^{\mathrm{T}}\mathbf{p}_k \tag{11}$$

where $\mathbf{w}_k^{\mathrm{T}}$ is a vector of lagged input and output variables and $\mathbf{p}_k$ is a vector of SDP parameters, defined as follows,

$$\begin{aligned}\mathbf{w}_k^{\mathrm{T}} &= \begin{bmatrix} -y_{k-1} & \cdots & -y_{k-n} & u_{k-1} & \cdots & u_{k-m}\end{bmatrix}\\ \mathbf{p}_k &= \begin{bmatrix} a_1\{\chi_k\} & \cdots & a_n\{\chi_k\} & b_1\{\chi_k\} & \cdots & b_m\{\chi_k\}\end{bmatrix}^{\mathrm{T}}\end{aligned}$$

Here $y_k$ is the output and $u_k$ the control input, while $a_i\{\chi_k\}\,(i=1,2,\ldots,n)$ and $b_j\{\chi_k\}\,(j=1,\ldots,m)$ are $n$ and $m$ state dependent parameters. The latter are assumed to be functions of a non–minimal state vector,

$$\chi_k^{\mathrm{T}} = \begin{bmatrix}\mathbf{w}_k^{\mathrm{T}}\ \mathbf{U}_k^{\mathrm{T}}\end{bmatrix} \tag{12}$$

in which $\mathbf{U}_k = [U_{1,k}, U_{2,k}, \ldots, U_{r,k}]$ is a vector of measured variables, potentially including other joint angles. Any pure time delay $\tau \geq 1$ is represented by setting the leading $b_1\{\chi_k\}\ldots b_{\tau-1}\{\chi_k\}$ terms to zero. To identify and estimate this model from experimental data, the authors have used the back–fitting approach of references [8] and [12].

Using this model, the first author and colleagues have developed a novel method for nonlinear pole assignment of SDP systems that guarantees closed–loop stability at the design stage: see references [6] and [9] for details. The approach is based on the definition of a suitable non–minimal state space (NMSS) form. For the present research, the algorithm has been modified to handle the integrating joint angle dynamics. In particular, the dead–zone calibration routine ensures that there is no movement when $u_k = 0$, hence external integral action is not required. In fact, the destabilizing nature of integral action has a negative impact on the control performance. For this reason, the following 'regulator' NMSS form is utilised:

$$\mathbf{x}_{k+1} = \mathbf{F}\{\chi_k\}\mathbf{x}_k + \mathbf{g}u_k \quad ; \quad y_k = \mathbf{h}\mathbf{x}_k \tag{13}$$

where the $n+m-1$ dimensional state vector is,

$$\mathbf{x}_k = \begin{bmatrix} y_k & \cdots & y_{k-n+1} & u_{k-1} & \cdots & u_{k-m+1}\end{bmatrix}^T \tag{14}$$

and the state matrices are defined to satisfy the SDP model (11), as shown by the example in section V. The state variable feedback control algorithm is,

$$u_k = -\mathbf{v}\{\chi_k\}\mathbf{x}_k + k_0\{\chi_k\}r_k \tag{15}$$

where,

$$\mathbf{v}\{\chi_k\} = [f_0\{\chi_k\}\ldots f_{n-1}\{\chi_k\}\,g_1\{\chi_k\}\ldots g_{m-1}\{\chi_k\}]$$

is a vector of scheduled control gains, $r_k$ is the command input and $k_0\{\chi_k\} = f_0\{\chi_k\} + \ldots + f_{n-1}\{\chi_k\}$. A vector $\mathbf{c}\{\chi_k\}$ is defined with a similar structure to $\mathbf{v}\{\chi_k\}$ but in which $\bar{g}_i\{\chi_k\}\,(i = 1\ldots m-1)$ differ from $g_i\{\chi_k\}$ as follows,

$$g_i\{\chi_k\} = \frac{b_\tau\{\chi_{k+\tau-i}\}}{b_\tau\{\chi_{k+\tau}\}}\cdot\bar{g}_i\{\chi_k\} \tag{16}$$

Here, $\mathbf{c}\{\boldsymbol{\chi}_k\}$ is determined by defining a suitable matrix $\Sigma_k\{\boldsymbol{\chi}_k\}$ of model coefficients and solving,

$$\Sigma\{\boldsymbol{\chi}_k\}\cdot\mathbf{c}\{\boldsymbol{\chi}_k\}=\mathbf{d}-\mathbf{p}\{\boldsymbol{\chi}_k\} \tag{17}$$

in which $\mathbf{p}\{\boldsymbol{\chi}_k\}$ and $\mathbf{d}$ are the open–loop and desired (time–invariant and stable) closed–loop coefficients respectively [6]. For brevity, full definitions of the above matrices are omitted but illustrative solutions for the HydroLek manipulator are presented in section V. The scaling (16) is necessary in order to achieve the desired stable closed–loop response and imposes a clear limitation on $b_\tau\{\boldsymbol{\chi}_{k+\tau}\}$. In fact, $b_\tau\{\boldsymbol{\chi}_{k+\tau}\}\neq 0$ corresponds directly to the pole assignability condition [6].

## V. Results

Preliminary open–loop step experiments (not shown) suggest that a first order linear difference equation, i.e.,

$$y_k=-a_1y_{k-1}+b_\tau u_{k-\tau} \tag{18}$$

provides an approximate representation of individual joints, with time–invariant parameters $\{a_1,b_\tau\}$ and the time delay $\tau$ depending on the sampling interval $\Delta t$. Such models have previously been utilized for the development of linear control systems for large scale hydraulic machinery in the construction industry [13]. However, further analysis quickly reveals limitations in the linear model. For the HydroLek, it is readily apparent that the numerical values of $\{a_1,b_\tau\}$ are not repeatable for different step experiments. Most notably, the value of $b_\tau$ depends on the magnitude of the applied voltage utilized for these step experiments, i.e. it is a SDP.

### A. System identification

Various candidate SDP structures have been investigated for open–loop movement of the right hand side manipulator shoulder and elbow pitch, moving in air with no additional loading terms, i.e. no objects are held by the gripper. For these conditions, trial and error experimentation suggests that a sampling interval of $\Delta t=0.07$ seconds yields a satisfactory compromise between a fast response and the following low–order model for control system design,

$$y_k=-a_1\{y_{k-1}\}\,y_{k-1}+b_2\{u_{k-2}\}\,u_{k-2} \tag{19}$$

The model is based on equation (11) with $n=1$, $m=\tau=2$, $b_1\{\boldsymbol{\chi}_k\}=0$ and initially $\boldsymbol{\chi}_k=[y_{k-1}u_{k-2}]$. In fact, the statistical estimates suggest that $a_1\{\boldsymbol{\chi}_k\}$ is relatively constant over time, with the pole close to unity. Assuming $a_1=-1$, the system essentially behaves as an integrator. In this case, $b_2\{\boldsymbol{\chi}_k\}$ represents the state–dependent angular velocity, approximated by the following second order polynomial,

$$b_2\{u_{k-2}\}=p_1u_{k-2}^2+p_2u_{k-2}+p_3 \tag{20}$$

The coefficients in equation (20) are optimized from open–loop experiments using fminsearch in Matlab®, yielding the estimates shown in Table I. The model responses (18) and (19) are compared with measured data for an illustrative open–loop shoulder joint experiment in Fig. 6. The SDP model is far superior, typically explaining over 95% of the output variance, compared with only 20–60% for the linear model.

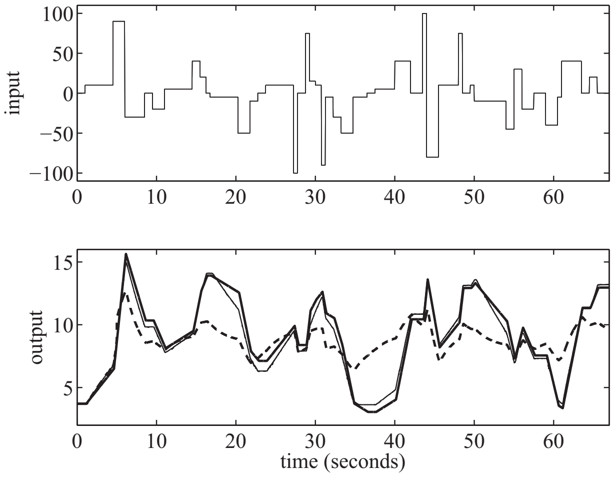


Fig. 6. Upper subplot: applied input voltage plotted against time. Lower subplot: potentiometer voltage, showing the optimized SDP model response (thick trace), experimental data (thin) and linear model (dashed).

TABLE I
Optimized polynomial coefficients

| Joint | $p_1$ | $p_2$ | $p_3$ |
|---|---|---|---|
| $\theta_2$ | $-2.2898e^{-007}$ | $1.2576e^{-005}$ | 0.0055 |
| $\theta_3$ | $-1.9290e^{-007}$ | $1.8214e^{-005}$ | 0.0051 |

### B. Control system design

The NMSS/SDP (13) representation of equation (19) is,

$$\mathbf{x}_k=\begin{bmatrix}y_k\\u_{k-1}\end{bmatrix}=\begin{bmatrix}-a_1 & b_2\{u_{k-2}\}\\0 & 0\end{bmatrix}\mathbf{x}_{k-1}+\begin{bmatrix}0\\1\end{bmatrix}u_{k-1} \tag{21}$$

and $y_k=\begin{bmatrix}1 & 0\end{bmatrix}\mathbf{x}_k$. For $b_2\{u_{k-2}\}>0$, the matrix $\Sigma\{\boldsymbol{\chi}_k\}$ in equation (17) can always be inverted, i.e.,

$$\begin{bmatrix}f_0\{\boldsymbol{\chi}_k\}\\\bar{g}_1\end{bmatrix}=\begin{bmatrix}0 & 1\\b_2\{u_{k-2}\} & a_1\end{bmatrix}^{-1}\begin{bmatrix}-p_1-p_2-a_1\\p_1p_2\end{bmatrix} \tag{22}$$

in which $(z-p_1)(z-p_2)=z^2-(p_1+p_2)z+p_1p_2$ is the desired closed–loop characteristic polynomial, with the poles $\{p_1,p_2\}$ chosen to lie within the unit circle of the complex $z$–plane. For the illustrative results considered in Figures 7, 8 and 9, $p_1=0.9$ and $p_2=0.94$ are chosen by trial and error simulation. Utilising (16), the control algorithm (15) is,

$$u_k=-f_0\{\boldsymbol{\chi}_k\}\,y_k-\frac{b_2\{u_{k-2}\}}{b_2\{u_{k-1}\}}g_1u_{k-1}+k_0\{\boldsymbol{\chi}_k\}\,r_k \tag{23}$$

where $f_0\{\boldsymbol{\chi}_k\}=k_0\{\boldsymbol{\chi}_k\}=(p_1p_2-g_1)/b_2\{u_{k-2}\}$ and $g_1=-p_1-p_2-a_1$. Note that $g_1$ is time–invariant in this case. For the HydroLek, $a_1=-1$ and $b_2\{u_{k-2}\}$ is defined by equation (20). Finally, three linear control systems are also considered. Experimental results suggests that the most robust linear algorithm is based on an operating level of $u_k=0$, hence utilising equation (20) and Table I, $b_2=0.0055$ and $b_2=0.0051$ for the shoulder and elbow joints respectively. For comparison, linear controllers are also developed for the maximum input signal each direction, i.e. equation (20) with $u_k=-100$ and $u_k=100$. In each case, equation (22) is solved off–line for these time–invariant coefficients.

### C. Experimental results

Typical practical results for the NMSS/SDP controller are illustrated by Fig. 7, showing a satisfactory response across a range of operating levels. The linear controller based on an operating level of $u_k = 0$ or 100 yields a slower response (not shown), since the model utilised tends to 'overestimate' the steady state gain of the system. By contrast, the linear controller based on $u_k = -100$ works adequately for large negative steps in the command (since these are associated with large negative values of the input) but tends to overshoot the set point at other times, as illustrated by Fig. 8.

Here, it should be emphasized that the control objective is to follow the particular response specified by the design poles. Although arbitrarily chosen for the present article, which focuses on the low–level joint control problem, the poles and command sequences are generally determined by the higher–level control module for optimising tasks using the attached tools. Hence, the faster speed of response and overshoot for the linear design in Fig. 8 represents an undesirable deviation from the required behaviour of the system.

Fig. 9 shows this error between the experimentally observed joint angle and the ideal or 'design' response. The latter is obtained by simulating a transfer function with the design poles in open–loop. In Fig. 9, the grey and black shading highlights the errors for the SDP and linear controllers respectively. Hence the black colour represents the *additional* errors associated with the linear controller in each case, demonstrating the improved performance of the NMSS/SDP design against all three linear controllers. The mean absolute errors between the joint angle and ideal response for each of the controllers (i.e. SDP and the linear controllers associated with each operating level) are summarised in Table II, for experiments similar to Fig. 7 and Fig. 8.

Finally, Figures 10 and 11 illustrate a resolved motion experiment based on the kinematic equations discussed in section III. Here, the end–effector is programmed to trace a circle in a clockwise motion, followed by a second lower circle in an anticlockwise direction, with point–to–point motion between. In fact, the circular movement is also based on point–to–point motion, with the trajectory defined by a series of small steps in the positional set point for $[p_x, p_z]$. The speed of response is determined by the number of iterations and the waiting time for each point. This straightforward approach to positional trajectory planning appears to work well in practice for hydraulically operated robotic manipulators [13].

Although the linear and nonlinear responses in Figure 10 are visually very similar, small joint angle errors are propagated in relation to the position of the end–effector for these resolved motion experiments. Once the experimental data are displayed in Cartesian form on the work–plane, as in Figure 11, the potential benefits of high performance nonlinear control become more apparent, particularly when relatively fast movement is required. For the illustrative results shown, each circle was programmed to be completed in 6 seconds, with the experiment repeated three times in Figure 11.

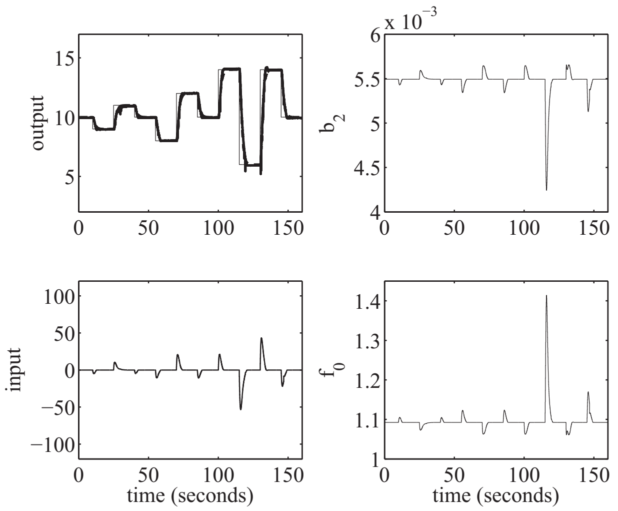


Fig. 7. Clockwise from upper left: closed–loop experiment showing the response of joint angle (thick trace) to a sequence of step changes in the command (thin), $b_2\{\chi_k\}$, $f_0\{\chi_k\}$ and $u_k$, all plotted against time.

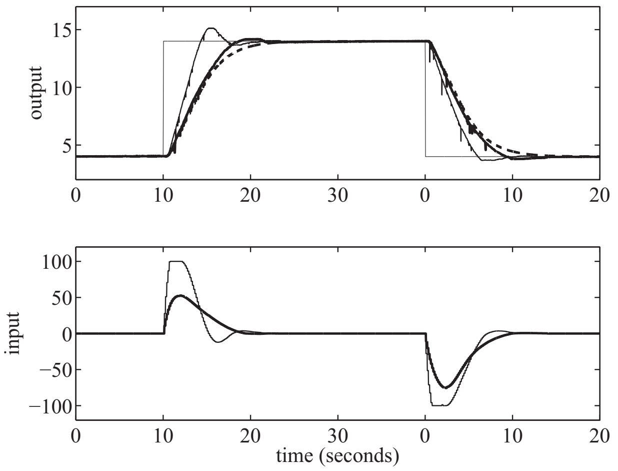


Fig. 8. Closed–loop response of joint angle to positive (joint opening) and negative (closing) step changes in the command input, showing experimental data obtained using the linear (thin trace) and nonlinear (thick) controllers, compared to the simulated design response (dashed), all plotted against time. Upper subplot: potentiometer voltage. Lower subplot: applied input voltage.

## VI. Conclusions

This article has considered state–dependent parameter (SDP) models and non–minimal state space (NMSS) control design for the hydraulically actuated joints of a mobile robot designed for nuclear decommissioning. The analysis suggests that a univariate SDP model with state–dependent gain, representing the angular velocity, is adequate for fast and smooth control of the manipulators. However, the authors are presently investigating other configurations and settings with a view to identifying additional state variables and potential multivariable state–dependencies.

To illustrate the modelling approach and typical closed–loop results, the analysis has concentrated on experimental data associated with the shoulder and elbow pitch of one

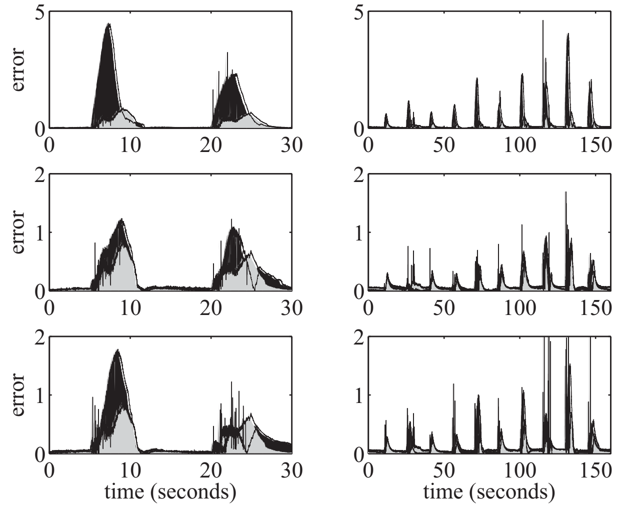


Fig. 9. Errors between the experimentally observed joint angle and the design response, plotted against time, comparing the nonlinear (grey shading) and linear (black) controllers for experiments similar to Fig. 7 (right hand side subplots) and Fig. 8 (left). Upper subplots: linear controller based on operating level $u_k = -100$. Middle subplots: $u_k = 0$. Lower subplots: $u_k = 100$.

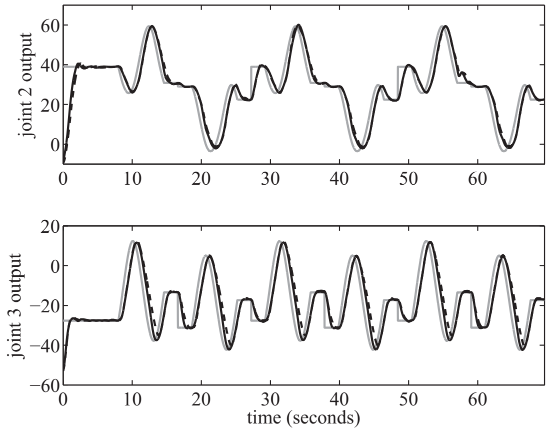


Fig. 10. Closed–loop response for a resolved trajectory, showing experimental data obtained using the linear (dashed) and nonlinear (black) controllers, together with the set point generated by the inverse kinematics (gray), plotted against sample number. Upper subplot: Joint 2. Lower subplot: Joint 3.

manipulator. The authors are now investigating the remaining joints, with a view to the development of a high–level control algorithm for resolved control of the dual–manipulators. In the future, these algorithms are intended to be part of a self–calibrating and self–tuning automatic control system.

## ACKNOWLEDGMENT

The authors are grateful for the support of Brokk UK Ltd. and HydroLek Ltd., who supplied the base robotic systems, and the Nuclear Decommissioning Authority (NDA), who have part–funded this research. The statistical tools have been assembled as the CAPTAIN toolbox within the Matlab® software environment and may be obtained from the second author [12].

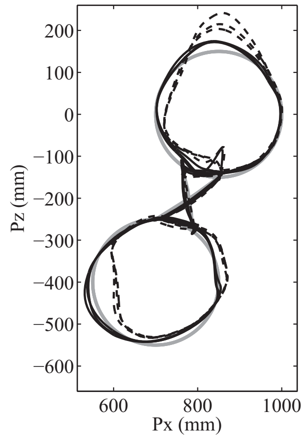


Fig. 11. End–effector location associated with Fig. 10, showing experimental data obtained using the linear (dashed) and nonlinear (black) controllers, together with the set point generated by the inverse kinematics (gray).

TABLE II
MEAN ABSOLUTE ERROR BETWEEN JOINT ANGLE AND IDEAL RESPONSE.

| | $SDP$ | $u_k = -100$ | $u_k = 0$ | $u_k = 100$ |
|---|---|---|---|---|
| Fig. 7 | 0.1640 | 0.5229 | 0.2414 | 0.2400 |
| Fig. 8 | 0.0921 | 0.2289 | 0.1030 | 0.1346 |

## REFERENCES

[1] D. W. Seward, C. M. Pace, and R. Agate, "Safe and effective navigation of autonomous robots in hazardous environments," *Journal of Autonomous Robots*, vol. 22, pp. 223–242, 1999.

[2] M. J. Bakari, D. W. Seward, and C. J. Taylor, "The development of a prototype of a multi–arm robotic system for decontamination and decommissioning applications within the nuclear industry," in *12th International Conference on Environmental Remediation and Radioactive Waste Management*, Liverpool, UK, October 2009.

[3] Q. Ha, D. Rye, and H. Durrant-Whyte, "Fuzzy moving sliding mode control with application to robotic manipulators," *Automatica*, vol. 35, pp. 607–616, 1999.

[4] M.-H. Chiang and H. Murrenhoff, *Adaptive servo–control for hydraulic excavators*, ser. Power Transmission and Motion Control. Professional Engineering Publishing Limited, UK, 1998.

[5] B. Fidan, Y. Zhang, and P. A. Ioannou, "Adaptive control of a class of slowly time varying systems with modeling uncertainties," *IEEE Transactions on Automatic Control*, vol. 50, pp. 915–920, 2005.

[6] C. J. Taylor, A. Chotai, and P. C. Young, "Nonlinear control by input–output state variable feedback pole assignment," *International Journal of Control*, vol. 82, pp. 1029–1044, 2009.

[7] C. J. Taylor, A. Chotai, and D. Robertson, "State dependent control of a robotic manipulator used for nuclear decommissioning activities," in *IEEE International Conference on Intelligent Robots and Systems (IROS)*, Taipei, Taiwan, October 2010.

[8] P. C. Young, P. McKenna, and J. Bruun, "Identification of nonlinear stochastic systems by state dependent parameter estimation," *International Journal of Control*, vol. 74, pp. 1837–1857, 2001.

[9] C. J. Taylor, A. Chotai, and K. J. Burnham, "Controllable forms for stabilising pole assignment design of generalised bilinear systems," *Electronics Letters*, vol. 47, pp. 437–439, 2011.

[10] D. Robertson, "HydroLek data acquisition & controller testing," Engineering Department, Lancaster University, UK, Tech. Rep., 2011.

[11] R. Manseur, *Robot Modelling & Kinematics*. Charles River, 2006.

[12] C. J. Taylor, D. J. Pedregal, P. C. Young, and W. Tych, "Environmental time series analysis and forecasting with the Captain Toolbox," *Environmental Modelling and Software*, vol. 22, no. 6, pp. 797–814, 2007.

[13] E. M. Shaban, S. Ako, C. J. Taylor, and D. W. Seward, "Development of an automated verticality alignment system for a vibro–lance," *Automation in Construction*, vol. 17, pp. 645–655, 2008.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Design and Control of Material Transport System for Automated Guided Vehicle

Wu Xing, Lou Peihuang, Cai Qixiang, Zhou Chidong, Shen ke, Jin chen
College of Mechanical and Electrical Engineering
Nanjing University of Aeronautics and Astronautics
Nanjing, P. R. China
Wustar5353@nuaa.edu.cn

*Abstract*—**Automated guided vehicle (AGV) is used widely in many industrial applications to transport materials. A long-travel material transport system is designed and controlled in this paper for a unit load AGV to transfer the material pallet from AGV to the load stand and back, including a both-side three-level push-pull load-transfer mechanism and a programmable logic controller (PLC). AGV follows the path of the magnetic tapes in the floor, and locates itself according to the RFID tags beside the load stand. The PLC detects the docking position of AGV and the load stand, and controls the horizontal and vertical movement of the load-transfer mechanism by using a stepper motor and an electric push rod. The load transfer experiments of our AGV testify the performance of this material transport system.**

***Keywords-automated guided vehicle; material transport; load transfer; mechanism control***

## I. Introduction

Automated guided vehicle (AGV) is a driverless, steerable, wheeled industrial vehicle that follows markers or wires in the floor, or uses vision or lasers. The largest consumer of AGVs is the automotive industry, yet AGVs are also common in other industries, including warehouses and distribution centers, paper, printing, textiles, and steel industries [1]. In these applications, AGVs have been found to increase routing flexibility, improve space utilization, ensure safety, and reduced overall operational cost [2].

According to the means of load transfer, AGVs can be classified as towing AGV, unit load AGV, pallet truck AGV, and fork truck AGV [3]. Towing AGV was the first type introduced and is still a very popular type today. It can tow single or multiple trailers by using hooks to provide traction. It has a simple structure and a capability of several times more materials than conventional unit load AGV. However, it has a difficulty to control the trailer position and has a larger turning radius. Unit load AGV is equipped with decks, which permits unit load transportation and often automatic load transfer. The decks can either be lift and lower type, powered or non-powered roller, chain or belt decks or custom decks with multiple compartments. Pallet truck AGV is designed to transport palletized loads to and from floor level, and it can eliminate the need for fixed load stands. Fork truck AGV is ideal for applications where automatic load pickup or delivery is required from floor or various height elevations. It has counterbalanced or straddled leg configurations depending on application requirements.

Initially, AGV paths are specified by electric wires placed approximately 1 inch below the ground [4]. It requires the expense of cutting the floor for the entire travel route, and it can not be easily removed and relocated if the course needs to change. Due to these disadvantages, Many AGVs begin to use tape in the floor as the guide path. The tapes can be magnetic, colored or other media. AGV uses the magnetic sensor to detect the magnetic tape and then follow the path [5], or uses the vision sensor (e.g. camera) to identify the colored tape [6-7]. In order to measure the position of AGV on the path, magnetic and visual coded signs can be used as artificial landmarks to express its absolute coordinates. Additionally, radio frequency identification (RFID) is regarded as a promising technology suitable for location and navigation of robot that can achieve a pervasive automation [8]. It can contain a large amount of data coded information compared to magnetic and visual signs.

We intend to design and develop an intelligent AGV guided by the magnetic tapes and located by the RFID tags to achieve the pervasive automation of materials transport and transfer, by combining the high guidance accuracy and setting convenience of magnetic taps with the high information capacity of RFID tags. The pervasive automation requires a close collaboration between vehicle guidance and material transport. This paper presents a long-travel material transport system for a unit load AGV having been equipped with the guidance control system. Section II proposes a material transport control methodology containing docking, retrieving and depositing for the unit load AGV. Section III designs a push-pull load-transfer mechanism that can pick up and down the material pallet in both sides. Section IV develops a PLC-based control system to detect the position of executive mechanisms and control their movements in the load transfer process. Section V presents the load transfer experiments, and a conclusion is given in section VI.

## II. Load Transfer Methodology

Unit load AGV has many kinds of decks (e.g. lift, roller, chain, etc) to transport materials. In the consideration of the transfer requirements of automobile assembly production lines, we suggest a material transport control methodology to locate AGV beside the load stand accurately and retrieve and deposit the pallet automatically.

This work was supported by the National Natural Science Foundation of China (No.61105114 and No. 51175262), and the Research Fund of Nanjing University of Aeronautics & Astronautics (No. NC2012005).

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

### A. Docking

A load stand is the fixed port where the material pallet is transferred from the AGV to the workstation and back. AGV movement near the load stand is controlled differently to its movement on the path since both a higher accuracy and a different obstacle avoidance technique are needed. In the close proximity to the load stand, AGV needs to navigate closer to obstacles and it must reach the designated point with greater precision. In order to achieve the accurate docking distance between AGV and the load stand, the real-time position of AGV needs to be updated continuously and AGV pose needs to be adjusted accurately.

Our AGV navigates on the specified paths by following the magnetic tapes, as shown in Fig.1. A magnetic guide sensor is mounted in the lateral center of AGV body, and it can detect the lateral deviation between AGV and the desired paths. Path tracking control is used to eliminate this deviation and keep AGV on the path. When the load stands are placed accurately beside the paths, the lateral position between AGV and the load stand can be guaranteed by controlling the lateral deviation. On the other aspect, the longitudinal position from AGV to the load stand is also very important to the docking process. It can be controlled by dead reckoning that accumulates the traveling distance by supposing pure rolling of AGV wheels. In fact, dead reckoning data tends to drift due to wheel slippage. It is necessary to reset the dead reckoning counters according to some artificial landmarks before AGV enters the docking area. Here we put several RFID tags on the path near the load stands, as shown in Fig.1. The longitudinal position begins to sum up by using dead reckoning after AGV reads the RFID tags. The RFID sensor is installed in the head of AGV, which is designed as a semi-open structure in order to avoid the shielded metal environment that may block the RFID signals.

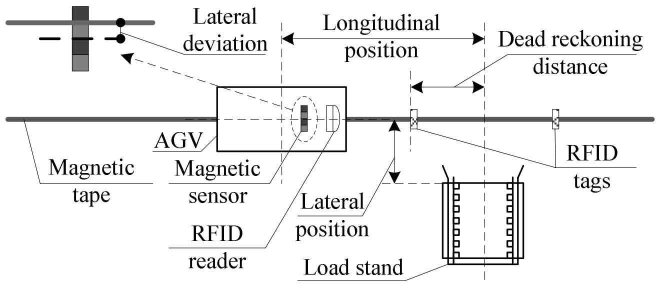


Figure 1. Materials transport path

### B. Retrieving

Retrieving means AGV picks up the material pallet from the load stand. After AGV arrives at the accurate docking spot, it should check whether there is the designated pallet in the load stand. If it exists, AGV begins to pull it by using its load-transfer mechanism. AGV locates itself by making its side facing the front of the load stand, and the pallet needs to be pulled to AGV from its side. AGV should dock in the close proximity to the load stand in the lateral direction to reduce the transfer distance, as shown in Fig.2. Since the width of the load-transfer mechanism is no more than that of AGV body, the retrieving process is proposed as following actions.

(1) The mechanism moves downwards to make its convex blocks below the sockets of the pallet in the vertical direction.

(2) The mechanism moves towards the pallet to make its convex block 2 aligned with socket 1 in the lateral direction.

(3) The mechanism moves upwards to make its convex block 2 inserted into socket 1. At this time the load-transfer mechanism has grasped the pallet.

(4) The mechanism moves backwards to pull the pallet on AGV gradually. At this time block 2 (shown as the dashed lines) has aligned with the initial position of block 1 (shown as the solid lines) in the lateral direction.

(5) The mechanism moves downwards to release the pallet. At this time, the pallet has been placed on AGV.

(6) The mechanism moves towards the pallet to make its convex block 1 and 2 aligned with socket 1 and 2 respectively in the lateral direction. Note that at this time they are aligned on AGV.

(7) The mechanism moves upwards to make its convex block 2 inserted into socket 2 (block 1 also inserted into socket 1). So the load-transfer mechanism can hold the pallet firmly when the pallet is carried in the travel.

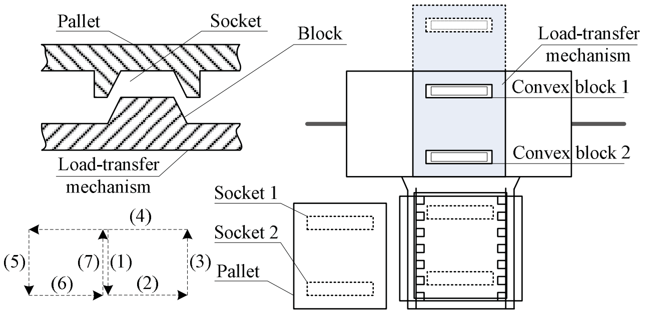


Figure 2. Load pickup approach

### C. Depositing

Depositing means AGV puts down the material pallet to the load stand. AGV should contact the workstation and verify the load stand is free before the material delivery process starts. The pallet is pushed to the load stand from the side of AGV. It is a similar process of retrieving as following actions, shown in Fig.3.

(1) The mechanism moves downwards to depart from its initial position, shown as step (1) in Fig.3.

(2) The mechanism moves away from the pallet to make its convex block 2 aligned with socket 1 in the lateral direction, shown as step (2) in Fig.3.

(3) The mechanism moves upwards to make its convex block 2 inserted into socket 1. At this time the load-transfer mechanism has grasped the pallet, shown as step (3) in Fig.3.

(4) The mechanism moves towards the load stand to push the pallet on it gradually, shown as step (4) in Fig.3.

(5) The mechanism moves downwards to release the pallet. At this time, the pallet has been placed on the load stand, shown as step (5) in Fig.3.

(6) The mechanism moves backwards to align itself with AGV body in the lateral direction, shown as step (6) in Fig.3.

(7) The mechanism moves upwards to return its initial position that is always kept when AGV follows the paths.

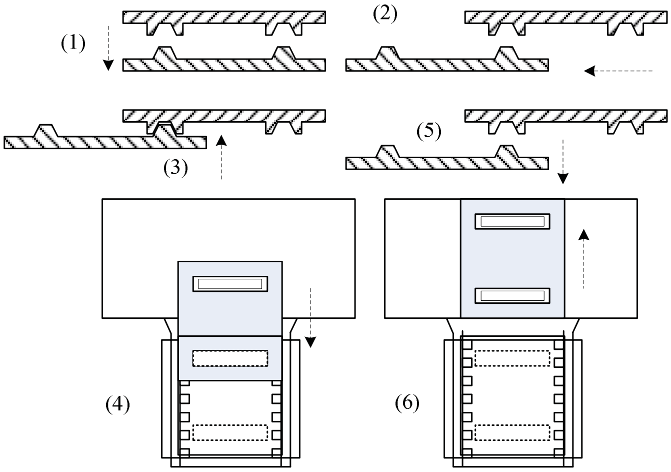


Figure 3. Load putdown approach

## III. LOAD-TRANSFER MECHANISM DESIGN

In order to execute the retrieving and depositing behaviors required in the long-travel materials transport process, a both-side three-level push-pull load-transfer mechanism is designed. This mechanism contains a lifting module and a translational module, as shown in Fig.4.

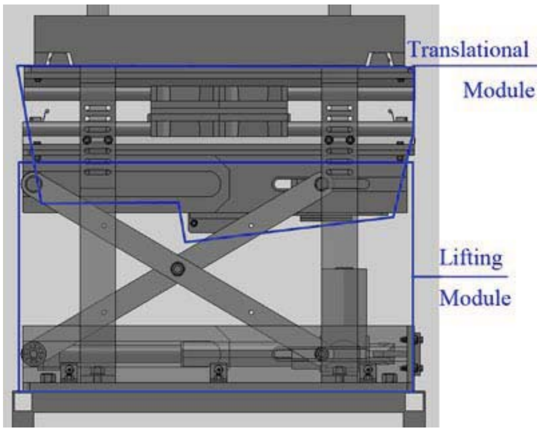


Figure 4. Load-transfer mechanism

### A. *Lifting module*

Lifting module is able to change the vertical height of the platform by using a cross-shaped mechanism, as shown in Fig.5. Two rods are linked at the middle point by using a hinge 3 to form a cross shape. Hinge 1 and rail 1 are fixed on the upper surface of AGV body. Hinge 2 and rail 2 are fixed on the lower surface of the platform. An electric push rod mounted on AGV surface impels slider 1 to make it move forwards and backwards along rail 1. When slider 1 moves forwards, hinge 3 is pulled to move along arc 1, slider 2 to move along arc 2, and hinge 2 to move downwards in the vertical direction, shown as dashed shapes in Fig.5. As a result, the platform follows their movements and lowers itself. When slider 1 moves backwards, these parts moves in reverse trajectories, and their movements lift the platform consequently. In order to control two limit positions of the platform in the vertical direction, two travel switches can be placed at two ends of rail 1 accurately.

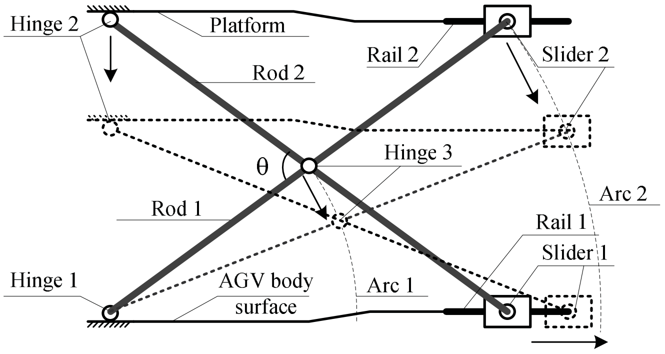


Figure 5. Lifting module

### B. *Translational module*

Translational module is fixed on the platform and moved by lifting module upwards and downwards. This module is responsible for the lateral movement of their convex blocks by using a three-level planar mechanism, as shown in Fig.6. The low-level plane is the platform, and a stepper motor is installed on its lower surface. Gear 1 is linked with the motor shaft by using a key, and gear 2 is supported by a stationary mandrel. On its upper surface, rail 1 is fixed near the right border, and rack 2 is in its middle line. The middle-level plane (plane 2) has slider 1 and slider 2 fixed on its upper and lower surfaces symmetrically, and has gear 3 supported by another stationary mandrel in the center of its upper surface. The high-level plane (plane 3) has rail 2 mounted near the right border on its lower surface, and rack 3 in its middle line. Convex block 1 and 2 are located on its upper surface, as shown in Fig.2.

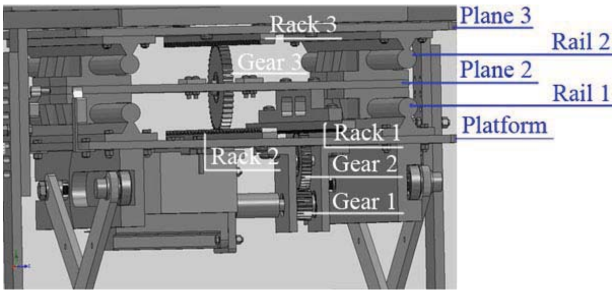


Figure 6. Translational module

The stepper motor drives gear 1 via the connecting key, and then gear 1 makes gear 2 roll on rack 1 by engaging tooth. Since the platform is fixed with lifting module, plane 2 moves relatively to the platform in the lateral direction. Gear 3 is pushed forwards by plane 2, and it rolls on the lower rack 2 of the platform and on the upper rack 3 of plane 3 simultaneously. If plane 2 is selected as the reference coordinate system, plane 2 and plane 3 have an opposite translational movement with the same speed rate. Therefore, when the platform is stationary, if plane 2 has a translational speed and displacement in one direction, plane 3 will have a double speed and displacement in the same direction. It is the basic principle on which the large-travel translational movement is implemented based by using the three-level planar mechanism with gears and racks. In order to control two limit positions and one initial position of plane 3 in the lateral direction, three travel switches can be placed at two ends and in the middle of rail 1 accurately.

## IV. Control System Development

A Siemens PLC S7-224XP CN is adopted as the controller of the material transport system, in charge of controlling the lateral movement of translational module by using the stepper motor, regulating the vertical movement of lifting module by using the electric push rod, and communicating with the on-board controller of AGV, as shown in Fig. 7.

The PLC has two high-speed impulse output ports Q0.0 and Q0.1, which can generate a PWM pulse or a PTO pulse. PWM means the duty cycle is variable by fixing pulse cycle and changing pulse width. They are accumulated by counters with the maximum value of 65535 and based on time unit of micro-second or millisecond. PTO is a pulse sequence of square wave with a fixed duty cycle of 50%. Their pulse number can change in the range of 10~65535μs or 2~65535ms. The stepper motor is controlled by a PTO pulse from Q0.0 and a direction signal from Q0.1. The PLC sends a pulse to make the stepper motor rotate over a step angle, and then checks whether the switch is triggered by the translational module.

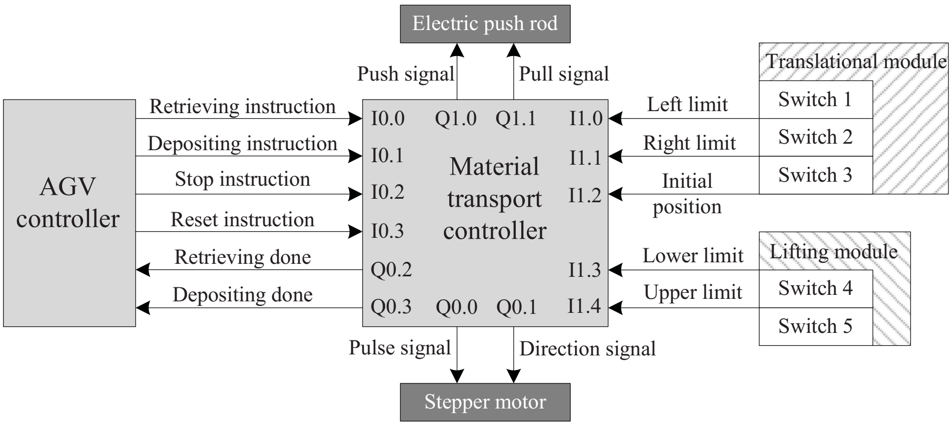


Figure 7. Material transport control system

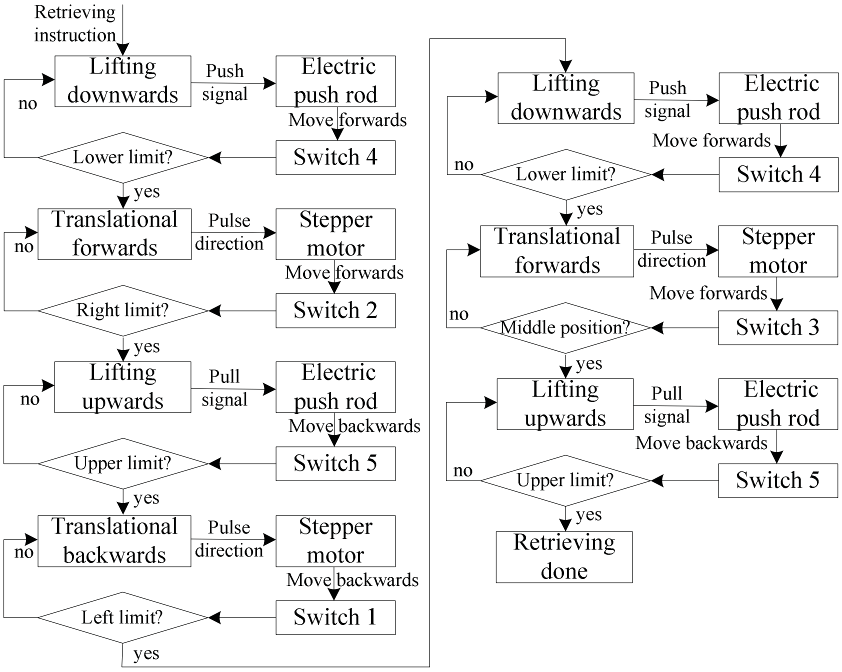


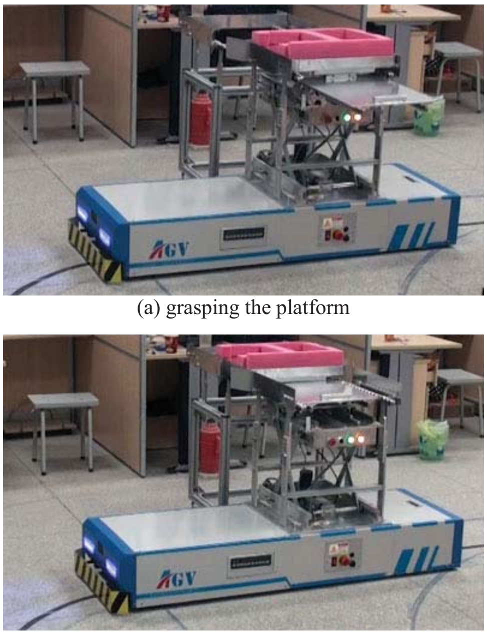
Figure 8. Retrieving control diagram

After AGV arrives at the accurate docking spot and checks its pallet target, AGV controller sends a retrieving instruction to the PLC that starts the entire control process, as shown in Fig8. Firstly, the PLC gives a push signal to the electric push rod, and it pushes sliders forwards to lower lifting module. When switch 4 sends a lower limit signal back to the PLC, it stops the push rod, and the convex blocks have moved below the sockets. Secondly, the PLC gives a forward pulse sequence to the stepper motor. The motor drives plane 2 and 3 towards the load stand. When switch 2 sends a right limit signal back to the PLC, it stops the stepper motor, and convex block 2 has aligned with socket 1. Thirdly, the PLC gives a pull signal to the electric push rod, and it pulls sliders backwards to raise lifting module. When switch 5 sends an upper limit signal back to the PLC, it stops the push rod, and convex block 2 has inserted into socket 1. Fourthly, the PLC gives a backward pulse sequence to the stepper motor. The motor drives plane 2 and 3 backwards AGV. When switch 1 sends a left limit signal back to the PLC, it stops the stepper motor, and block 2 has aligned with the initial position of block 1 in the lateral direction. Fifthly, the PLC gives a push signal to the electric push rod until switch 4 feeds back a lower limit signal. Convex block 2 has retracted from socket 1 and plane 3 has released the pallet. Sixthly, the PLC gives a forward pulse sequence to the stepper motor until switch 3 feeds back a middle position signal. Convex block 1 and 2 has aligned with socket 1 and 2 respectively in the lateral direction. Lastly, the PLC gives a pull signal to the electric push rod until switch 5 detects an upper limit signal. Convex block 1 and 2 has inserted into socket 1 and 2, and plane 3 has grasped the pallet again.

## V. Load Transfer Experiments

In order to test the material transport system of AGV, the load transfer experiments are carried out in our laboratory. A ring closed path is laid out in the floor by using magnetic tapes. One load stand is placed on a designated spot accurately beside the path, which is used as a load pickup workstation as well as a load delivery workstation. Firstly, AGV starts at an arbitrary point of the path and moves towards the load stand. After it arrives at the accurate docking spot, it picks up the pallet from the load stand automatically. Then it continues to run on the ring path and comes back to the load stand when it finishes the circle of path. After it checks the load stand is free, it delivers the pallet to the load stand automatically, as shown in Fig.9. Fig.9.(a) shows the grasp operation that plane 3 holds the platform by inserting convex block 2 into socket 1 when the electric push rod pulls sliders backwards to raise lifting module. Fig.9.(b) shows the push operation that plane 3 pushes the platform to the load stand when the stepper motor drives rack 1 and 3 move forwards.

In the load transfer experiments, retrieving and depositing operations of AGV are executed continuously in a periodic way up to 8 hours. This satisfactory experiment result can only be achieved when the material transport system already has two control capacities at the same time. One is that AGV can locate itself accurately on the same spot beside the load stand at each time. So the longitudinal position and the lateral position between AGV and the load stand can be guaranteed strictly. Usually, the longitudinal position error is less than ±15mm, and the lateral position error is less than ±10mm. The other is that the load-transfer mechanism can arrive at the precise position to grasp the platform, push the platform to the correct position of the load stand, and pull the platform back to itself accurately. The experiments show that the translational module has a position error less than ±1mm, and the lifting module has a position error less than ±3mm. It is seen that the performance of the material transport system can be guaranteed by the high repeatable accuracy of AGV and its load-transfer mechanism.



(a) grasping the platform

(b) pushing the platform

Figure 9. Load transfer experiments

## VI. Conclusions

This paper designs a long-travel material transport system for a unit load AGV. AGV follows the path of the magnetic tapes in the floor, and locates itself according to the RFID tags beside the load stand. A load transfer methodology containing docking, retrieving and depositing is proposed firstly. A push-pull load-transfer mechanism is then designed to pick up and down the pallet in both sides, including a lifting module that changes the vertical height of the platform and a translational module that moves their convex blocks in the lateral direction. Thirdly, a PLC-based control system is developed to detect the position of executive mechanisms and control their movements. Lastly, the load transfer experiments of our AGV testify the performance of this material transport system.

## References

[1] I.F.A. Vis. "Survey of research in the design and control of automated guided vehicle systems," European Journal of Operational Research, 2006, vol. 170, pp. 677–709.

[2] T. Ganesharajah, N.G. Hall and C. Sriskandarajah. “Design and operational issues in AGV-served manufacturing systems,” Annals of Operation Research, 1998, vol. 76, pp. 109-154.

[3] A. Kelly, B. Nagy, D. Stager and R. Unnikrishnan. “An infrastructure-free automated guided vehicle based on computer vision,” IEEE Robotic and Automation Magazine, 2007, vol. 14, pp. 24-34.

[4] Chen C, Wang B and Ye Q T. “Application of automated guided vehicle (AGV) based on inductive guidance for newsprint rolls transportation system,” J. of Dong Hua Univ. (Engl. Ed.), 2004, Vol. 21, pp. 88-92.

[5] F. Tomoya, O. Jun, A. Tamio and et al. "Semi-guided navigation of AGV through iterative learning," IEEE Int. Conf. Intell. Rob. Syst., 2001, vol.2, pp. 968-973.

[6] G Beccari, S Caselli, F Zanichelli and et al. “Vision-based line tracking and navigation in structured environments,” IEEE Int. Symposium on computational intelligence in robotics and automation, 1997, pp. 406-411.

[7] Wu X, Lou P H and Tang D B. “Integrated motion control of path tracking and servo control for an automated guided vehicle,” Chinese J. of Mech. Eng., 2011, vol. 47, pp. 43-48.

[8] T. Deyle, N. Hai, M. S. Reynolds and et al. “RFID-guided robots for pervasive automation,” IEEE Pervasive Comput., 2010, vol. 9, pp. 37-45.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# A Novel Adaptive Approach for Home Care Ambient Intelligent Environments with an Emotion-Aware System

Sherief Mowafey and Steve Gardner
Glamorgan Intelligent Home Care
Faculty of Advanced Technology, University of Glamorgan
Pontypridd, United Kingdom
smowafey@glam.ac.uk and sgardner@glam.ac.uk

*Abstract*— **The elderly population worldwide has an increasing expectation of well-being and life expectancy. The monitoring of elderly people on an individual basis, in a medical sense, will not be a viable proposition in the future. The infrastructure available is not adequate to meet all expectations and subsequently people will continue to live at home with inadequate care. Prior research has shown an accelerated need for the expansion in the Ambient Intelligence (AmI) domain and to that end we present a novel learning technique for intelligent agents that are embedded in Ambient Intelligent Environments (AIEs). A novel agent that combines an emotion recognition system with a fuzzy logic based learning and adaptation technique provides for an automated self-learning system that constantly adapts to individual requirements. This agent, entitled Health Adaptive Emotion Fuzzy Agent (HAOEFA), has the ability to model and learn the user behaviour in order to control the environment on their behalf with respect to his/her emotional preferences. In addition, the agent incorporates temporal adaption in order to facilitate changing behaviour and preferences within the environment. The results show that such architecture can both provide monitoring and ambient environmental control features such that users with limited physical or cognitive functions can have their well-being advanced with limited external resources.**

***Keywords: Fuzzy Logic Systems, Ambient Assisted Living Systems, Adaptive Intelligent Agents, Ambient Intelligent Home Care Environments, Well-being.***

## I. INTRODUCTION

The world is being overtaken by different demographic trends that will affect the human life over the coming years. One of these trends is labelled as global ageing where in the next few decades it promises to affect everything from business psychology and workforce productivity to the shape of the family and the direction of global capital flows [1]. For instance, according to the Office for National Statistics (ONS) [2] the percentage of the population aged 65 and over within the UK has increased from 15% in 1984 to 16% in 2009, which is an increase of 1.7 million people. More important is in the same period of time, the percentage of the population aged under 16 decreased from 21% to 19%. Future estimation were calculated where it presents that by the year 2034, 23% of the population is projected to be aged 65 and over compared to 18% aged under 16. Also, a chart presented by the HelpAge International [3, 4] showed in Fig. 1 shows different world countries in different regions ageing around the globe as years pass. The vertical axis measures the percentage of each country's population between 65-79 years, the horizontal axis measures shows the countries' population get "older" as they move upwards and across the years [4].

Fast ageing of populations is presenting different challenges for developed and developing countries. These challenges include economic challenges, strains on pension and social security systems, bigger demand on long-term care as well as health care, and bigger demand for trained-health workforce. To address such challenges computer scientists attempted to develop a variety of interactive systems that can communicate with humans as well as helping people through their daily life activities. So, the use of new technologies such as ambient sensors can help doctors to monitor certain activities of elderly people within their houses in order to collect more information about the person. It can help them to understand the alteration of the human well-being status.

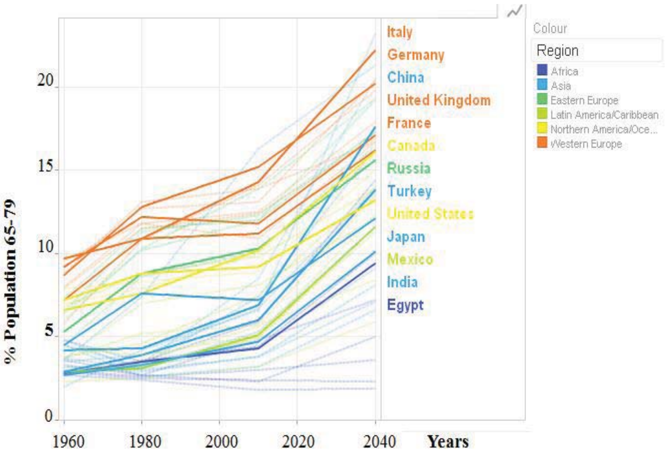


Figure 1. World population over the years [4]

With intelligent home care assistance, care providers can use these aspects and functionalities of sensors (such as kitchen sensor activated, fridge door opened, cold water run for 20 seconds, kettle switched on for 60 seconds) in order to clearly understand the individual's wellbeing current state, as well as answering judgmental questions (such as “is occupant sleeping regularly”, “has the occupant any repeating eating disorder in the last few months”, etc.) [5].

Such techniques of intelligent data monitoring, information clustering and analysis can enable doctors to make accurate diagnosis before it is “late”, and by “late” it is meant the doctor was able to identify accurately the current situation before it reaches a critical state [6]. These efforts of advancing technology to pervade everyday life and to foster wide availability and acceptance was yielded in 1991 when Mark Weiser [7] introduced his vision of ubiquitous computing in his famous article "The Computer for the 21st Century" [7]. Thus, different approaches developed have attempted to help decreasing the global ageing trend where applications such as Tele-care systems developed by Martin et al. [5] tries to produce emergency alerts system that helps in critical situations. Also, research scientists in the Ambient Intelligent Center (AMIC) located in the German University in Cairo [8] have developed an AIE that can learn, and model the user behaviour as well as adapt to the changes of his needs, habits, and gestures in an invisible none interrupting way. In Essex, the Computational Intelligence Centre research team [9] used a soft computing approach based on computational intelligence techniques in order to control the environment on the user behalf and adapt to his/her needs. According to Riva [10] the Intelligent Mixed Reality (IMR) is the most determined illustration to ambience intelligence where it can be invasively integrated and embedded into physical real life environments. It will allow the user to interact with other persons and the environment itself in a usual way. The Research group in University of Bristol identified the three generations of telecare as the first generation systems are typically for “panic alarms situations” such as falling on the ground, the second generation systems are more complicated where they use sensors to detect needed assistance in certain situations and send alarm to family members, friends or to a monitoring agency. Finally, the third generation vision where it tries to predict early warnings for possible emergency situations by observing intrusively significant changes in the person’s “well-being” [5, 11].

Most of the research in the intelligent environments field is focused on controlling and adapting the environment to its user according to their behaviour preferences [12], less focus has been given to the user emotional state within the environment. Therefore, this paper aspect is to show a developed embedded intelligent agent that has the ability to understand the user behaviour in a real-life physical environment pervasively. In addition to, the emotions recognition that take place over time in order to understand the domain of the word well-being and how does it affect the user behavioural actions within the environment. For instance, a person notices that his house has been burglarised. The person assesses the situation, which triggers anger. His heart rate, respiration rate starts to increase rapidly. Consequently, the person contacts the police. Such a situation handling strategy shapes the emotional reactions by adjusting the link between a person and their environment [13].

Various experiments had took place in which the intelligent agent has learned the user behaviour and their emotions during an extended experiment of 14 consecutive days in Glamorgan intelligent home care (Glam i-HomeCare) that acts like a real world AIE test bed. The rest of this paper is structured as follows. In Section II, we will introduce our test bed for AIEs which is the Glam i-HomeCare. Section III presents a description of the life-long learning and adaptation agent entitled HAOEFA. Section IV presents the experiments and results obtained while testing HAOEFA in the environment. Finally, Section V provides the conclusion and future work.

## II. Glamorgan i-HomeCare Overview

The Glam i-HomeCare is one of the real physical AIE test beds shown in Fig. 2, located in the University of Glamorgan. It aims at Home Care AIEs where it looks like any other ordinary room containing furniture. However, it consists of a big number of embedded actuators, processors and sensors that are connected to each other through an assorted network [14]. The Glam i-HomeCare is a single user space that can be used in a wide range of activities such as reading books, watching TV, and eating dinner, etc.

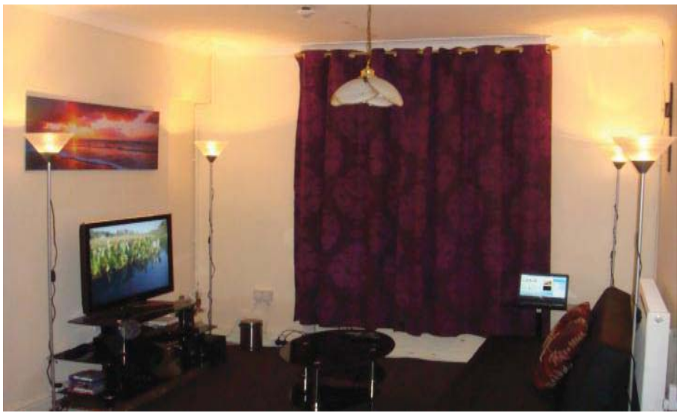


Figure 2. The Glam i-HomeCare Environment

Any networked computer that can run a standard Java process can access the Glam i-HomeCare easily. Thus, this multimedia PC shown in Fig. 2 can also act as an interface controlling the actuators inside room wirelessly. The Glam i-HomeCare Network Infrastructure shown in Fig. 3 is equipped with various embedded actuators and sensors that consists of: date, time of day, internal light level sensor, external light level sensor, internal temperature sensor, external temperature sensor and an occupancy sensor. The developed agent can control four dimmable floor standing up lights and the TV inside the room. These actuators and sensors are embedded within the room infrastructure in order to keep the user completely unaware of the intelligent infrastructure developed, which is required to reach the main aim of AIEs [15].

Although the Glam i-HomeCare environment looks like any other room, the walls and the celling hide numerous networked embedded devices residing on two different networks. Since, it is essential to manage the access of the devices, gateways between different networks critical components in such systems, while combining appropriate granularity with security. Glam i-HomeCare network architecture provides the diverse infrastructure present in ubiquitous computing environments that allow us to improve network independent solutions [16].

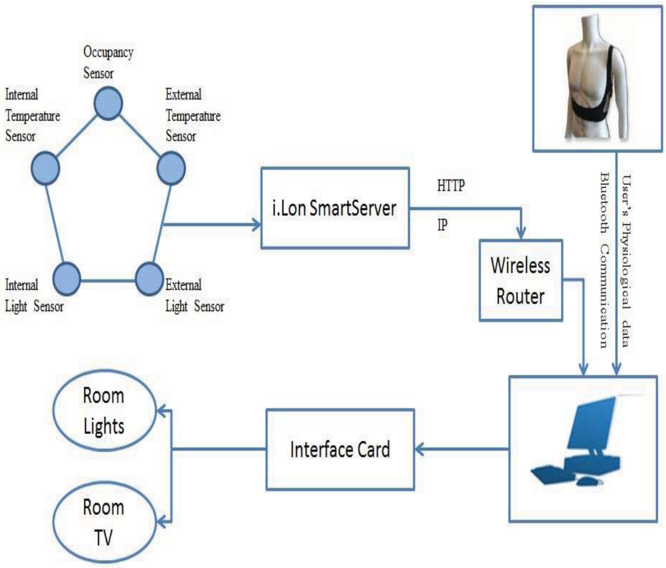


Figure 3. Glam i-HomeCare Network Infrastructure

Fig. 4 shows photos of the various sensors located in the Glam i-HomeCare where its values are displayed on our fuzzy agent interface shown in Fig. 7(b) that operates from the standard multimedia PC in the room presented in Fig. 2.

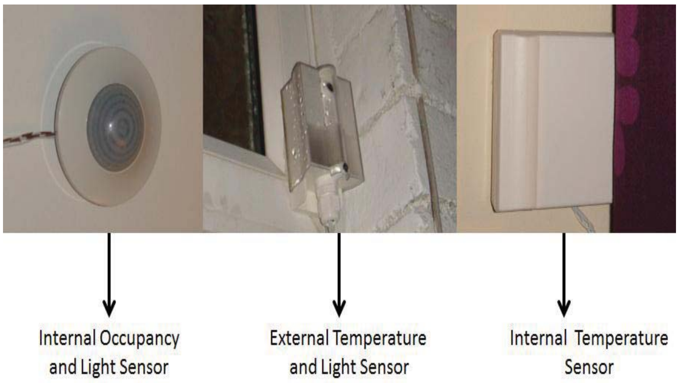

Internal Occupancy and Light Sensor
External Temperature and Light Sensor
Internal Temperature Sensor

Figure 4. Glam i-HomeCare Sensors

## III. The Health Adaptive Online Emotion Fuzzy Agent (HAOEFA)

The proposed agent, entitled as Health Adaptive Online Emotion Fuzzy Agent (HAOEFA), is a Type-1 Fuzzy Logic Controller (FLC). HAOEFA uses an unsupervised methodology to extract fuzzy membership functions and fuzzy rule sets from the data gathered during the learning process, in order to build a fuzzy logic controller that will be capable of learning and modelling the user's behaviour while taking in consideration his/her emotional preferences. These data are gathered by monitoring the user interaction inside the real life environment over a certain period of time. HAOEFA allows its interface to take controls actions in the Glam i-HomeCare environment based on the present situation (state of inputs) and the user emotional state. Additionally, the adaptive agent have the ability to control and manage the room on behalf of the user as well as allowing the rules set to be extended online and adapted over time, achieving a life-long learning technique that adjust to environmental changes and user's behaviour adjusting to it. The HAOEFA consist of the five following phases as shown in Fig. 5:

1. Monitoring the user's behavior and capturing the input/output data that are associated with user's actions with in the environment.
2. Extraction of fuzzy membership functions from the data.
3. Extraction of the user emotional states.
4. Extraction of the fuzzy rules from the captured data.
5. The agent controls the AIE on behalf of the user
6. Life-long learning and adaptation mechanism.

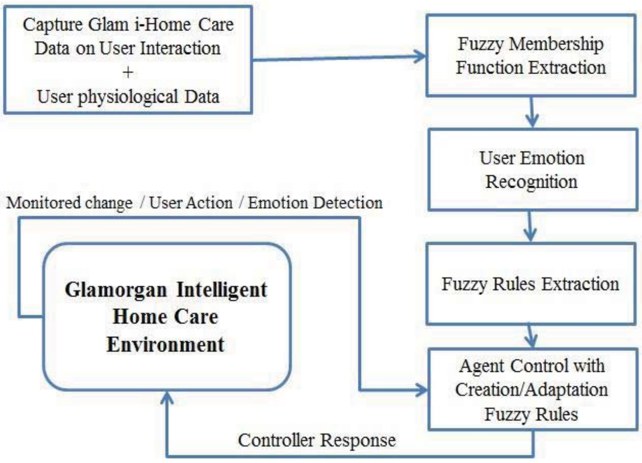
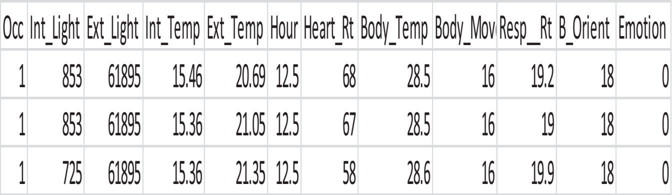

Capture Glam i-Home Care Data on User Interaction + User physiological Data
Fuzzy Membership Function Extraction
User Emotion Recognition
Monitored change / User Action / Emotion Detection
Glamorgan Intelligent Home Care Environment
Fuzzy Rules Extraction
Agent Control with Creation/Adaptation Fuzzy Rules
Controller Response

Figure 5. HAOEFA Flow Diagram.

As presented in Fig. 5, the last two phases of this algorithm are control loops that once started, it receives the environment inputs as changes of sensor values. Subsequently, HAOEFA produces the right control response based on the set of learned rules. Moreover, if the user is not satisfied by the agent control, he/she can take over with a simple change to the system Graphical user interface (GUI) that will lead to a modification in their environment behavioural actions. Therefore, the rules learned by the system will be adapted to the new preference instantly.

### A. *Input/Output Data Capture*

In the learning process the agent displays the different sensor values shown in Fig. 7(b) and monitors the user actions in the room. Every five minutes the agent records a "snapshot" of the current sensor values. For instance, a snapshot of the Glam i-HomeCare environment would look like the following statement:



| Occ | Int_Light | Ext_Light | Int_Temp | Ext_Temp | Hour | Heart_Rt | Body_Temp | Body_Mov | Resp_Rt | B_Orient | Emotion |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 853 | 61895 | 15.46 | 20.69 | 12.5 | 68 | 28.5 | 16 | 19.2 | 18 | 0 |
| 1 | 853 | 61895 | 15.36 | 21.05 | 12.5 | 67 | 28.5 | 16 | 19 | 18 | 0 |
| 1 | 725 | 61895 | 15.36 | 21.35 | 12.5 | 58 | 28.6 | 16 | 19.9 | 18 | 0 |

where Occupancy will be equal to 1 if the user is in the room, else it's going to be 0, the internal and external light levels are measured in LUX, internal and external temperatures are measured in degree Celsius (°C), the hour is the time of the day. In addition, the agent captures the measurements of the heart rate, skin temperature, respiration rate, and body posture using a wireless belt that that user wears around his chest. In the learning process the system won't be able to detect the user emotional state, but once the adaptation process takes place, HAOEFA will be able to interpret the emotions of the user such as happy, angry and Neutral.

It also observes the outputs (actuator states with new values of which any actuators were modified by the user). For instance, the output or the action of the user that took place for the input or the situation discussed previously in the snapshot taken is:

| ACTION_Light_value | Action_TV_value |
|---|---|
| 0 | 0 |
| 0 | 0 |
| 0 | 0 |

where the ACTION_Light_value are the control light values of the lights in the room where they range from 0 to 100 (0 equals off, 50 equals to halfway, and 100 is on). The TV value is either 0 or 1 (0 equals off, and 1 is on). All these inputs are recorded by the agent in order to model the user behavior, detect his/her emotional state while living in the Glam i-HomeCare.

### *B. Extraction of Fuzzy Rule Sets*

The proposed agent combines the set of membership functions with the input/output data so that it can model the behavior of the user in order to extract the fuzzy rules. This fuzzy rule extraction methodology used by the HAOEFA extracts multi-input multi-output rule that describe the relation between y = $(y_1, ..., y_k)$ and x = $(x_1, \dots, x_n)^T$ as in (1):

$$if\ x_1\ is\ A_1^{(l)}\ and\ ...\ and\ x_n\ is\ A_n^{(l)} then\ y_1 \text{ is } \mathrm{B}_1^{(l)}\text{and ... and } \mathrm{y_n} \text{ is } \mathrm{B}_n^{(l)} \tag{1}$$

where l = 1, 2, ... , M, and M is the number of rules and l is the index of rules. There are V fuzzy sets $A_s^q$, q = 1, ... , V, defined for each input $x_s$. There are W fuzzy sets $B_c^h$, h = 1, ... , W defined for each output $y_c$ [9].

### *C. The Emotion Recognition System*

Emotion recognition computing systems should operate in a similar way to how humans recognize emotions during communication. Thus, detecting people emotions during conversations could lead to an intelligent human to human interaction. For example, a person notices that a friend is feeling sad so he turns off the music being played instantly. Inspired by this idea the intelligence of HAOEFA is not only limited to the environment conditions, weather, time, etc. but it also includes the recognition of the user emotional state as it tries to achieve a certain level of comfortably within the environment by controlling it on the user's behalf. In addition, it monitors the user's health indicators as an attempt to better understand the user's well-being over time.

During Glam i-HomeCare experiments, the user wears the equivital belt which uses Bluetooth communication to measure the physiological conditions of the human body. Following the learning period and the extraction of the fuzzy membership functions for the FLC, HAOEFA is able to identify the user emotions by monitoring the user physiological data where it detects any changes that occurs in a real-time mode. Consequently, the agent starts controlling the environment on the user behalf. For instance, play some music while the user is happy. Currently, only "angry", "happy" and "neutral" emotions are being considered since more differentiated emotion recognition would lead to lower accuracies [14, 17].

### *D. The Agent Controller*

The proposed agent uses singleton fuzzification, max-product composition, product implication and height defuzzification, as it maps a crisp input n into a crisp output y = f(x) where it applies as in (2) as follows:

$$\mathrm{y(x)} = \mathrm{f_s(x)} = \frac{\sum_{l=1}^{M} y^{-l} \prod_{i=1}^{n} \mu_{F_i^l}(x_i)}{\sum_{l=1}^{M} \prod_{i=1}^{n} \mu_{F_i^l}(x_i)} \tag{2}$$

where M is the total number of rules in the rule base, y-l is the maximum membership value point in the ith rule output fuzzy Bl, $\prod_{i=1}^{n} \mu\, F_i^l\, (x_i)$ set is the product of the membership values of each rule's input fuzzy sets, and n is the number of inputs. As to get multiple outputs, equation (2) is repeated for each output parameter. Consequently, after extracting the membership function and building the set of rules from the user's environment multi-input/output data. The agent has learned and modeled the user behavior. Consequently, HAOEFA starts controlling the environment on the user's behalf with respect to their needs and emotional state by monitoring the Glam i-HomeCare environment input values (such as sensors and emotional state) and as a result it controls the actuators based on the rule model built before from the user interactions that have been made during the learning process in the intelligent environment.

### *E. Life Long Learning and Online adaptation*

In the above subsection, we have explained how the agent can learn the user behaviors. Nevertheless, HAOEFA is a flexible system, where the initially learned rules can be easily extended to change both existing rules as well as adding new rules as user's behavior might alter or change over time. The agent allows capturing wide range of various values for each input and output constrains. This methodology offers a continuous operation even if a gradual change in the environment exists (such as temperature drop off in winter). If, however, a significant alteration in the environment or the user's behavior acquires, that has not been introduced to the system during the learning phase. Consequently, the agent will intelligently create new rules that satisfy the current state in an unobtrusive way in order to extend its behavior to satisfy the user [14].

In other words, that the rule fired, and would therefore have contributed to the overall control response generated by the agent's FLC. The consequent membership functions that give the highest membership values to the user defined actuator values are selected to replace the consequent sets of all fired rules in the rule base [15]:

$$\mu_{B_c^{h*}}(y_c) \geq \mu_{B_c^h}(y_c) \tag{3}$$

where $h = 1, 2, \dots, W$, and W is the number of fuzzy sets, $B_c$ is chosen as $B_c^{h*}$, where $c = 1, 2, \dots, k$. The fired rules are therefore adapted to better reflect the user's updated actuator preferences given the current state of the environment and this leads to a grid of identified fuzzy set(s) for each input parameter. From this grid, new rules are constructed based on each unique combination of consecutive input fuzzy sets [15]. The consequent fuzzy sets for each of the new rules are determined as in (3). This allows new rules to be gradually added to the rule base. The agent will also add new rules when the currently monitored environmental state is undefined by the existing rules in the rule base; i.e., none of the existing rules fired. In this case, the agent will create new rules where the antecedent sets reflect the current input states of the environment and the consequent fuzzy sets are based on the current state of the actuators. The agent adopts life-long learning, where it adapts its rules as the state of the environment and the preferences of the user change over a significantly long period of time [14, 15].

## IV. Experiments and Results

This section demonstrates the results obtained by the HAOEFA in the Glam i-HomeCare which is a real AmI test bed. In the beginning, we are going to introduce the scenario performed during experiments. Consequently, the HAOEFA results obtained in the Glam i-HomeCare will be presented, where the agent has learned and modeled the user behavior through a certain period of time. Afterwards, the agent controls the environment on the user behalf with respect to the user emotion preferences. The results were examined where the rules are compared with the readable format rules with user's journal parameters to guarantee that the agent has successfully learned the behavior the user was anticipating.

### *A. Glam i-HomeCare Scenario*

The experiments scenario takes the following route: whenever a user enters the Glam i-HomeCare, there are several steps to be maintained as shown in Fig. 6. Firstly, the user enters the username and password in order to authenticate with an online server as shown in Fig. 7(a) as well as for the system to identify the user in action. Subsequently, on one hand, if the user is already registered with the Glam i-HomeCare, HAOEFA will start immediately to control the environment on the his/her behalf accordingly to the behavior learned and the emotion preference. It can be easily adapted online to whatever changes could happen wither in the environment or in the user behavior. On the second hand, if he/she is a new user to Glam i-HomeCare, the HAOEFA will start monitoring the user for x days as it enters the learning mode. The experiments presented in this section take a period of 14 days while the first day is the learning day then the agent start computing in order to build the rule base. Subsequently, HAOEFA will enter the live adaption online control phase for 13 more days where the agent controls the environment and whenever the user is unsatisfied with agent control he can easily adapt it using the Glam i-HomeCare GUI shown in Fig 7(b).

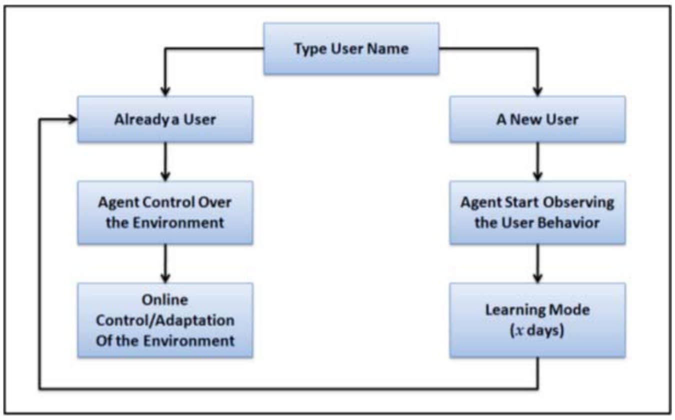
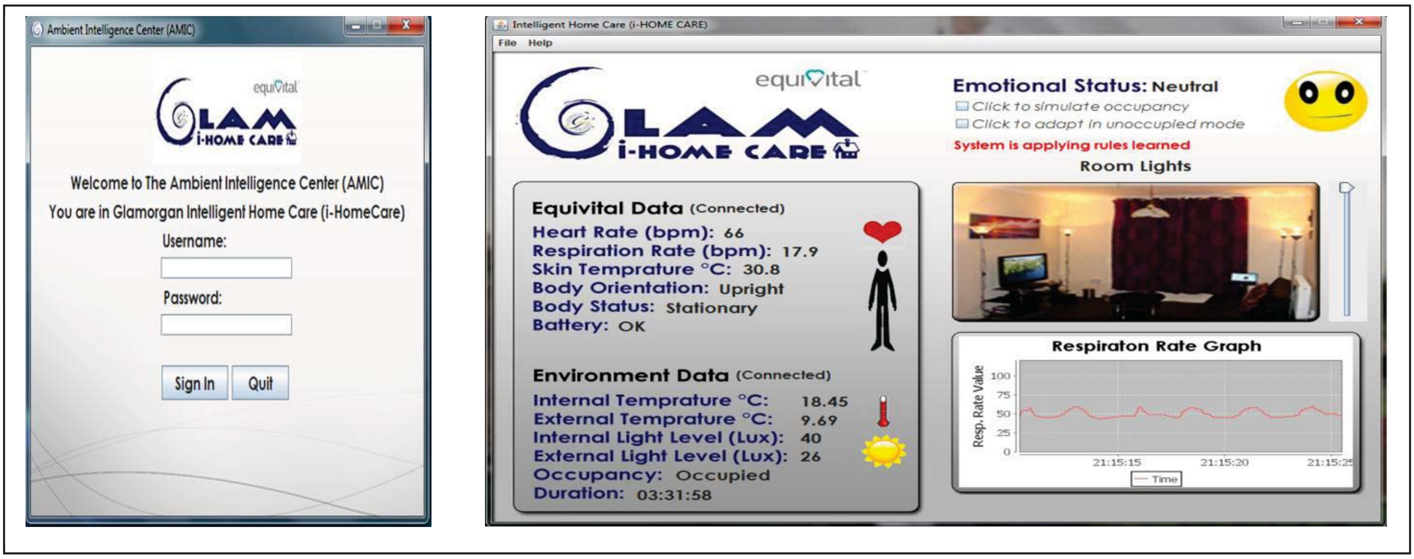

Type User Name
Already a User
A New User
Agent Control Over the Environment
Agent Start Observing the User Behavior
Online Control/Adaptation Of the Environment
Learning Mode ($x$ days)

Figure 6. The Glam i-HomeCare Scenario

In addition, the user uses HAOEFA's GUI installed on the multimedia PC to monitor the real-time physiological body changes and the emotions recognized with it, as well as the display of the environment's sensors data while the agent the environment control taken as shown in Fig. 7(b). In order to test the agent capabilities a history record of the user decisions was saved in a journal and using a parsing tool that convert the .csv file containing the fuzzy rules sets into a readable format.

The complete dataset of 2880 instances acquired from the user's interactions in the Glam i–Home Care over the initial period of one day was used by HAOEFA to learn an initial FLC set of rules. The membership functions were built using fuzzy c-means and the results are shown in Fig. 8 (a) and (b). The agent then runs online for a further 13 days, during which it monitors the user's activities and controls the environment on their behalf. During this time, the user is allowed to override and adapt the agent's learned control responses, if it was necessary to modify and tune them further. As mentioned previously, one of the main characteristics of the Glam i-HomeCare System is that the user is always in control of the environment and capable of overriding the agent's control at any time and his instructions are executed immediately, in order to achieve the responsive property implied in the ambient intelligence vision.


Ambient Intelligence Center (AMIC)

Welcome to The Ambient Intelligence Center (AMIC)

You are in Glamorgan Intelligent Home Care (i-HomeCare)

Username:

Password:

Intelligent Home Care (i-HOME CARE)

Emotional Status: Neutral

□ Click to simulate occupancy

□ Click to adapt in unoccupied mode

System is applying rules learned

Equivital Data (Connected)

Heart Rate (bpm): 66

Respiration Rate (bpm): 17.9

Skin Temprature °C: 30.8

Body Orientation: Upright

Body Status: Stationary

Battery: OK

Environment Data (Connected)

Internal Temprature °C: 18.45

External Temprature °C: 9.69

Internal Light Level (Lux): 40

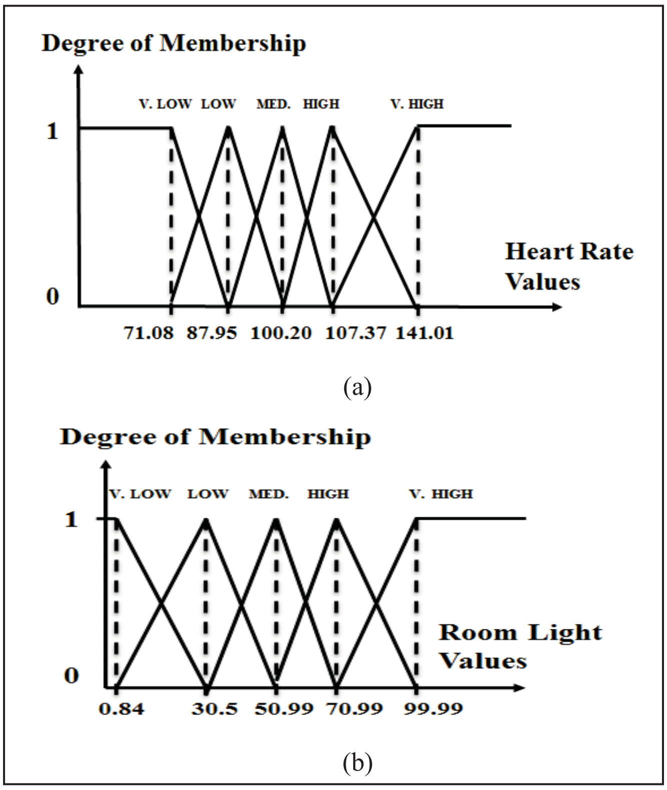
External Light Level (Lux): 26

Occupancy: Occupied

Duration: 03:31:58

Figure 7. (a) User Login Screen (b) Glam i-Home Care Graphical User Interface

Thus, whenever changes to controls were made by the user, the agent received the request, generated new rules or adjusted previously learned rules, and allowed the action through. The performance of the agent could be gauged on the number of occasions when the user had to override the agent's control responses and adapt the rules over time, as this can reflect how satisfied the user is with the agent control actions.



(a)

(b)

Figure 8. (a) User Heart Rate Membership Function (b) User Heart Rate Membership Function

The success of HAOEFA could be measured by the ability of understanding human dialogue words such as very low, low, medium, high, etc. as presented in the built membership of different attributes like heart rate and the control indoor room light shown in Fig. 8(a) and Fig. 8(b) respectively. In addition to, the monitoring of how efficient the agent adjusted the environment to the user's preferences where the user intervention was reduced over time. Fig. 9 shows a graph plotting the number of online rule adaptations against time measured in hours. The agent initially learned 16764 rules in the first 24 hours from the user's dataset and over the course of the subsequent rest of the days. It also demonstrates the number of the user's induced rule adaptations that occurred over the duration of the experiment, where 341 rules were added in the second day, and then 11 rules were added in the third day. Later, a total of 729 new rules were added by end of the seventh day. At that time, no more adaptations occurred which illustrate that HAOEFA was able to fully stabilize by the end of the first week. Moreover, given the theoretical maximum number of rules which is 15625000 rules, the agent generated a total of 17493 rules. This shows that HAOEFA was able model the user's behaviour within Glam i-HomeCare using 0.001% of the actual theoretical number using a nonintrusive technique to learn most of the user's preferences for various weather and environmental conditions over the duration of the first day of the experiment, including behavioural changes associated with user emotion changes.

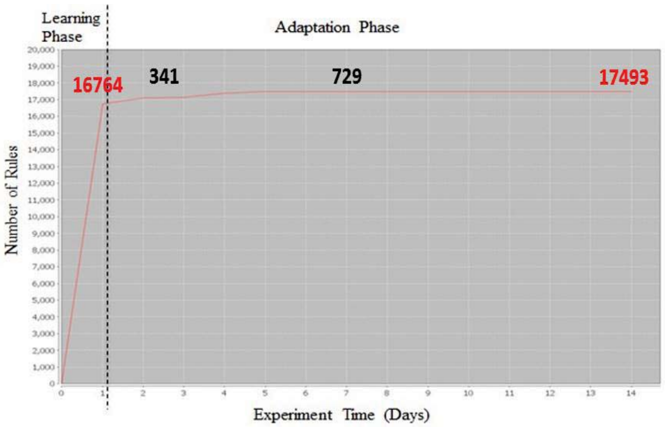


Figure 9. HAOEFA rules learned plotted against the experiment time.

## V. CONCULISION AND FUTURE WORK

In Conclusion, the Glam i-HomeCare agent succeeded in interacting within a real life physical environment. HAOEFA has effectively produced the appropriate control responses based on its learned rules and the emotion preference of the user detected at a time. In the Future, more experiments will take place in order to enhance the emotional model and to add more emotions such as sadness, and fear to the agent. Over and above that, a more efficient Type-2 Fuzzy Logic agent will be developed in order to enhance the FLC to model the user behavior with respect to his emotional preference as an attempt to define the word "Well-Being" in order to help the elder people to live a better life in their homes with care they deserve.

## REFERENCES

[1] W. H. Organization. (2011, 25-02-2011). *Ageing and life course* Available: http://www.who.int/ageing/en/

[2] O. N. Statistics. (2005, 25-02-2011). *Older People*. Available: http://www.statistics.gov.uk/cci/nugget.asp?id=1263

[3] H. International. (2011, 25-02-2011). *Ageing in motion*. Available: http://www.helpage.org/resources/ageing-data/ageing-in-motion/

[4] S. Mowafey, "Investigation Into The Creation Of An Ambient, Physiology Measurement Environment To Facilitate The Modelling Of "Well-Being"." PhD, Faculty of Advanced Technology, University of Glamorgan, University of Glamorgan, 2011.

[5] T. Martin, B. Majeed, B. Lee, and N. Clarke, "A Third-Generation Telecare System using Fuzzy Ambient Intelligence," *Computational Intelligence for Agent-based Systems,* pp. 155-175, 2007.

[6] E. Braithwaite, *Neural networks for medical condition prediction: an investigation of neonatal respiratory disorder*: Citeseer, 1998.

[7] M. Weiser, "The computer for the 21st century," *Scientific American,* vol. 265, pp. 94-104, 1991.

[8] S. Mowafey, A. Schmitt, H. Hagras, and W. Minker, "Creating an Ambient Intelligent Environment with an Emotion-Aware System," in *The 5th International Conference on Intelligent Environments*, Barcelona, Spain, 2009.

[9] F. Doctor, H. Hagras, and V. Callaghan, "A fuzzy embedded agent-based approach for realizing ambient intelligence in intelligent inhabited environments," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on,* vol. 35, pp. 55-65, 2004.

[10] G. Riva, "Ambient intelligence in health care," *CyberPsychology & Behavior,* vol. 6, pp. 295-300, 2003.

[11] R. Lee and V. Loia, *Computational intelligence for agent-based systems*: Springer Verlag, 2007.

[12] D. Cook, H. Hagras, V. Callaghan, and A. Helal, "Making our environments intelligent," *Journal of Pervasive and Mobile Computing,* vol. 5, pp. 556-557, 2009.

[13] R. S. Lazarus, *Emotion and adaptationn: Conceptual and empirical relations*: Oxford University Press, USA, 1991.

[14] S. Mowafey, "A Novel Approach for Developing an Ambient Intelligent Environment with an Emotion-Aware System based on Computational Intelligence," The Faculty of Postgraduate Studies and Scientific Research, The German University in Cairo, Cairo, 2009.

[15] F. Doctor, H. Hagras, V. Callaghan, and A. Lopez, "An adaptive fuzzy learning mechanism for intelligent agents in ubiquitous computing environments," 2005.

[16] H. Hagras, V. Callaghan, M. Colley, G. Clarke, A. Pounds-Cornish, and H. Duman, "Creating an ambient-intelligence environment using embedded agents," *Intelligent Systems, IEEE,* vol. 19, pp. 12-20, 2004.

[17] H. Meng, J. Pittermann, A. Pittermann, and W. Minker, "Combined speech-emotion recognition for spoken human-computer interfaces," 2008, pp. 1179-1182.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# A Convergence Analysis of D-ILC Algorithm

Inam Ul Hasan Shaikh
(*Student Member IEEE*)
Control Systems Centre
The University of Manchester
Manchester, UK
Inamulhasan.Shaikh@postgrad.manchester.ac.uk

Martin Brown
Control Systems Centre
The University of Manchester
Manchester, UK
Martin.Brown@manchester.ac.uk

***Abstract*—ILC is an emerging technique for learning control. The D-ILC algorithm is the generic ILC scheme which captures the error trend of a batch to update the control input for the next batch or batches. The 2-dimensional nature requires in-depth convergence analysis of the algorithm. This paper addresses these issues in detail. This paper deals with the convergence properties of ILC algorithms with emphasis on control input. Discrete-time linear state space representation of a linear time-invariant system has been considered along with usual assumptions which ensure D-type ILC algorithm converges in terms of output error. The convergence for control input sequence is investigated up to component level.**

***Keywords: Iterative Learning Control; Rate of Convergence;***

## I. Introduction

Iterative Learning Control (ILC) is one of the Intelligent Control schemes. The transient performance is improved for systems which operate in a repetitive manner. The repetitions occur after fixed intervals of time. ILC has achieved better performance of control systems especially when dealing with uncertain/stochastic systems[1]. The concept of learning through repeated trials evolved in late 1970s for improving the motion control of mechanical arms [2, 3]. The D-type algorithms for Iterative Learning Control for linear time-varying systems with application to robotic manipulators were developed [4-6]. However, the D-type ILC suffered with the problem of differentiation of high frequency noise. Later, the P-type ILC improved upon by using only the error instead of its derivative [7].

Current Iteration Tracking Error (CITE) was introduced which formulated ILC in line with feedback control paradigm and helped to overcome large overshoot thus convergence could be accelerated [8]. The discrete-time version of D-type ILC was formulated for MIMO linear systems which possessed global robustness against state disturbances, measurement noise and re-initialisation error at the beginning of each iteration[9]. The ILC law has been employed for non-linear time varying systems having affine input and linear output. Uniform convergence of input and state was achieved when there were no disturbances [10].

In real world applications, the plants are usually non-linear with higher order dynamics and can not be modelled accurately. The mathematical model is linearised around a known equilibrium point for LTI systems. It then becomes easier to design the controllers. However, the model uncertainties restrict the controller from achieving optimal solutions.

Iterative Learning Controllers provide an adaptive solution. These controllers utilize the error information of each batch/iteration and update the control input accordingly for the next batch or batches. The control input signal using ILC converges to the desired value of the control input which is the inverse solution. Hence, it is also termed as Iteration Inversion Process.

The error information at each time instance can be used in a variety of ways to generate the update for control input. Hence, there are D, PD, PI, PID, Gradient based etc. many types of ILC controllers in which the output error, its derivative, integral or some combination of these is added to the current control input to generate input for the same time instant in next batch [7, 9-19]. The rate of convergence up to component level inside batches has not been covered in earlier works. This has been carried out in this paper.

## II. Discrete-time System with ILC

The discrete-time LTI system having one relative degree is considered as follows:

$$\begin{aligned} x(i+1,k) &= A\,x(i,k) + B\,u(i,k) \\ y(i,k) &= C\,x(i,k) \end{aligned} \tag{1}$$

where $k$ denotes the Batch/iteration/trial number having $M$ number of samples in each trial, $i \in [1, M]$ is the time index or sample number during each batch, state vector $x \in \mathbb{R}^n$, input $u(i,k) \in \mathbb{R}^r$ and output of the system is $y(i,k) \in \mathbb{R}^p$. $A, B$ and $C$ are the real-valued state, input and output matrices respectively, having appropriate dimensions. The initial conditions have been selected as follows:

- $x(1,k) = x_0$ is same at the start of each batch. Hence zero initial error is maintained.
- $u(i,1) = u_0(i)$ is the control input vector for first batch which may be externally specified or left to be zero [9, 10, 19, 20].

Here the control input for the first batch is assumed zero.

### *A. D-ILC Control Input Update*

Control input is updated using D-ILC as follows:

$$u(i,k+1) = u(i,k) + K_d\{e(i+1,k) - e(i,k)\} \tag{2}$$

where $K_d$ is the real-valued learning gain matrix and the derivative of error $\dot{e}(i,k)$ has been approximated using forward difference:

$$\dot{e}(i,k) = e(i+1,k) - e(i,k) \tag{3}$$

where $e(i,k)$ is the error between the desired/reference and the actual outputs:

$$e(i,k) = y^*(i) - y(i,k) \quad \text{for } 1 \le i \le M \tag{4}$$

Furthermore, using the zero initial error assumption we have $e(1,k) = 0$ i.e., $y^*(1) = y(1,k)$.

Thus, the objective of the D-type ILC algorithm is to find the sequence $u(i,k)$ so that:

$$\lim_{k\to\infty} u(i,k) = u^*(i) \quad \text{for } \forall i = 1,2,\ldots,M \tag{5}$$

For a linear system, the convergence of input sequence corresponds to the convergence of the output sequence:

$$\lim_{k\to\infty} y(i,k) = y^*(i) \quad \text{for } \forall i = 2,3,\ldots,M+1 \tag{6}$$

### *B. Convergence of Control Input Sequence using D-ILC*

In this section, the component-level relationship for $u(i,k+1)$ at $(k+1)^{\text{th}}$ batch is presented in terms of static and dynamic components as well as the control input $u(i,k)$ from previous batch *k*. Convergence of control input $u(i,k)$ approaching the desired input $u^*(i)$ is investigated. So that the desired output sequence $y^*(i)$ is generated when sequence $u^*(i)$ is applied to the system:

$$y^*(i) = G(z)u^*(i) = \left\{C\,(zI - A)^{-1}\,B + D\right\}u^*(i) \tag{7}$$

The convergence condition for control input and rate of convergence for individual components of the control input have been derived, where bounds for the convergence rates have been formulated as well.

#### *1) Batch to Batch Control Input Sequence*

The D-type ILC algorithm generates the batch to batch control input sequence, i.e. from $1,2,\ldots,k,k+1,\ldots$ for each time index $i$. In this section a recurrence relationship is derived to perform a convergence analysis on $u(i,k)$. The control input sequences for individual time indices are derived using (2) as follows:

$$u(1,k+1) = u(1,k) + K_d\{e(2,k) - e(1,k)\} \tag{8}$$

Using the zero initial error, i.e., $e(1,k) = 0$, we get:

$$\begin{aligned} u(1,k+1) &= u(1,k) + K_d\{e(2,k)\} \\ &= u(1,k) + K_d\left\{y^*(2)\right\} - K_d C \begin{Bmatrix} Ax(1,k) \\ +Bu(1,k) \end{Bmatrix} \\ &= \left(I - K_d CB\right)u(1,k) + K_d\left\{y^*(2)\right\} - K_d CAx_0 \end{aligned} \tag{9}$$

It can be further expressed as follows:

$$\begin{aligned} u(1,k+1) &= \left(I - K_d CB\right)u(2,k) + K_d\left[y^*(3) - y^*(2)\right] \\ &\quad + K_d C\,(I - A)\,x(2,k) \end{aligned} \tag{10}$$

The sequence in (10) can be generalized as follows:

$$u(i,k+1) = \begin{pmatrix} \left(I - K_d CB\right)u(i,k) + K_d \begin{bmatrix} y^*(i+1) \\ -y^*(i) \end{bmatrix} \\ + K_d C\,(I - A)\,x(i,k) \end{pmatrix} \tag{11}$$

Solution for (11) is expressed as a recurrence relation:

$$\begin{aligned} u(i,k) &= \left(I - K_d CB\right)^{k\text{-}1} u(i,1) + \\ &\sum_{j=1}^{k\text{-}1}\left(I - K_d CB\right)^{k\text{-}j\text{-}1} \begin{bmatrix} K_d\left\{y^*(i+1) - y^*(i)\right\} \\ +K_d C\,(I - A)\,x(i,k) \end{bmatrix} \end{aligned} \tag{12}$$

For zero control input at first batch, (i.e. $u(i,1) = 0$), (12) can be reduced as follows:

$$u(i,k) = \sum_{j=1}^{k\text{-}1}\left(I - K_d CB\right)^{k\text{-}j\text{-}1} \begin{bmatrix} K_d\left\{y^*(i+1) - y^*(i)\right\} \\ +K_d C(I - A)x(i,k) \end{bmatrix} \tag{13}$$

As batch number $k$ increases, the control input components achieve convergence one after the other with increasing time index $i$ inside a batch. From re-writing the sequences in (11) to show the dependency of $u(i,k+1)$ on $u(i,k)$ from previous batch including all the other previous control input components $u(i-1,k)$, $u(i-2,k),\ldots,u(1,k)$ occurring in the same batch as follows:

$$\begin{aligned} u(i,k+1) &= \left(I - K_d CB\right)u(i,k) + K_d C \begin{pmatrix} (I - A) \\ Ax(i-1,k) \end{pmatrix} \\ &\quad + K_d C \begin{pmatrix} (I - A) \\ Bu(i-1,k) \end{pmatrix} + K_d\left[y^*(i+1) - y^*(i)\right] \end{aligned} \tag{14}$$

On further simplification, we have:

$$
\begin{aligned}
u(i,k+1) = \left(I - K_d CB\right) u(i,k) + \ K_d C\,(I - A)\,Bu(i-1,k) \\
+ \cdots + K_d C\,(I - A)\,A^{i\text{-}2} Bu(1,k) \\
+ K_d C\,(I - A)\,A^{i\text{-}1} x(1,k) + K_d\left[y^*(i+1) - y^*(i)\right]
\end{aligned}
\tag{15}
$$

It can be observed that convergence $u(1,k) \rightarrow u^*(1)$ occurs first. After that the output convergence $y(2,k) \rightarrow y^*(2)$ is achieved. Then after one or few batches, $u(2,k) \rightarrow u^*(2)$ convergence is achieved followed by $y(3,k) \rightarrow y^*(3)$. Convergence of other inputs $u(i,k)\text{'s}$ and corresponding outputs $y(i+1,k)\text{'s}$ occurs in further batches. The batch to batch sequential convergence of samples is shown in Figure 1.

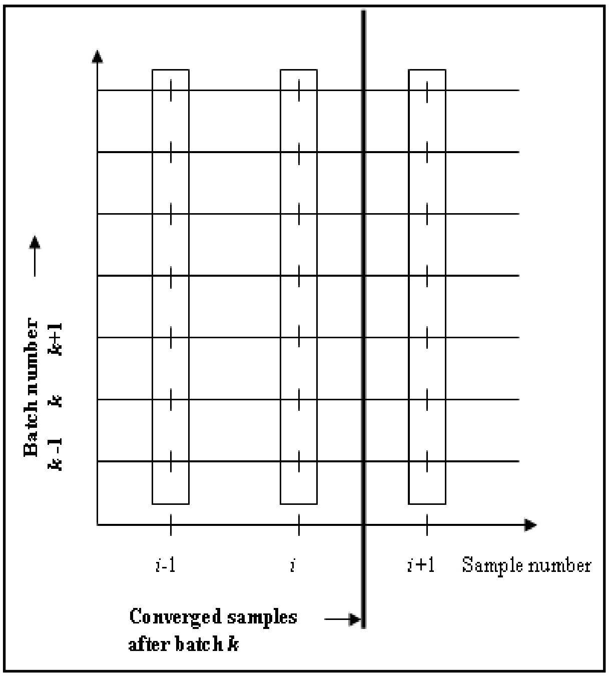


**Figure 1 Sequential Convergence in ILC**

*2) Convergence Condition for Control Input Sequence*

To obtain a stable and bounded sequence of control input $u(i,k)$ in (13), the Eigenvalues of $\left(I\text{ - }K_d CB\right)$ must lie inside the unit circle. Hence, the maximum absolute eigenvalue should be less than unity as follows:

$$
\max\left|\lambda\left(I\text{ - }K_d CB\right)\right| < 1 \tag{16}
$$

For monotonic convergence, the condition in (16) is expressed as norm [21]:

$$
\left\|I\text{ - }K_d CB\right\| < 1 \tag{17}
$$

The (17) gives the necessary and sufficient condition for convergence of D-ILC algorithm. It is pointed that since this condition does not depend on the system matrix $A$, it marks the ability of ILC algorithm to achieve convergence even when the model parameters are unknown.

*3) Rate of Convergence of Control Input Errors*

Using D-ILC algorithm, the component-wise control input errors between desired control input $u^*(i)$ and $u(i,k+1)$ are calculated as follows:

$$
\begin{aligned}
u^*(i) - u(i,k+1) &= u^*(i) - \begin{bmatrix} u(i,k) \\ +K_d\{e(i+1,k) \\ -e(i,k)\} \end{bmatrix} \\
&= u^*(i) - u(i,k) \\
&\quad - K_d C\left[\begin{Bmatrix} \left(Ax^*(i) + Bu^*(i)\right) \\ -\left(Ax(i,k) + Bu(i,k)\right) \end{Bmatrix}\right] \\
&\quad + K_d\left\{y^*(i) - y(i,k)\right\}
\end{aligned}
\tag{18}
$$

Let's denote $u^*(i) - u(i,k) = \Delta u(i,k)$ and use in (18) and after simplification we get:

$$
\begin{aligned}
\Delta u(i,k+1) &= \left(I - K_d CB\right)\left\{\Delta u(i,k)\right\} \\
&\quad + K_d C\,(I - A)\left\{x^*(i) - x(i,k)\right\}
\end{aligned}
\tag{19}
$$

Since initial conditions are preserved & initial error is zero, i.e. $x^*(1) = x(1,k)$, hence at 1st time index the error is:

$$
\Delta u(1,k+1) = \left(I - K_d CB\right)\left\{\Delta u(1,k)\right\} \tag{20}
$$

At $i^{th}$ time index, the error becomes:

$$
\begin{aligned}
\Delta u(i,k+1) &= \left(I\text{ - }K_d CB\right)\Delta u(i,k) \\
&\quad + K_d C\,(I\text{ - }A)\,B\Delta u(i\text{ - }1,k) \\
&\quad + \cdots + K_d CA^{i\text{-}1}\,(I\text{ - }A)\,B\Delta u(2,k) \\
&\quad + K_d CA^{i\text{-}2}\,(I\text{ - }A)\,B\Delta u(1,k)
\end{aligned}
\tag{21}
$$

By writing (20) and (21) in matrix form, the evolution of vectors of control input errors from batch to batch is given as follows:

$$
\Delta u(k+1) = T\ \Delta u(k) \tag{22}
$$

where, the operator matrix $T$ controls the evolution of control input errors from batch to batch as follows:

$$\begin{bmatrix} (I - K_d CB) & 0 & 0 & \cdots & 0 \\ K_d C(I - A)B & (I - K_d CB) & 0 & \cdots & 0 \\ K_d CA(I - A)B & K_d C(I - A)B & \ddots & 0 & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ K_d CA^{M-2}(I - A)B & \cdots & \cdots & \cdots & (I - K_d CB) \end{bmatrix} \quad (23)$$

The solution of (22) is given as follows:

$$\Delta u(k) = T^{k-1} \Delta u(1) \quad (24)$$

*Lemma:*

The D-ILC algorithm produces a sequence of bounded control inputs which over the long term converge component-wise to the desired control input sequence at the rate equal to the magnitude of the Eigenvalue of the matrix relating the evolution of control input error provided initial conditions are same and desired output matches with the measured output at the beginning of each batch.

*Proof of Lemma:*

Consider the solution for evolution of control input errors in (24). Due to repeated eigenvalues, matrix $T$ cannot be diagonalised using eigen-decomposition or SVD because there are repeated eigen vectors which make the matrix of Eigen vectors singular. Therefore, the following assumptions have been considered:

1. The rate of convergence of each component $\Delta u(i,k)$ can be found if $(M \times M)$ matrix $T$ is decomposed into Jordan Normal form. For a non-singular matrix $Q$ such that $Q^{-1} \times Q = I$, Jordan Decomposition is $T = Q^{-1} D Q$.

2. The matrix $D$ has the eigenvalues of matrix $T$ as its diagonal elements along with a sub-diagonal containing all 1's.

3. To calculate the convergence rates of individual components of control input error vector $\Delta u(k)$ in (24) we have to decompose as follows [22]:

$$T^{k-1} = Q^{-1} D^{k-1} Q \quad (25)$$

4. The control input error in (22) can be written as follows:

$$\Delta u(k) = T^{k-1} \Delta u(1) = Q^{-1} D^{k-1} Q \Delta u(1) \quad (26)$$

5. The rate of convergence $\Gamma_{\Delta u(i,k)}$ is the ratio of $\infty$-norm at batch $k$ with respect to the $\infty$-norm at batch $k-1$:

$$\Gamma_{\Delta u(i,k)} = \frac{\left\|\Delta u(i,k)\right\|_\infty}{\left\|\Delta u(i,k-1)\right\|_\infty}$$

$$= \frac{\left[\begin{array}{l} \left|\lambda^{k-1}\right| + \left|(k-1)\lambda^{k-2}\right| \\ + \left|\dfrac{(k-1)(k-2)}{2!}\lambda^{k-3}\right| \\ + \cdots + \left|\dfrac{(k-1)\ldots(k-i+1)}{(i-1)!}\lambda^{k-i}\right| \end{array}\right]}{\left[\begin{array}{l} \left|\lambda^{k-2}\right| + \left|(k-2)\lambda^{k-3}\right| \\ + \left|\dfrac{(k-2)(k-3)}{2!}\lambda^{k-4}\right| \\ + \cdots + \left|\dfrac{(k-2)\ldots(k-1-i+1)}{(i-1)!}\lambda^{k-1-i}\right| \end{array}\right]} \quad (27)$$

Consequently, after simplification, (27) can be written as follows:

$$\Gamma_{\Delta u(i,k)} = |\lambda| \sum_{j=0}^{i-1} \left\{ \left| \binom{k-1}{j} \right| \left| \left(\frac{1}{\lambda}\right)^j \right| \right\} \bigg/ \left\{ \left| \binom{k-2}{j} \right| \left| \left(\frac{1}{\lambda}\right)^j \right| \right\} \quad (28)$$

where

$$\binom{k-1}{j} = \frac{(k-1)!}{(k-1-j))!\,j!} \quad (29)$$

In the limit $k \to \infty$, the ratio to the right of summation in (28) is unity, so that every component $\Delta u(i,k)$ has long term convergence rate $\approx |\lambda| = \left|(I - K_d CB)\right|$. □

## III. Case Study

A damped pendulum in Figure 2 is an interesting control problem which has been widely studied. Here we apply the D-type ILC algorithm for learning to track the desired angle & corresponding angular velocity by generating the desired control signal at each sample time. The set of desired control signal for a complete swing has been obtained from a fine-tuned PD controller. The angle $\theta$ is measured anti-clock-wise. The control input torque $u$ is applied in anti-clock-wise direction as well. The pendulum is at rest with initial position $\theta_0 = \pi / 4$ radians at the left side. The initial angular velocity, $\omega_0 = 0$ rad/s. The pendulum has been simulated using following parameters:

Length $L$ = 1m, mass $m$ = 0.5 Kg, acceleration due to gravity $g$ = 9.81 m/s2 and damping co-efficient $b$ = 0.25

N-s/m. Using state vector $x = [\theta, \quad \omega]^T$, the linearised discrete-time state space matrices for the pendulum, sampled at 0.05 seconds are as follows:

$$A_d = \begin{bmatrix} 0.9879 & 0.0492 \\ -0.4824 & 0.9633 \end{bmatrix}, \quad B_d = \begin{bmatrix} 0.0012 \\ 0.0492 \end{bmatrix}, \\ C_d = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad D_d = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{30}$$

The time period for the pendulum is 2 seconds, so there are 40 samples for each swing from left to right and back.

For ILC update, the error considered is the state error. The selected learning gain matrix $K_d = 0.5 \times [0,1]$ allows the forward difference error of the angular velocity only to update the control inputs. The tracking of the desired control input $u$, angle $\theta$ and angular velocity $\omega$ has been monitored for 2500 swings in Figure 3 and Figure 4.

The RMS values reached minima around 2000 swings:

Final RMS of error in angle = 1.2015e-015,

Final RMS of error in angular velocity = 2.2210e-015,

Final RMS of control input error = 1.6876e-014

These values show that limit of precision has been reached. The evolution of the control input errors at selected time indices is shown in Figure 5. The semi-log plot in Figure 6 give the convergence rate of control input errors at selected time indices. The rate of convergence varies in the earlier swings. The convergence occurs sequentially as the control input for earlier time indices converge before the later ones. The kinks in convergence rates for initial swings occur due to oscillations or zero-crossings of the control input errors. Finally, the D-ILC algorithm achieves its limit at large batch numbers. As shown in Figure 6, that just before the control input errors are near the minimum threshold, all the individual components $\Delta u(i,k)$ achieve the same rate (0.9754) equal to the eigenvalue of matrix $T$.

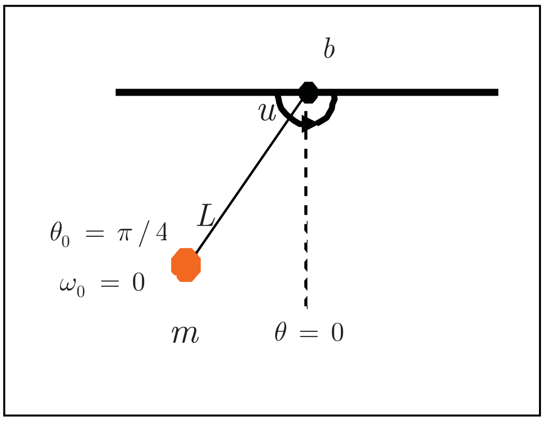


**Figure 2 Damped Pendulum**

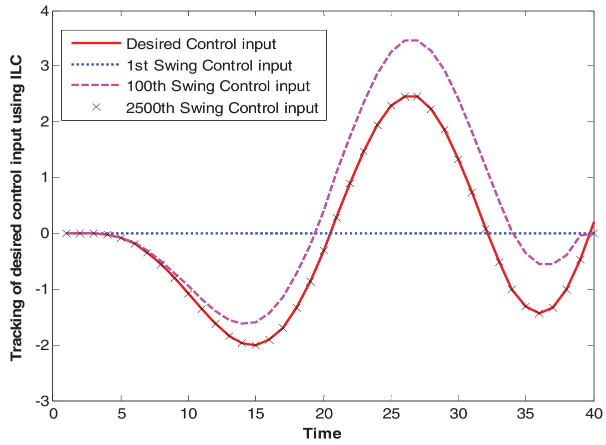


**Figure 3 Tracking control input**

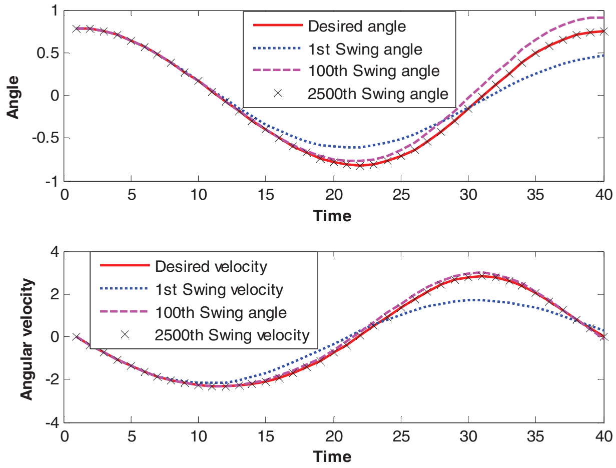


**Figure 4 Tracking angle and velocity**

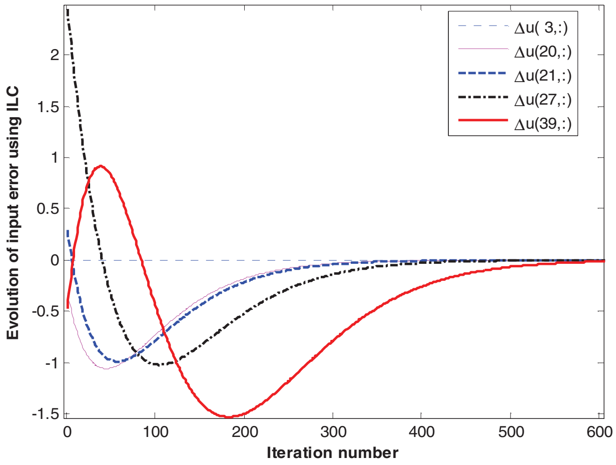


**Figure 5 Evolution of control input errors**

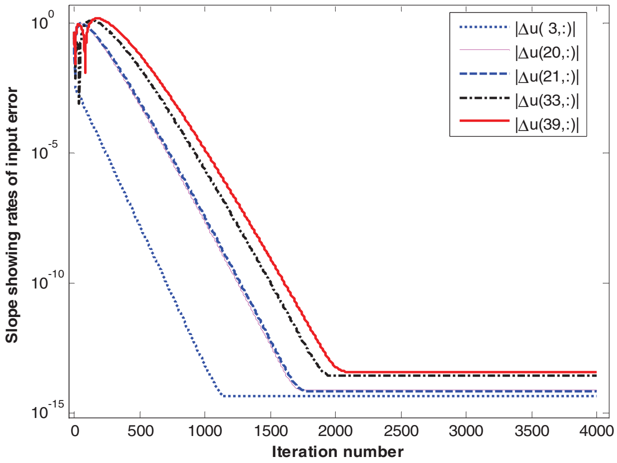


**Figure 6 Convergence rate of control input errors**

## IV. Conclusion

The rate of convergence of D-ILC algorithm has been investigated in terms of input errors at component level which evolve sequentially. In the long term, all input components acquire same convergence rate equal to the eigenvalue of the Toeplitz matrix which relates the input errors from batch to batch. This knowledge can be helpful in designing the Iterative Learning controllers and their performance comparison.

## References

[1] Hyo-Sung, A., C. YangQuan, and K.L. Moore, *Iterative Learning Control: Brief Survey and Categorization.* Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 2007. **37**(6): p. 1099-1121.

[2] UCHIYAMA, M., *Formation of high speed motion pattern of mechanical arm by trial.* . Transactions of the Society of Instrumentation and Control Engineers, 1978. **19**: p. 706-712.

[3] Craig, J.J. *Adaptive control of manipulators through repeated trials*. in *Proceedings of American Control Conference*. 1984. San Diego, CA,.

[4] Arimoto, S., S. Kawamura, and F. Miyazaki, *Bettering operation of Robots by learning.* Journal of Robotic Systems, 1984. **1**(2): p. 123-140.

[5] Kawamura, S., F. Miyazaki, and S. Arimoto. *Applications of learning method for dynamic control of robot manipulators*. in *24th IEEE Conference on Decision and Control*. 1985.

[6] Arimoto, S., S. Kawamura, and F. Miyazaki, *Convergence, stability and robustness of learning control schemes for robot manipulators*, in *Proceedings of the International Symposium on Robot Manipulators on Recent trends in robotics: modeling, control and education*. 1986, Elsevier North-Holland, Inc.: Albuquerque, New Mexico, United States. p. 307 - 316.

[7] Saab, S.S., *On the P-type learning control.* Automatic Control, IEEE Transactions on, 1994. **39**(11): p. 2298-2302.

[8] Owens, D.H. *Iterative learning control-convergence using high gain feedback*. in *Proceedings of the 31st IEEE Conference on Decision and Control*. 1992.

[9] Saab, S.S., *A discrete-time learning control algorithm for a class of linear time-invariant systems.* Automatic Control, IEEE Transactions on, 1995. **40**(6): p. 1138-1142.

[10] Samer, S.S., *Robustness and convergence rate of a discrete-timelearning control algorithm for a class of nonlinear systems.* International Journal of Robust and Nonlinear Control, 1999. **9**(9): p. 559-571.

[11] Arimoto, S., S. Kawamura, and F. Miyazaki. *Bettering operation of dynamic systems by learning: A new control theory for servomechanism or mechatronics systems*. in *The 23rd IEEE Conference on Decision and Control*. 1984.

[12] Bondi, P., G. Casalino, and L. Gambardella, *On the iterative learning control theory for robotic manipulators.* IEEE Journal of Robotics and Automation, 1988. **4**(1): p. 14-22.

[13] Wang, D., *On D-type and P-type ILC designs and anticipatory approach.* International Journal of Control, 2000. **73**(10): p. 890.

[14] Xu, J.-x. and Y. Tan, *On the P-type and Newton-type ILC schemes for dynamic systems with non-affine-in-input factors.* Automatica, 2002. **38**(7): p. 1237-1242.

[15] Shao-Juan, Y., W. Ju-Hua, and Y. Xue-Wen. *A PD-type open-closed-loop iterative learning control and its convergence for discrete systems*. in *Machine Learning and Cybernetics, 2002. Proceedings. 2002 International Conference on*. 2002.

[16] Zhang, B., G. Tang, and S. Zheng, *PD-type iterative learning control for nonlinear time-delay system with external disturbance.* Journal of Systems Engineering and Electronics, 2006. **17**(3): p. 600-605.

[17] YangQuan, C. and K.L. Moore. *PI-type iterative learning control revisited.* in *Proceedings of the American Control Conference*. 2002.

[18] Lequin, O., et al., *Iterative feedback tuning of PID parameters: comparison with classical tuning rules.* Control Engineering Practice, 2003. **11**(9): p. 1023-1033.

[19] Madady, A., *PID Type Iterative Learning Control with Optimal Gains.* International Journal of Control, Automation, and Systems, 2008. **6**(2): p. 194-203.

[20] Heinzinger, G., et al. *Robust learning control.* in *Proceedings of the 28th IEEE Conference on Decision and Control*. 1989.

[21] Yongqiang, Y., A. Tayebi, and L. Xiaoping. *A Unit-Gain D-type Iterative Learning Control Scheme: Application to a 6-DOF Robot Manipulator.* in *Intelligent Control, 2007. ISIC 2007. IEEE 22nd International Symposium on*. 2007.

[22] Galor, O., *Multi-Dimensional, First-Order, Linear Systems: Characterization*, in *Discrete Dynamical Systems*. 2007, Springer Berlin Heidelberg. p. 59-91.

[23] Nocedal, J. and S.J. Wright, *Numerical Optimization*, in *Springer Series in Operations Research and Financial Engineering*, T.V. Mikosch, S.I. Resnick, and S.M. Robinson, Editors. 2006, Springer: New York.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Stability Analysis and Control of Linear Periodic Time-Delay Systems with State-Space Models Based on Semi-Discretization

Chenhui Shao
Department of Mechanical Engineering
University of Michigan, Ann Arbor
Ann Arbor, MI, USA
chshao@umich.edu

Jie Sheng
Department of Automation
University of Science and Technology of China
Hefei, China
jsheng@ustc.edu.cn

*Abstract*—**Stability analysis and control for linear periodic time-delay systems described by state space models are investigated in this paper. Semi-discretization method is used to develop a mapping of the system response in a finite-dimensional state space. The stability region and stability boundary can be found by comparing the maximum absolute value of the mapping's eigenvalues with 1. More importantly, an efficient stability criterion is presented for linear periodic neutral systems. Besides, minimization of the maximum absolute value of the mapping's eigenvalues leads to optimal control gains. Two numerical examples are given to illustrate the proposed method's effectiveness.**

***Keywords-periodic time-delay systems; periodic neutral systems; stability analysis; feedback control; semi-discretization.***

## I. Introduction

Generally speaking, linear periodic time-delay systems can be divided into three categories according to the system complexity: (i) systems with a single time-delay, (ii) systems with multiple time-delays and (iii) neutral systems. Linear periodic systems with a single or multiple time-delays are actually special cases of linear periodic neutral systems.

In recent years, the stability analysis of linear time-delay systems, especially neutral systems, has attracted considerable attention. Among different derivation methods for stability criteria, one main method is based on Lyapunov-Krasovskii functional (LKF) and linear matrix inequality (LMI) [1-5]. In [1] delay-dependent stability conditions were obtained for neutral systems with time-varying delays in terms of LMIs and descriptor model transformation. Reference [2] derived a stability criterion which was formulated in an LMI for uncertain neutral systems with norm-bounded or time-varying uncertainty. A discretized LKF approach was developed to analyze the stability of linear neutral systems with mixed neutral and discrete delays in [3]. Combining the parameterized model transformation method with a method taking the relationships between the terms in the Leibniz-Newton formula into account and using LMI, [4] put forward delay-dependent robust stability criteria and a stabilizing method for neutral systems. Reference [5] studied the stability of neutral type systems with uncertain time-varying delays and norm-bounded uncertainties by utilizing the input-output approach and LKF. Besides, some other methods have also been introduced, e.g. [6] presented necessary and sufficient conditions for delay-dependent stability by principle of the argument.

Meanwhile, control for linear time-delay systems has also been investigated intensively using LMI, LKF and sliding-mode control (SMC) [7-12]. Reference [7] proposed a sufficient condition for the solvability of the design of memoryless state feedback controllers in terms of LMI and gave an explicit expression for the desired controller. In [8], a criterion for the existence of dynamic output feedback controllers was derived based on the LMI and LKF, and a parameterized characterization of the controllers was also given. In [9], LMI optimization approach is used to design the robust output dynamic observer-based controls. Reference [10] raised a robust control design method with LMI and discretized parameter-dependent LKF. In [11], the design of SMC for a class of neutral delay systems with uncertainties in both the state matrices and the input matrix was studied. Reference [12] put forward a robust adaptive SMC design scheme for discrete-time state-delay systems with mismatched uncertainties and external disturbances.

Although there are a large number of references discussing linear systems with time-delay, one type of time-delay system, i.e. linear periodic time-delay systems, has not received much attention. Semi-discretization has been introduced to study this type of time-delay systems. This method proposes to discretize some spatial or temporal variables and treat the rest of them as continuous variables, and thus the exact solution of linear systems can be obtained to construct a very accurate mapping of the state vector over a mapping time step. Therefore semi-discretization is able to handle periodic time-delay systems. Semi-discretization was utilized for the stability analysis of second order linear periodic time-delay system with a single delay represented by ordinary differential equations (ODEs) in [13]. In [14] the stability analysis of higher order systems with both discrete and continuous time-delays was studied based on semi-discretization. Reference [15] reported an application of semi-discretization to the feedback controls and optimal gain design of linear periodic systems with a single time-delay.

This work was supported by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry (WF2100100001) and the Fundamental Research Funds for the Central Universities (WK2100100003).

Reference [16] extended this idea to the stability analysis of linear neutral systems represented by ODEs.

In this paper, we further extend this method to the investigation regarding stability analysis and control design of all types of linear periodic time-delay systems described by state space models.

The rest of this paper is organized as follows. In Section II, semi-discretization method is deployed to derive stability criteria for different types of linear periodic time-delay systems. Also, optimal feedback control of such systems is discussed. Section III presents two numerical examples to illustrate this method's effectiveness. The conclusions are presented in Section IV.

## II. Main Results

In this section, two topics are covered: Subsection A presents stability criteria for three types of linear periodic time-delay systems. Subsection B introduces one scheme for the optimal feedback control design.

### *A. Stability Criteria*

Linear periodic time-delay systems have a general state space form as shown by (1).

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \sum_{j=1}^{m}[\mathbf{B}_j(t)\mathbf{x}(t-\tau_j) + \mathbf{C}_j(t)\dot{\mathbf{x}}(t-\tau_j)], \quad (1)$$

where $\mathbf{x} \in \mathbf{R}^N$, and the coefficient matrices $\mathbf{A}(t)$, $\mathbf{B}_j(t)$ and $\mathbf{C}_j(t)$ are all periodic matrices with period $T$. $m$ is the number of time-delays. Without loss of generality, it is assumed that $\tau_j$'s are already arranged such that $\tau_1 < \tau_2 < \cdots < \tau_m$.

With different constraints on $m$ and $\mathbf{C}(t)$'s , (1) is able to represent different types of linear periodic time-delay systems. The general forms of periodic time-delay systems with a single delay and periodic time-delay systems with multiple delays are given by (2) and (3).

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{x}(t-\tau) \quad (2)$$

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \sum_{j=1}^{m}\mathbf{B}_j(t)\mathbf{x}(t-\tau_j) \quad (3)$$

Because (2) and (3) are special cases of (1), stability criteria for linear periodic systems with a single or multiple time-delays are derivable from the criterion for linear periodic neutral systems. Hence Theorem 1 will be first introduced as a stability criterion for linear periodic neutral systems, and then the stability criteria for the other two types of systems are given in Corollaries 1 and 2, respectively.

*Theorem 1*: Suppose for one linear periodic time-delay system, the mapping of the state vector over one period is obtained using semi-discretization, and is shown as

$$\mathbf{y}_{p+1} = \mathbf{\Phi}\mathbf{y}_p, \quad (4)$$

where $p$ indicates the index of one period, and $\mathbf{y}_p$ is the state vector at the beginning of the $p$th period. $\mathbf{\Phi}$ is the product of transition matrices over the period, namely,

$$\mathbf{\Phi} = \left(\prod_{i=0}^{k-1}\mathbf{H}_i^T\right)^T, \quad (5)$$

where $\mathbf{H}_i$ is the transition matrix over a time interval $[t_i, t_{i+1}]$, and satisfies (6). Note that the period $T$ is equally partitioned into $k$ intervals.

$$\mathbf{y}_{i+1} = \mathbf{H}_i\mathbf{y}_i, \quad (6)$$

where $i$ indicates the index of one time interval, and $\mathbf{y}_i$ is the state vector at the beginning of the $i$th interval.

The transition matrices $\mathbf{H}_i$'s are given in (7).

$$\mathbf{H}_i = \begin{bmatrix} \mathbf{H}_{i,11} & \mathbf{H}_{i,12} \\ \mathbf{H}_{i,21} & \mathbf{H}_{i,22} \\ \mathbf{H}_{i,31} & \mathbf{H}_{i,32} \\ \mathbf{H}_{i,41} & \mathbf{H}_{i,42} \end{bmatrix}_{2(n_m+1)N\times 2(n_m+1)N} \quad (7)$$

where

$$\mathbf{H}_{i,11} = \begin{bmatrix}\mathbf{Q}_i & \mathbf{0}_{N\times(n_1-1)N} & \mathbf{P}_i\mathbf{B}_{1,i} & \cdots & \mathbf{0}_{N\times(n_m-n_{m-1}-1)N} & \mathbf{P}_i\mathbf{B}_{m,i}\end{bmatrix}_{N\times(n_m+1)N},$$
$$\mathbf{H}_{i,12} = \begin{bmatrix}\mathbf{0}_{N\times n_1 N} & \mathbf{P}_i\mathbf{C}_{1,i} & \cdots & \mathbf{0}_{N\times(n_m-n_{m-1}-1)N} & \mathbf{P}_i\mathbf{C}_{m,i}\end{bmatrix}_{N\times(n_m+1)N},$$
$$\mathbf{H}_{i,21} = \begin{bmatrix}\mathbf{I}_{n_m N\times n_m N} & \mathbf{0}_{n_m N\times N}\end{bmatrix}_{n_m N\times(n_m+1)N},$$
$$\mathbf{H}_{i,22} = \mathbf{0}_{n_m N\times(n_m+1)N},$$
$$\mathbf{H}_{i,31} = \begin{bmatrix}\mathbf{A}_i\mathbf{Q}_i & \mathbf{0}_{N\times(n_1-2)N} & \mathbf{B}_{1,i} & \mathbf{A}_i\mathbf{P}_i\mathbf{B}_{1,i} & \cdots & \mathbf{B}_{m,i} & \mathbf{A}_i\mathbf{P}_i\mathbf{B}_{m,i}\end{bmatrix}_{N\times(n_m+1)N},$$
$$\mathbf{H}_{i,32} = \begin{bmatrix}\mathbf{0}_{N\times(n_1-1)N} & \mathbf{C}_{1,i} & \mathbf{A}_i\mathbf{P}_i\mathbf{C}_{1,i} & \cdots & \mathbf{C}_{m,i} & \mathbf{A}_i\mathbf{P}_i\mathbf{C}_{m,i}\end{bmatrix}_{N\times(n_m+1)N},$$
$$\mathbf{H}_{i,41} = \mathbf{0}_{n_m N\times(n_m+1)N},$$
$$\mathbf{H}_{i,42} = \begin{bmatrix}\mathbf{0}_{n_m N\times N} & \mathbf{I}_{n_m N\times n_m N}\end{bmatrix}_{n_m N\times(n_m+1)N}.$$

In (7):

$$\begin{aligned}&\mathbf{A}_i = \mathbf{A}(t_i), \mathbf{B}_{j,i} = \mathbf{B}_j(t_i), \mathbf{C}_{j,i} = \mathbf{C}_j(t_i),\\ &\mathbf{Q}_i = e^{\mathbf{A}_i\Delta t}, \mathbf{P}_i = \int_{t_i}^{t_{i+1}} e^{\mathbf{A}_i(\Delta t-\hat{t})}d\hat{t},\\ &n_j = \tau_j/\Delta t, j = 1,2,...,m.\end{aligned} \quad (8)$$

Let $|\lambda|_{\max}$ denote the maximal absolute value of eigenvalues of $\mathbf{\Phi}$. Then the stability region is determined by

$$|\lambda|_{\max} < 1, \quad (9)$$

and the stability boundary is determined by

$$|\lambda|_{\max} = 1. \quad (10)$$

*Proof*: First let us discretize the period $T$ into an integer $k$ intervals of equal length $\Delta t$ such that $T = k\Delta t$.

Consider (1) in a time interval $t \in [t_i, t_{i+1}]$, where $t_i = i\Delta t, i = 0,1,2,...,k$. In each small time interval $[t_i, t_{i+1}]$, the delayed responses $\mathbf{x}(t-\tau_j)$ and time-dependent coefficient matrices are assumed to be constant. Thus we have the notations in (8). Besides we denote

$$\begin{aligned} &\mathbf{x}(t_i) = \mathbf{x}_i, \\ &\mathbf{x}(t_{i+1}) = \mathbf{x}_{i+1}, \\ &\mathbf{x}(t-\tau_j) = \mathbf{x}(i\Delta t - n_j \Delta t) = \mathbf{x}_{i-n_j}. \end{aligned} \tag{11}$$

Within one particular interval, (1) becomes:

$$\dot{\mathbf{x}}(t) - \mathbf{A}_i \mathbf{x}(t) - \sum_{j=1}^{m} [\mathbf{B}_{j,i} \mathbf{x}(t-\tau_j) + \mathbf{C}_{j,i} \dot{\mathbf{x}}(t-\tau_j)] = 0. \tag{12}$$

The general solution of (12) is given by

$$\mathbf{x}(t) = e^{\mathbf{A}_i (t-t_i)} \mathbf{x}_i + \int_{t_i}^{t} e^{\mathbf{A}_i (t-t_i-\hat{t})} \sum_{j=1}^{m} [\mathbf{B}_{j,i} \mathbf{x}(\hat{t}-\tau_j) + \mathbf{C}_{j,i} \dot{\mathbf{x}}(\hat{t}-\tau_j)] d\hat{t}, \tag{13}$$

where $t \in [t_i, t_{i+1}]$.

Let $t$ be $t_{i+1}$ in (13), then

$$\mathbf{x}_{i+1} = \mathbf{Q}_i \mathbf{x}_i + \mathbf{P}_i \sum_{j=1}^{m} (\mathbf{B}_{j,i} \mathbf{x}_{i-n_j} + \mathbf{C}_{j,i} \dot{\mathbf{x}}_{i-n_j}), \tag{14}$$

where the definition of $\mathbf{Q}_i$ and $\mathbf{P}_i$ is given in (8).

In order to construct a mapping over one time interval, one $2(n_m + 1) \times N$ dimensional state vector is defined as:

$$\mathbf{y}_i = \begin{bmatrix} \mathbf{x}_i^T & \mathbf{x}_{i-1}^T & \cdots & \mathbf{x}_{i-n_m}^T & \dot{\mathbf{x}}_i^T & \dot{\mathbf{x}}_{i-1}^T & \cdots & \dot{\mathbf{x}}_{i-n_m}^T \end{bmatrix}^T. \tag{15}$$

To find the transition matrix $\mathbf{H}_i$ is equivalent to find a mapping from $\mathbf{y}_i$ to $\mathbf{y}_{i+1}$. $\mathbf{y}_{i+1}$ has the form in (16).

$$\mathbf{y}_{i+1} = \begin{bmatrix} \mathbf{x}_{i+1}^T & \mathbf{x}_i^T & \cdots & \mathbf{x}_{i-n_m+1}^T & \dot{\mathbf{x}}_{i+1}^T & \dot{\mathbf{x}}_i^T & \cdots & \dot{\mathbf{x}}_{i-n_m+1}^T \end{bmatrix}^T. \tag{16}$$

All entries except $\mathbf{x}_{i+1}^T$ and $\dot{\mathbf{x}}_{i+1}^T$ in $\mathbf{y}_{i+1}$ also appear in $\mathbf{y}_i$, which indicates that corresponding entries in $\mathbf{H}_i$ are identity matrices with appropriate dimensions.

Moreover, (14) actually provides a representation of $\mathbf{x}_{i+1}$ using the entries in $\mathbf{y}_i$. Hence the last problem lies in the representation of $\dot{\mathbf{x}}_{i+1}$.

Let $t$ be $t_{i+1}$ in (12), and it becomes:

$$\dot{\mathbf{x}}_{i+1} = \mathbf{A}_i \mathbf{x}_{i+1} + \sum_{j=1}^{m} (\mathbf{B}_{j,i} \mathbf{x}_{i-n_j+1} + \mathbf{C}_{j,i} \dot{\mathbf{x}}_{i-n_j+1}). \tag{17}$$

Plug (14) into (17) and we get (18).

$$\begin{aligned} \dot{\mathbf{x}}_{i+1} = {} & \mathbf{A}_i \mathbf{Q}_i \mathbf{x}_i + \\ & \sum_{j=1}^{m} [\mathbf{A}_i \mathbf{P}_i \mathbf{B}_{j,i} \mathbf{x}_{i-n_j} + \mathbf{B}_{j,i} \mathbf{x}_{i-n_j+1} + \mathbf{A}_i \mathbf{P}_i \mathbf{C}_{j,i} \dot{\mathbf{x}}_{i-n_j} + \mathbf{C}_{j,i} \dot{\mathbf{x}}_{i-n_j+1}]. \end{aligned} \tag{18}$$

Combining the results above together, and the transition matrix over $[t_i, t_{i+1}]$ can then be expressed as (17).

The mapping of the state vector over one period $T = k\Delta t$ is therefore shown by (4), and $\mathbf{\Phi}$ is given by (5).

The stability of the system is determined by the eigenvalues of $\mathbf{\Phi}$, i.e.

$$|\mathbf{y}_{p+1}| \le |\lambda|_{\max} |\mathbf{y}_p|. \tag{19}$$

When $|\lambda|_{\max} < 1$, $\mathbf{\Phi}$ is a contraction, and the system is asymptotically stable. Then the stability region and stability boundary can be found using (9) and (10), respectively.

Moreover, (19) also indicates that the smaller $|\lambda|_{\max}$ is, the faster the system response converges to zero. Hence $|\lambda|_{\max}$ also provides an indicator of the control performance.

□

Furthermore, as particular cases of the previous result, the stability criteria for periodic time-delay systems with a single delay and multiple delays are presented in the sequel.

*Corollary 1*: For a linear periodic time-delay system with one single time-delay, the stability region and stability boundary are determined by (9) and (10), respectively. The mapping over one period and the corresponding mapping matrix $\mathbf{\Phi}$ are the same as in Theorem 1, and transition matrices $\mathbf{H}_i$'s can be instead calculated using (20).

$$\mathbf{H}_i = \begin{bmatrix} \mathbf{H}_{i,1} \\ \mathbf{H}_{i,2} \end{bmatrix}_{(n+1)N \times (n+1)N}, \tag{20}$$

where

$$\begin{aligned} &\mathbf{H}_{i,1} = \begin{bmatrix} \mathbf{Q}_i & \mathbf{0}_{N \times (n-1)N} & \mathbf{P}_i \mathbf{B}_i \end{bmatrix}, \\ &\mathbf{H}_{i,1} = \begin{bmatrix} \mathbf{I}_{nN \times nN} & 0_{nN \times N} \end{bmatrix}, \\ &n = \tau / \Delta t, \end{aligned}$$

and $\mathbf{Q}_i$, $\mathbf{P}_i$ and $\mathbf{B}_i$ are the same as in (8).

*Corollary 2*: For a linear periodic time-delay system with multiple time-delays, the stability region and stability boundary are determined by (9) and (10), respectively. The mapping over one period and the corresponding mapping matrix $\mathbf{\Phi}$ are the same as in Theorem 1, and transition matrices can be found with (21).

$$\mathbf{H}_i = \begin{bmatrix} \mathbf{H}_{i,1} \\ \mathbf{H}_{i,2} \end{bmatrix}, \tag{21}$$

where

$$\mathbf{H}_{i,1} = \begin{bmatrix} \mathbf{Q}_i & \mathbf{0}_{N\times(n_1 N - N)} & \mathbf{P}_i \mathbf{B}_{1,i} & \cdots & \mathbf{0}_{N\times(n_m N - n_{m-1} N)} & \mathbf{P}_i \mathbf{B}_{m,i} \end{bmatrix},$$
$$\mathbf{H}_{i,2} = \mathbf{I}_{(n_m+1)N\times(n_m+1)N},$$
$$n_j = \tau_j / \Delta t,\ j = 1, 2, ..., m,$$

and $\mathbf{Q}_i$, $\mathbf{P}_i$ and $\mathbf{B}_{j,i}$ are the same as in (8).

### *B. Optimal Feedback Control*

In this section, we consider the optimal feedback control design for linear periodic time-delay systems.

Equation (22) is a general linear periodic time-delay system with feedback control, which can be either delayed or not.

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \sum_{j=1}^{m} [\mathbf{B}_j(t)\mathbf{x}(t-\tau_j) + \mathbf{C}_j(t)\dot{\mathbf{x}}(t-\tau_j)] + \mathbf{D}(t)\mathbf{u}(t), \tag{22}$$

where $\mathbf{x} \in \mathbf{R}^N$, $\mathbf{u} \in \mathbf{R}^M$. $\mathbf{A}(t)$, $\mathbf{B}_j(t)$, $\mathbf{C}_j(t) \in \mathbf{R}^{N\times N}$, and $\mathbf{D}(t) \in \mathbf{R}^{N\times M}$ are periodic matrices with period $T$.

According to whether delays exist in $\mathbf{u}(t)$ or not, $\mathbf{u}(t)$ have two different forms which are represented by (23).

$$\mathbf{u}(t) = -\mathbf{K}\mathbf{x}(t) \text{ or } \mathbf{u}(t) = -\mathbf{K}\mathbf{x}(t-\tau_u). \tag{23}$$

where $\mathbf{K} \in \mathbf{R}^{M\times N}$ is the gain matrix.

Obviously, (22) is able to be transformed into the form of (1). Therefore, the stability region and stability boundary can be obtained by using stability criteria presented in Subsection A. Furthermore, optimal control gains can also be found, and the search scheme is discussed as follows.

If the problem is restricted in a finite and compact region $\mathbf{\Omega}$ in the parametric space $\mathbf{K}$, we can find the regions of stability and optimal control gains in the region to minimize $|\lambda|_{\max}$. This leads to one optimization problem:

$$\min_{\mathbf{K}\in\mathbf{\Omega}}[\max|\lambda(\mathbf{\Phi})|] \text{ subject to } |\lambda|_{\max} < 1. \tag{24}$$

This optimization formulation offers an approach to design feedback controls for linear periodic time-delay systems, and the control performance criterion is chosen as the decay rate of the mapping $\mathbf{\Phi}$ over one period. For some further discussion, please refer to Sheng and Sun [15].

## III. Examples

In this section, two examples are presented to illustrate semi-discretization's validity. All computations are carried out by Matlab R2010b.

Example 1: Equation (25) describes a linear time-invariant neutral system.

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{x}(t-\tau) + \mathbf{C}\dot{\mathbf{x}}(t-\tau), \tag{25}$$

where

$$\mathbf{A} = \begin{bmatrix} -2 & 0 \\ 0 & -0.9 \end{bmatrix},\ \mathbf{B} = \begin{bmatrix} -1 & 0 \\ -1 & -1 \end{bmatrix},\ \mathbf{C} = \begin{bmatrix} c & 0 \\ 0 & c \end{bmatrix},\ 0 \le c \le 1.$$

When $c = 0$, this system is simplified to be a linear time-delay system with a single time-delay, so Corollary 1 is applied to get the upper stability bound of $\tau$; when $c \neq 0$, this system is a linear neutral system, and Theorem 1 is used.

The upper stability bound is calculated with different criteria, which are given in [1, 2, 4 and 6]. The comparison with the criteria presented in this paper is listed in Table I.

Table I suggests that the upper bounds obtained by the methods in this paper are much less conservative and much more accurate than those of [1, 2 and 4], and meanwhile they are comparable with those of [6].

TABLE I. Comparison of Time-Delay's Upper Bound for Example 1

| Criteria \ c | 0 | 0.5 | 0.9 |
|---|---|---|---|
| [1] | 4.47 | 1.14 | 0.13 |
| [2] | 4.35 | 3.62 | 0.99 |
| [4] | 4.47 | 3.67 | 1.41 |
| [6] | 6.15-6.2 | 4.735-4.74 | 1.57-1.575 |
| Semi-Discretization | 6.17 | 4.67 | 1.52 |

Example 2: A linear periodic neutral system is given by (26).

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \sum_{j=1}^{2} [\mathbf{B}_j\mathbf{x}(t-\tau_j) + \mathbf{C}_j\dot{\mathbf{x}}(t-\tau_j)], \tag{26}$$

where

$$\mathbf{A}(t) = \begin{bmatrix} -0.9 & -1 \\ 1+0.3\sin 2t & 0.9+0.4\cos 4t \end{bmatrix},$$
$$\mathbf{B}_1 = \begin{bmatrix} 0 & 0 \\ -k_1 & -k_2 \end{bmatrix},\ \mathbf{B}_2 = \begin{bmatrix} 1 & 0.4 \\ 0 & 1 \end{bmatrix},$$
$$\mathbf{C}_1 = \begin{bmatrix} 0 & -0.4 \\ -0.5 & 0 \end{bmatrix},\ \mathbf{C}_2 = \begin{bmatrix} -0.5 & 0 \\ 0 & -0.5 \end{bmatrix},$$
$$\tau_1 = \frac{\pi}{10},\ \tau_2 = \frac{\pi}{5}.$$

In this example, the coefficient matrix $\mathbf{A}(t)$ is periodic with period $\pi$, and the feedback controller is delayed with delay $\pi/10$.

Theorem 1 is utilized for searching the stability region on the $k_1 - k_2$ plane. Besides, the optimal pair of control gains is also found by solving the optimization problem stated by (24).

The stability region and the optimal control gain pair are shown by Fig. 1, where the labels of contours are the maximum absolute values of eigenvalues of the mapping matrices, i.e. $|\lambda|_{\max}$. The stability region is within the bold black line, and the optimal gain pair is $(k_1,\ k_2) = (0.3556,\ 2.8256)$, with $|\lambda|_{\max}$ = 0.3234, which is indicated by red "+". Besides, the blue circle and black triangle indicate $(0.8,\ 3)$ and $(1,\ 2)$, respectively.

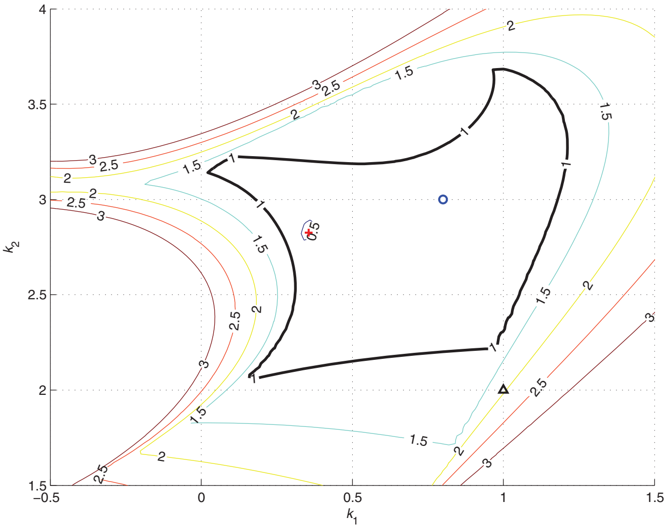


Figure 1. Stability region and the optimal pair of control gains.

Fig. 2 shows simulation results when different pairs of control gains are applied for (26), where the horizontal axis is simulation time, and the vertical axis is the norm of state, which is defined as $\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2}$. The initial conditions are set to be $\mathbf{x}_0 = [0 \quad 0.1]^T$. Note that $(0.3556,\ 2.8256)$ is the optimal controller, and therefore the corresponding system response converges to zero fastest. $(0.8,\ 3)$ is within the stability region, but the corresponding $|\lambda|_{\max}$ is 0.8403, which is larger than that when the optimal controller is applied, so the system response converges to zero asymptotically yet slowly. $(1,\ 2)$ is one pair causing the system to be unstable, with $|\lambda|_{\max} = 1.9439 > 1$, and hence the system response increases very quickly.

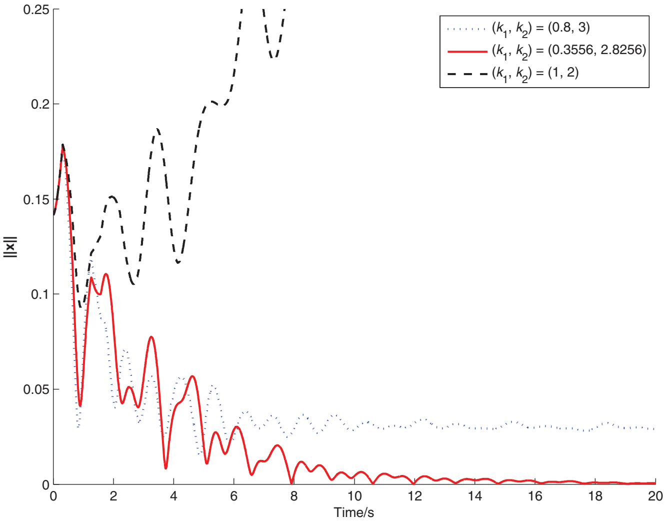


Figure 2. System responses with different control gains.

## IV. Conclusion

In this paper, stability analysis and control design of linear periodic time-delay systems have been investigated. Based on the semi-discretization method, stability criteria have been derived for different types of linear periodic time-delay systems, and optimal feedback control design scheme has also been presented. Two illustrative examples have been given to demonstrate the merits of the obtained results.

## References

[1] E. Fridman and U. Shaked, "Delay-dependent stability and H-infinite control: constant and time-varying delays," International Journal of Control, vol. 76, pp. 48-60, 2003.

[2] Q. Han, "Robust stability of uncertain delay-differential systems of neutral type," Automatica, vol. 38, pp. 719-723, 2002.

[3] Q. Han, "On stability of linear neutral systems with mixed time delay: A discretized Lyapunov functional approach," Automatica, vol. 41, pp. 1209-1218, 2005.

[4] M. Wu, Y. He and J. She, "New delay-dependent stability criteria and stabilization method for neutral systems," IEEE Transactions on Automatic Control, vol. 49, pp. 2266-2270, 2004.

[5] E. Fridman, "On robust stability of linear neutral systems with time-varying delays," IMA Journal of Mathematical Control and Information, vol. 25, pp. 393-407, 2008.

[6] G. Hu and M. Liu, "Stability criteria of linear neutral systems with multiple delays," IEEE Transactions on Automatic Control, vol. 52, pp. 720-724, 2007.

[7] S. Xu, J. Lam and C. Yang, "Robust $H_\infty$ control for uncertain linear neutral delay systems," Optimal Control Applications & Methods, vol. 23, pp. 113-123, 2002.

[8] J. H. Park, "Design of dynamic controller for neutral differential systems with delay in control input," Chaos, Solitons and Fractals, vol. 23, pp. 503-509, 2005.

[9] J. Chen, "Robust output observer-based control of neutral uncertain systems with discrete and distributed time delays: LMI optimization approach," Chaos, Solitons and Fractals, vol. 34, pp. 1254-1264, 2007.

[10] F. O. Souza, R. M. Palhares and V. J. S. Leite., "Improved robust $H_\infty$ control for neutral systems via discretised Lyapunov-Krasovskii functional," International Journal of Control, vol. 81, pp. 1462-1474, 2008.

[11] Y. Niu, J. Lam and X. Wang, "Sliding-mode control for uncertain neutral delay systems," Control Theory and Applications, vol. 151, pp. 38-44, 2004.

[12] Y. Xia, Z. Zhu, C. Li, H. Yang and Q. Zhu, "Robust adaptive sliding mode control for uncertain discrete-time systems with time delay," Journal of the Franklin Institute, vol. 347, pp. 339-357, 2010.

[13] T. Insperger and G. Stepan, "Semi-discretization of delayed dynamical systems," Proceedings of ASME Design Engineering Technical Conference, vol. 6B, pp. 1227-1232, 2001.

[14] T. Insperger and G. Stepan, "Semi-discretization method for delayed systems," International Journal for Numerical Methods in Engineering, vol. 55, pp. 503-518, 2002.

[15] J. Sheng and J. Q. Sun. "Feedback controls and optimal gain design of delayed periodic linear systems," Journal of Vibration and Control, vol. 11, pp. 277-294, 2005.

[16] J. Sheng, C. Shao, L. Wang, Z. Chen and J. Chao, "Stability analysis and control of linear neutral systems based on method of semi-discretization," 2010 Chinese Control and Decision Conference, CCDC 2010, pp. 723-728, 2010.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# IMAGE RESOLUTION ENHANCEMENT USING MULTI-WAVELET AND CYCLE-SPINNING

*P. Bagheri Zadeh, A. Sheikh Akbari*

*Faculty of Computing, Engineering and Technology, Staffordshire University, Staford, UK*
*E-mails: (p.bagheri-zadeh, a.s.akbari)@staffs.ac.uk*

**Keywords:** Super-resolution, multi-wavelets, cycle-spinning.

## Abstract

In this paper a multi-wavelet and cycle-spinning based image resolution enhancement technique is presented. The proposed technique generates a high-resolution image for the input low-resolution image using the input image and an inverse multi-wavelet transform (all multi-wavelet high frequency subbands' coefficients are set to zero). The concept of the cycle spinning algorithm in conjunction with the multi-wavelet transform is then used to generate a high quality super-resolution image for the input image from the resulting high resolution image, as follows: A number of replicated images with different spatial shifts from the resulting high-resolution image is first generated; Each of the replicated images is de-correlated into its subbands using a multi-wavelet transform; The multi-wavelet high frequency subbands' coefficients of each of the de-correlated images are set to zero and then a primary super-resolution image for each of these images is produced using an inverse multi-wavelet transform; The resulting primary super-resolution images are then spatially shift compensated and the output super-resolution image is created by averaging the resulting shift compensated images. Experimental results were generated using four standard test images and compared to the state of art techniques. Results show that the proposed technique significantly outperforms the classical and non-classical super-resolution methods both subjectively and objectively.

## 1 Introduction

The Super-Resolution (SR) techniques aim at displaying or printing an image at a resolution higher than that the original image without severely affecting the visual quality of the image. Since natural images exhibit abrupt discontinuities at object boundaries, traditional image interpolation methods such as linear and higher-order polynomial interpolation algorithms (which implicitly assume the underlying image is locally smooth) may not be able to offer the required visual quality [1].

Wavelets have been applied to image resolution enhancement to address the over smoothing problem of the traditional super-resolution techniques and promising results have been reported in the literature [2, 4, 7, 8, 9, 10]. The common feature of the wavelet based techniques is the assumption that the image to be enhanced is the wavelet approximation band of the high-resolution image. These techniques generate the high frequency wavelet coefficients from the given low resolution image [9]. A wavelet-based interpolation method was proposed by Carey et al. in [1]. Their proposed method estimates the wavelet high frequency coefficients of the image by exploiting the regularity of edges across the scales. They reported promising results. However, this method may not be able to estimate wavelet high-frequency coefficients with small values. A wavelet based image resolution enhancement algorithm using an adapted cycle-spinning technique was reported in [8]. This algorithm uses the information in the low-resolution image wavelet subbands to estimate the local edge orientation of the image. This local edge orientation information is then used to determine the cycle spinning parameters to generate the enlarged image. Authors reported superior results compared to Kinebuchi and Muresan method. Temizel and Vlachos proposed another wavelet based image resolution enhancement using cycle-spinning [7], which gives superior results to the state of the art. Another wavelet based image resolution enhancement algorithm, which operates in a quad-tree wavelet decomposition framework and exploits wavelet coefficient correlation in a local neighbourhood sense, was presented in [9]. This method employed linear least-squares regression algorithm to estimate the wavelet high-frequency coefficients. This method gives superior results compared to the conventional methods for a wide range of test images. Another image resolution enhancement method in wavelet domain using inter-subband correlation was presented by Piao et al. [4]. They designed filters to estimate high frequency subbands from the lower bands. They reported superior results, up to 0.36dB, to the competing methods.

Wavelets filters generally have poor frequency characteristic. Hence, wavelet baseband of the images usually contain some of the image high-frequency information [4]. This property of the wavelets is usually used to improve the quality of the enlarged image especially around their edges. Applications of multi-wavelet transform [6] that use two or more wavelets in image decomposition generate four or more approximation bands from the input image. Similar to wavelets, multi-wavelet filters also have poor frequency characteristics. Therefore, their resulting approximation bands carry different high-frequency information of the image. This feature of multi-wavelets makes them a potential tool in enhancing image resolution. The application of multi-wavelets in enlarging images and also application of cycle spinning algorithm in multi-wavelet domain for enlarging images have

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

not been reported in the literature. Hence, in this paper, cycle-spinning algorithm is adapted into multi-wavelet domain and used for enlarging images. The rest of the paper is organized as follows: In Section 2, a brief review of the multi-wavelet transform is presented. The proposed image resolution enhancement technique is discussed in Section 3. Experimental results are presented in Section 4. Finally the paper is concluded at Section 5.

## 2 Multi-wavelet transform

Multi-wavelet transforms are very similar to scalar wavelet transforms with some vital differences. Classical wavelet theory is based on the following refinement equations:

$$\begin{aligned} \phi(t) &= \sum_{k=-\infty}^{k=\infty} h_k \phi(mt-k) \\ \psi(t) &= \sum_{k=-\infty}^{k=\infty} g_k \psi(mt-k) \end{aligned} \qquad (1)$$

where $\phi(t)$ is a scaling function, $\psi(t)$ is a wavelet function, $h_k$ and $g_k$ are scalar filters, $m$ represents the subband number and $k$ is the shifting parameter. In contrast to wavelet transforms, multi-wavelets have two or more scaling and wavelet functions. The set of scaling and wavelet functions of a multi-wavelet in vector notation can be defined as:

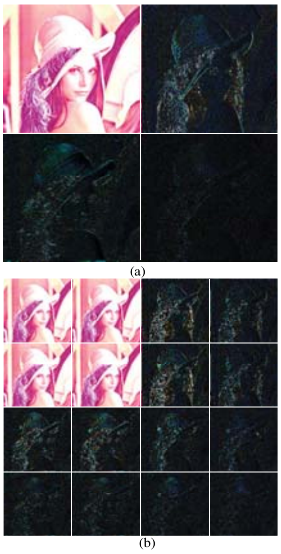
$$\begin{aligned} \Phi(t) &\equiv [\phi_1(t) \quad \phi_2(t) \quad \phi_3(t) \quad \ldots \quad \phi_r(t)]^T \\ \Psi(t) &\equiv [\psi_1(t) \quad \psi_2(t) \quad \psi_3(t) \quad \ldots \quad \psi_r(t)]^T \end{aligned} \qquad (2)$$

where $\Phi(t)$ and $\Psi(t)$ are the multi-scaling and multi-wavelet functions, respectively, with $r$ scaling- and wavelet-functions. In the case of scalar wavelets their multiplicity order $r$ is equal to 1, while multi-wavelets support $r \geq 2$. To date, most multi-wavelets have multiplicity order, $r = 2$. A multi-wavelet with two scaling and wavelet functions can be defined as [6]:

$$\begin{aligned} \Phi(t) &= \sqrt{2} \sum_{k=-\infty}^{k=\infty} H_k \Phi(m\,t-k) \\ \Psi(t) &= \sqrt{2} \sum_{k=-\infty}^{k=\infty} G_k \Psi(m\,t-k) \end{aligned} \qquad (3)$$

where $H_k$ and $G_k$ are $2 \times 2$, ( $r \times r$ ), matrix filters and $m$ is the subband number. Similar to wavelet transforms, multi-wavelets can be implemented using Mallat's filter bank theory [3]. A visual comparison of the resulting subbands for a 2D wavelet and respectively multi-wavelet, is shown in Figures 1(a) and 1(b). From Figure 1, the multi-wavelet transform generates four subbands instead of each subband created by the wavelet transform, and these four subbands carry different spectral content of the input image due to the properties of multi-wavelet filters.

The major advantage of multi-wavelets over scalar wavelets is their ability to possess symmetry, orthogonality and higher order of approximation simultaneously, which is not possible for scalar wavelets. Furthermore, the multi-channel structure of the multi-wavelet transform is a closer approximation to the human visual system than what wavelets offer. These features of multi-wavelets make them a potential tool in enhancing image resolution. Further information about the generation of multi-wavelets, their properties and their applications can be found in [5, 6].


(a)

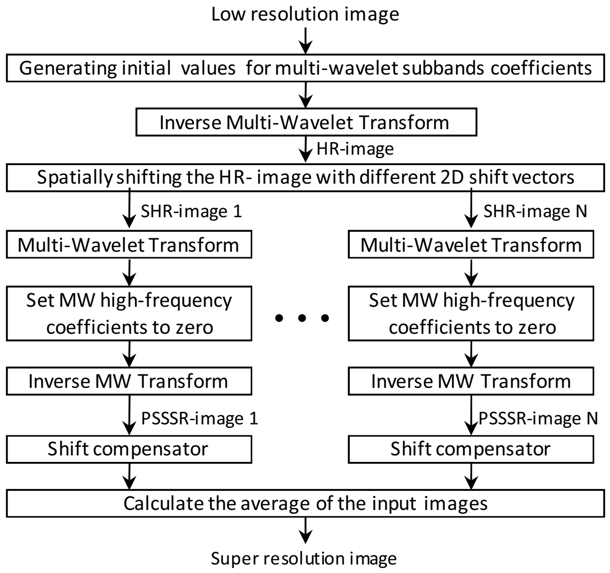
(b)

Figure 1: Single level decomposition of Lena test image (a) Antonini 9/7 wavelet transform, (b) balanced bat01 multiwavelet transform.

## 3 Multi-wavelet based image resolution enhancement technique

Figure 2 shows a block diagram of the proposed multi-wavelet based image resolution enhancement technique. A low resolution image is input to the system. The input image coefficients are weighted to generate four multi-wavelet basebands. The weighting factor for each of the multi-wavelet basebands equals to the overall multi-wavelet filters gain for that subband. The Inverse Multi-Wavelet Transform (IMWT) block produces a High-Resolution (HR) image for the input image using the information in the resulting basebands (the coefficients in multi-wavelet high frequency subbands are set to zero.). The concept of the cycle spinning technique is then combined with the multi-wavelet transform to generate a super-resolution image for the input image from the resulting HR-image, as follows: The resulting HR-image is first spatially shifted with different 2D shift values, generating N Shifted High Resolution (SHR) images. Then the N 2D-shift vectors are determined using equation 4:

$$\begin{bmatrix} (-k,-k) & (-k,-k+1) & \cdots & (-k,k) \\ (-k+1,-k) & (-k+1,-k+1) & \cdots & (-k+1,k) \\ \cdots & \cdots & \cdots & \cdots \\ (k,-k) & (k,-k+1) & \cdots & (k,\ k) \end{bmatrix} \qquad (4)$$



Low resolution image
Generating initial values for multi-wavelet subbands coefficients
Inverse Multi-Wavelet Transform
HR-image
Spatially shifting the HR- image with different 2D shift vectors
SHR-image 1
SHR-image N
Multi-Wavelet Transform
Multi-Wavelet Transform
Set MW high-frequency coefficients to zero
• • •
Set MW high-frequency coefficients to zero
Inverse MW Transform
Inverse MW Transform
PSSSR-image 1
PSSSR-image N
Shift compensator
Shift compensator
Calculate the average of the input images
Super resolution image

Figure 2: Block diagram of the multi-wavelet based image enlargement algorithm.

where $k$ is the maximum number of pixel shifts in both horizontal and vertical directions. In this paper $k$ was set to 3, as system exhibits its best performance with this value. Number of shifted images, N, can be calculated using equation 5:

$$N = (2k + 1)^2 \qquad (5)$$

where $k$ is the maximum number of pixel shifts in both horizontal and vertical directions.

Each of the resulting N Shifted High Resolution (SHR) images is then processed as follows: I) a 2D Multi-Wavelet Transform (MWT) is applied on the SHR-image, de-correlating the SHR-image into its multi-wavelet subbands; II) the coefficients in the resulting multi-wavelet high frequency subbands are set to zero; III) a 2D Inverse Multi-Wavelet Transform (IMWT) is used to generate a Primary Spatially Shifted Super Resolution (PSSSR) image from the resulting coefficients; IV) the PSSSR-image is spatially inverse shifted, bringing back the image coefficients onto its original positions.

The output super-resolution image is created by averaging the resulting N-shift compensated PSSSR-images.

# 4 Experimental results

To generate experimental results, four standard test images called: Lena, Elaine, Baboon and Peppers are taken. Each of these four images, is first lowpass filtered using a 2D Blackman filter, and then down sampled by a factor of 2 in both horizontal and vertical directions, generating a replica low resolution image for each of them. The Blackman 2D FIR filter coefficients are tabulated in Table 1.

| | | |
|---|---|---|
| 0.0381 | 0.1051 | 0.0381 |
| 0.1051 | 0.4273 | 0.1051 |
| 0.0381 | 0.1051 | 0.0381 |

Table 1: The Blackman 2D FIR filter coefficients.

To evaluate the performance of the proposed method, the resulting low resolution Lena, Elaine, Baboon and Peppers test images are enlarged using the proposed method, Nearest-neighbourhood, Bilinear, Bicubic, Sinc, Cycle Spinning (CS) [8], Directional Cycle Spinning (DSC) [9], and Yinji [4] methods. In this research, ghm multi-wavelet was used to generate the experimental results. The Peak Signal to Noise Ratio (PSNR) measurements for the enlarged images were calculated and tabulated in Table 2.

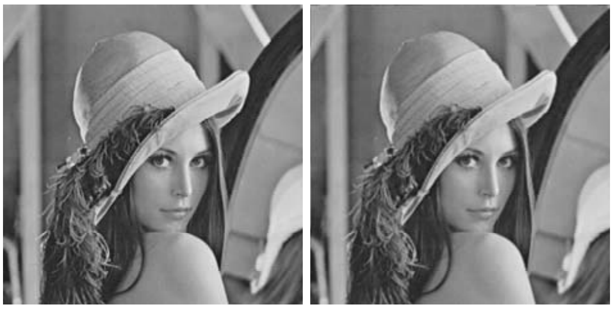
| | Lena | Elaine | Pepper | Baboon |
|---|---|---|---|---|
| Nearest | 29.15 | 30.15 | 29.20 | 25.30 |
| Bilinear | 29.90 | 30.45 | 29.69 | 26.28 |
| Bicubic | 30.25 | 30.59 | 29.90 | 26.91 |
| Sinc | 34.49 | 32.89 | 33.26 | 31.55 |
| Yinji | 33.75 | 32.50 | 32.79 | 30.17 |
| CS | 34.31 | 32.92 | 33.22 | 31.16 |
| DCS | 34.77 | 33.19 | 33.54 | 32.16 |
| Proposed | 34.96 | 33.35 | 33.63 | 32.55 |

Table 2: The PSNR comparison of different methods.

From table 2, it can be seen that the proposed method outperforms the competing methods when enlarging the images that contain significant high frequency information. The proposed technique produces almost the same objective quality when enlarging images containing less high frequency components.

To give a visual perception of the resulting enlarged images, original Lena test image and its enlarged image using the proposed technique are shown in Figure 3. From Figure 3, it can be seen that the enlarged image is almost the same as the original image and it is hard to differentiate between the original and the enlarged image.

To facilitate the assessment of different enlargement techniques, the enlarged image using the proposed method, Nearest neighbourhood, Bilinear, Bicubic, Sinc, Cycle Spinning (CS), Directional Cycle Spinning (DSC), and Yinji methods are subtracted from the ground truth image and magnified by a factor of 10 (shown in Figure 4). From these figures it is obvious that the proposed technique exhibits the lowest residuals compared to the other techniques.



(a) (b)

Figure 3: a) Original Lena test image; b) enlarged Lena test image using the proposed technique.

## 4 Conclusions

In this paper, a new image resolution enhancement technique was proposed. The proposed technique adopts the cycle-spinning scheme into multi-wavelet domain, increasing the quality of the enlarged images. Experiments on natural test images showed the proposed technique outperforms the state of the art techniques especially when the enlarged images contain high frequency information.

## References

[1] W.K. Carey, D.B. Chuang and S.S. Hemami, "Regularity preserving Image Interpolation, " *IEEE Trans. on Image Process*, **volume 8**, pp.1295-1297, (Sep. 1999).

[2] K. Kinebuchi, D. D.Muresan, T. W. Parks, "Image interpolation using wavelet-based hidden Markov trees," in *Proc. ICASSP 2001*, pp. 1957-1960, (May 2001).

[3] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, (1999).

[4] Y. Piao , L. H. Shin and H. W. Park, "Image Resolution Enhancement using Inter-Subband Correlation in Wavelet Domain," in *Proc. IEEE Int. conf. on Image processing*, pp. 445-448, (September 2007).

[5] V. Strela and A.T. Walden, "Signal and image denoising via wavelet thresholding: orthogonal and biorthogonal, scalar and multiple wavelet transforms," *In Nonlinear and Nonstationary Signal Process.*, pp. 124-157, (1998).

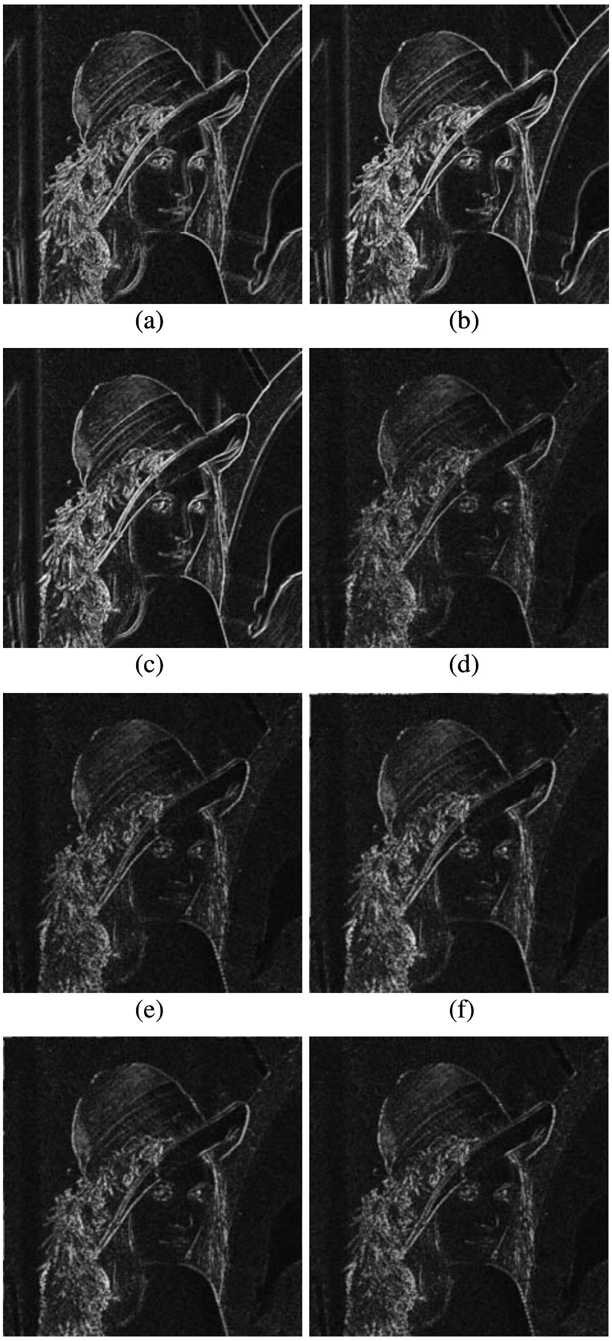
[6] V. Strela, "Multiwavelets: theory and applications," PhD thesis, MIT, (1996).

[7] A. Temizel and T. Vlachos, "Wavelet Domain Image Resolution Enhancement using Cycle-spinning," *Electronics Letters*, **volume 41,** No.3, (Feb. 2005).

[8] A. Temizel and T. Vlachos, "Image resolution up-scaling in the wavelet domain using directional cycle spinning," *Jour. of Electronic Imaging*, **volume 14(4),** (Dec. 2005).

[9] A. Temizel and T. Vlachos, "Wavelet Domain Image Resolution Enhancement," IEE *Proc. Vis. Image Signal Processing*, **volume 153,** No.1, pp. 25-30, (Feb. 2006).

[10] D.H. Woo, I.K. Eom and Y.S. Kim, "Image Interpolation Based on Inter-scale Dependency in Wavelet Domain," in *Proc. IEEE Int. conf. on Image processing,* pp. 1687-1690, (October 2004,).



(a) (b)

(c) (d)

(e) (f)

(g) (h)

Figure 4: Difference images magnified by a factor of ten, (a) Nearest neighbourhood, (b) Bilinear, (c) Bicubic, (d) Sinc, (e) Cycle Spinning (CS) [8], (f) Directional Cycle Spinning (DSC) [9], (g) Yinji [4] and (h) the proposed methods.

# Configurations of Binary Distillation column for optimal control

**Abdelmadjid KHELASSI, Riad BENDIB, Abdelhai BENHALLA**
University M'Hamed Bougara of Boumerdes,
Faculty of Hydrocarbons and Chemistry, Laboratory of Applied Automatic
E-mail :madjidk@hotmail.com

*Abstract-* **Choice of the control structure is a major concern and considerable research activity has been devoted to finding the best control configuration. In this work eight distillation control structures have been investigated by simulation based on models that are deduced using a software package for distillation column. These programs use transformations between the structures. The Dynamic Relative Magnitude Criterion, DRMC has been used to evaluate the interactions and disturbances propagation between the two interacting controlled loops of distillation column. Comparison between the results given by DRMC and those given by the Relative Gain Array, RGA has been also given in this work.**

***Keywords: Configuration;, Distillation; Interaction; DRMC; RGA; Optimal control.***

## I. INTRODUCTION

Recently there has been considerable interest in developing process control strategies for multivariable control system, that is, problems where several process variables are to be controlled and several variables can be manipulated [1]. Although a variety of advanced control techniques are available, only a few industrial applications have been reported. The conventional industrial approach to multivariable control problems is to use a "multiloop control system" consisting of *m* conventional PI or PID controllers where *m* is the number of controlled variables [2].

A big part of monovariable control problems can be solved by the classic PID-type compensator and fuzzy controllers. A number of investigators have developed measures of interaction that allow a control system designer to determine the proper input-output pairing for a set of single-input/single-output (SISO) controllers. There are many cases , however, where the measures of interaction show the absence of an adequate control configuration.

The goal of this paper is to give insight into the different structures of controlling binary distillation column and give the best scheme that allows us to minimise the energy consumed by this column

The configurations considered here are the Energy balance "LV-configuration" [3], the material balance " DV-configuration" [3], the ratio configuration such as (L/D,V/B) (L/D, V/F), (D/ (L+D), V), (L/D, V), (D/ (L+D), V/B), and LB [4], [5], [6].

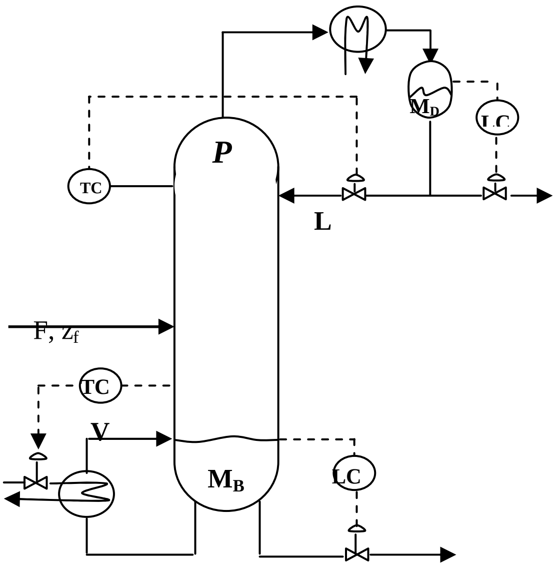


**Fig (1)** LV configuration

## II. DISTILLATION CONTROL STRUCTURES

### *A. Energy balance structure (L V)*

The energy balance structure can be considered to be the standard control configuration for dual composition control of distillation. In this control structure, the reflux flow rate *L* and the boil-up manipulator *V* are used to control the outputs concentrations or temperatures.

A typical distillation column with LV configuration is shown in Fig (1).

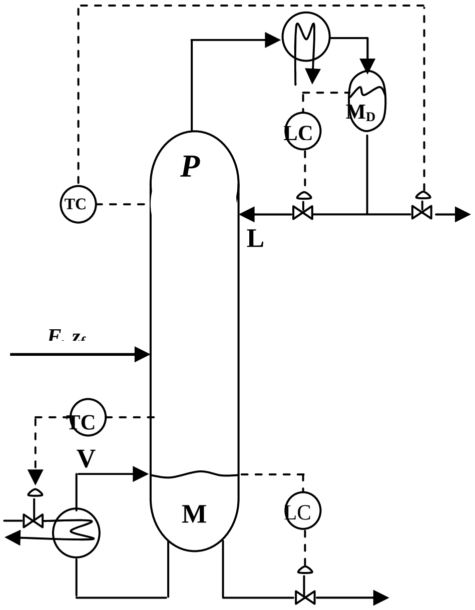


**Fig (2)** DV configuration

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

### *B. Material balance structures (D,V) and (L,B)*

Two other frequently used control structures are the material balance structures (D, V) and (L, B)**.** In the (D, V) structure, D and V are used as primary manipulators whereas L and B usually are used as inventory control manipulators. The implementation of this control scheme is shown in Fig (2).

### *C. The ratios scheme* $\left({}^{L}\!/_{D}, {}^{V}\!/_{B}\right)$

Ratio control configurations have been used in industry for at least forty years [7]. Condenser level is adjusted with both L and D such that their ratio is constant and reboiler level with both V and B such that their ratio is constant. The simplest justification for using ratios as inputs follows from steady state considerations: to keep the compositions constant, the ratio $L/V$ inside the column (slope of the operating line on the McCabe- Thiel diagram [8],[9]) should be constant.

### *D. The ratios scheme* $(L/D, V/F)$

The $(L/D, V/F)$ is an example of a control structure where a primary manipulator include a measurable disturbance. The inclusion of $F$ in the manipulator $V/F$ means that there is such a built-in feed forward from measured disturbances in *F* that V is changed in the same proportion as *F*. This results in a better rejection of disturbances in *F* than is the case in other structures.

### *E. Ryskamp's control scheme*

This control structure is suggested by Ryskamp (10), where the primary outputs are Controlled by $D/(L+D)$ and $V$. This scheme holds the reflux ratio constant if the top composition controller is constant. An increase of heat input from the bottom composition controller does not make top product as impure as would occur with reflux constant (conventional control) nor as over-pure as would occur with distillation flow constant (material balance control).Thus, this property of the scheme is sometimes said to result in "implicit decoupling", in contrast to "explicit decoupling" accomplished by external decoupling elements.

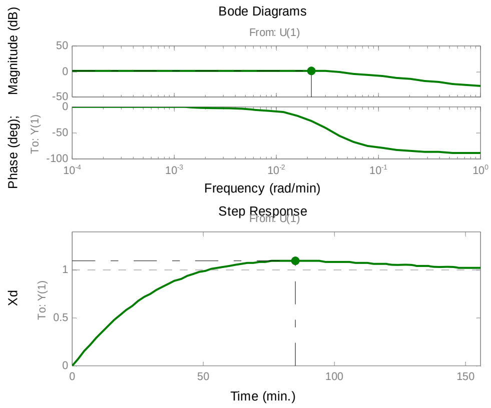


**Fig (3-a)** The closed loop bode diagram and step response 1st loop (LV scheme)

### *F. (L/D, V) structure*

This control structure is equivalent to Ryskamp's structure, because $\Delta\left(D/(L+D)\right) = -(\overline{D}/(\overline{L}+\overline{D}))^2\Delta(L/D)$ [7]; the only difference is a different scaling of gains associated with the top primary manipulator.

### *G. (D/(L+D),V/B) structure*

This control scheme is an extension of Ryskamp's structure, is suggested and studied by Takamatsu, Hashimoto, and Hashimoto (11).

## III. ANALYSIS OF INTERACTIONS IN THE LISTED STRUCTURES USING DRMC

### *A. Conventional control structure (LV) Configuration*

To specify parameters for the PI controllers root locus method has been used. In our case the design objective is to get a closed loop response where the percent overshot is small or equal to 07% . The closed loop step responses and bode diagrams for both loops are shown in Fig(3),

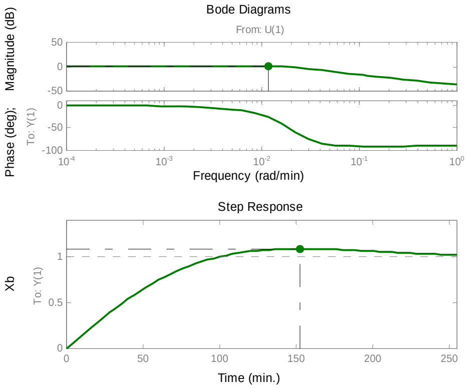
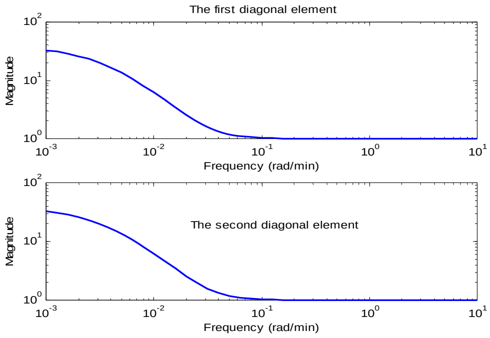


**Fig (3-b)** The closed loop bode diagram and step response 2nd loop (LV scheme)

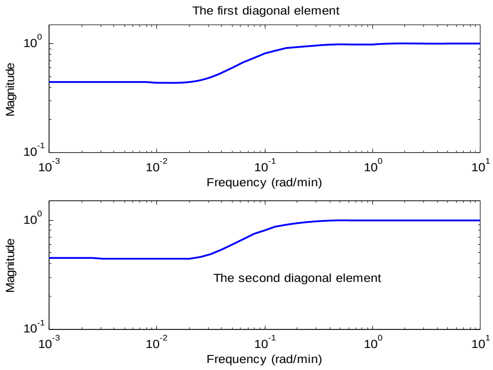
As it is shown in Fig (4) the magnitude of the diagonal elements for the range where the system works (i.e. the resonant frequency) are far from unity $\delta_{11}$ *and* $\delta_{22} \approx 7$ which means that strong interactions exist between the loops, the fact that let the (LV) configuration to be not recommended for two point control ( i.e. where all loops are in automatic). The off-diagonal elements $\delta_{12} = 0.8$ *and* $\delta_{21} \approx 1$ in the resonant frequencies which indicate that there exist large disturbances between the two loops and propagate approximately by the same magnitude.



**Fig(4)** The DRMC diagonal elements

*B. Material balance (D,V) structure*

The DRMC values, the diagonal elements

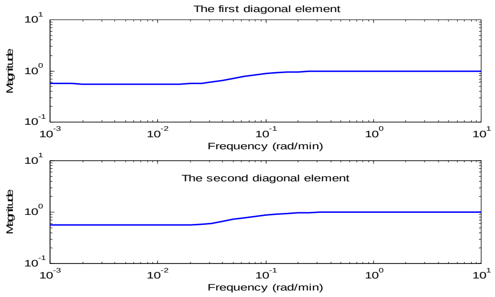


Fig(5) The DRMC elements of DV configuration

The distillation column under this configuration is slightly interactive as it is indicated by the DRMC diagonal elements shown in Fig (5) ( $\delta_{11}$ and $\delta_{22} = 0.45$ are close to unity). For the off-diagonal elements we notice that $\delta_{21} = 0.1, \delta_{12} = 0.09$ i.e. $\delta_{21} > \delta_{12}$ which means that there exist disturbances that propagate from the top loop to the bottom loop. In general this configuration is a good choice in the sense of interactions but in reality there are problems that affect the operation of the column under this control scheme [7],[12], the major problem that may rise with this configuration is the effect of level control, such that with fast condenser level control, the increase in boil-up goes up the column, then returned back as a reflux through the action of the condenser level controller (since D is constant), and we have an increase in the internal flows only.

*C. Material balance (L,B) structure*

The DRMC elements, the diagonal elements

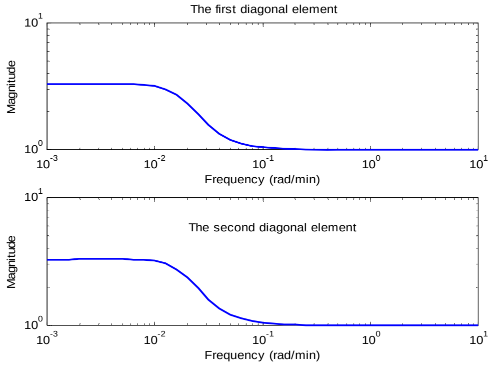


Fig(6) The DRMC diagonal elements
LB configuration

The distillation column under this configuration is almost non-interactive as it is indicated by the DRMC diagonal elements shown in Fig (6) ( $\delta_{11}$ and $\delta_{22} = 0.6$ are close to unity). For the off-diagonal elements $\delta_{21} \approx \delta_{12} = 0.1$ which means that there are disturbances that propagate between the loops by the same magnitude. As the first material balance configuration there is a problem of level control, another practical problem is that this configuration is extremely sensitive to disturbances in feed flow rate [7].

*D. The two ratios scheme* $\left(L/_D, V/_B\right)$

The DRMC values, the diagonal elements



Fig(7) The DRMC diagonal elements (L//D V/B) configuration

The distillation column under this configuration is interactive as it is indicated by the diagonal elements of the DRMC shown in Fig (7) ( $\delta_{11}$ and $\delta_{22} = 3.2$ ), but compared with the energy balance configuration the degree of interactions is smaller (this also can be deduced using the RGA). The examination of the off-diagonal elements show that $\left(\delta_{21} = 0.6\right)$ is greater than $\left(\delta_{12} = 0.35\right)$ which indicates that there is disturbance propagation from the top loop to the bottom loop. The main disadvantages of this configuration is the need for measurements of all flows L, D, B and V which makes it more failure sensitive and more difficult to implement.

### *E. The two ratios scheme* $\left(L/_{D}, V/_{F}\right)$

The (L/D,V/F) model is deduced from the LV steady state model.

**The DRMC values, The diagonal elements**

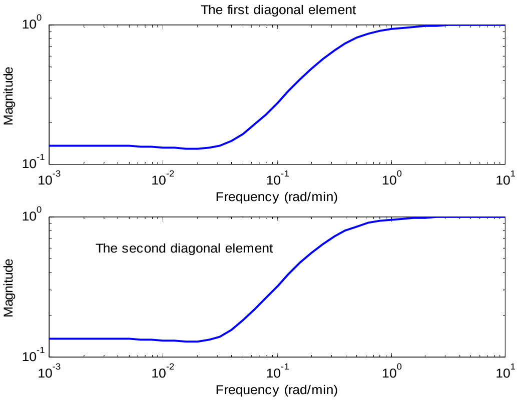
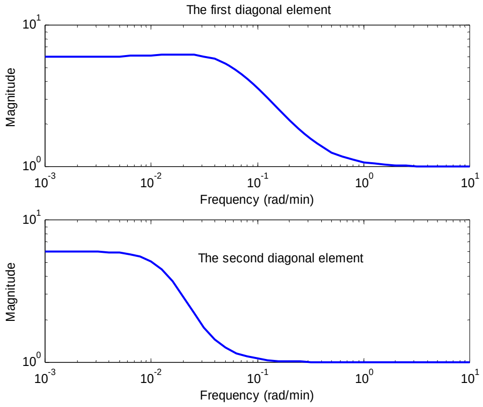


Fig (8) The DRMC diagonal elements(L//D V/F) configuration

The examination of DRMC values (at the resonant frequencies) shows that ($\delta_{11}$ and $\delta_{22} \approx 0.15$) which indicates that more interaction compared with the previous two ratios configuration is expected, for the off-diagonal elements $(\delta_{21} = 0.2)$ is greater than $\left(\delta_{12} \approx 1.6\text{x}10^{-2}\right)$ which indicates that there exists a disturbance propagation from the top loop to the bottom loop, in general we can say that this configuration behaves between the (DV) and (L/D,V/B) , but in this configuration we avoid the disturbances created by the flow rate ( because F is included in the primary manipulator). The major drawback for this control scheme is the need of measurements for all flows which makes it more difficult to implement.

### *F. Ryskamp's scheme* $\left(D/(L+D), V\right)$

The DRMC values, the diagonal elements



**Fig(9)** The DRMC diagonal elements Ryskamp's configuration

The diagonal elements at the resonant frequencies shown in Fig (9) are far from unity $\delta_{11}$ and $\delta_{22} = 5.5$, but there values are small compared with those of LV control scheme, which indicates that the degree of interactions exist in this configuration is small compared with conventional control, this is according to decoupling (implicit) effect which results from the property that the scheme holds the reflux ratio constant if the top composition controller is constant. The DRMC off-diagonal elements that indicate that there is a large disturbance propagation from the bottom loop to the top loop (since$(\delta_{12} = 0.6) > (\delta_{21} = 0.1)$).

### *G. The (L/D,V) control scheme*

The properties of $(L/D,V)$ structure are similar to those of the $(L/D,V/F)$ structure, particularly in our case where the feed rate is $F = 1$. The $(L/D,V)$ model is deduced from the LV steady state model.

### *H. The (D/(L+D),V/B) control structure*

The DRMC values, the diagonal elements

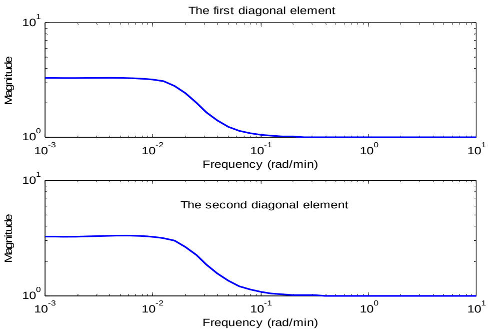


**Fig (10)** The DRMC diagonal elements (D/(L+D),V/B) configuration

Fig (10) presents DRMC diagonal elements for this structure, the examination of the diagonal elements shows that there exist interactions between the two loops, but with a magnitude smaller than those of Ryskamp's scheme $\left(\delta_{11}, \delta_{22} = 3.2\right)$ (this is what Shinskey (1984) and Hashimoto have shown [7]), and disturbances propagate from the top loop to the bottom loop as it is given by the off-diagonal elements $\left(\delta_{21} = 0.5\right) > \left(\delta_{12} = 0.3\right)$

## IV. ANALYSIS OF INTERACTIONS USING RGA

TABLE I
COMPARISON BETWEEN THE RGA VALUES AND DRMC VALUES FOR THE STUDIED STRUCTURES

| **Structure** | $\lambda_{11}$ | $\delta_{11}$ |
|---|---|---|
| **LV** | **36.06** | **7** |
| **DV** | **0.446** | **0.45** |
| **LB** | **0.56** | **0.6** |
| **(L/D V/B)** | **3.28** | **3.2** |
| **(L/D V/F)** | **5.98** | **0.15** |
| **Ryskamp's** | **5.98** | **5.5** |
| **(L/D, V)** | **5.98** | **0.15** |
| **(D/(L+D) V/B)** | **3.28** | **3.2** |

As it is shown in Table I, there are deviations between the values given by the RGA and those given by the DRMC, since the first method gives information about the steady state behavior of the system, whereas the DRMC deals with the dynamic behavior of the system.

## CONCLUSION

In order to overcome the problem of interactions several distillation control schemes are suggested by many contributors. DRMC is used to assess interactions in eight well known control configurations for binary distillation columns. The mathematical model for each configuration is deduced by using transformations between control schemes taking the (LV) configuration as a base model. It has been shown from the analysis that the conventional control structure (LV) suffers from severe interactions the diagonal elements are far from unity and the off diagonal elements show that there exist large disturbances that propagate between the loops approximately by the same magnitude. The material balance control (DV) is non-interactive the diagonal elements are close to unity, the off diagonal elements show that there exist disturbances that propagate from the top loop to the bottom loop since $\delta_{12} > \delta_{21}$, the same remark with the second material balance (LB) the diagonal elements are close to unity which indicates that for this configuration low interaction is expected, the off diagonal elements show that there exist disturbances that propagate between the loops by the same magnitude.

The ratio schemes are introduced to reduce the degree of interactions exist in the conventional control scheme (LV) and that is what it has been shown using DRMC for the remaining control schemes. For the two ratios control scheme $\left(L/D, V/B\right), \left(L/D, V/F\right)$ and $\left(L/D, V\right)$ are non-interactive with respect to (LV) control (as it is indicated by the DRMC diagonal elements for the three cases). Ryskamp's configuration is an interactive one but with smaller degree than that of (LV) control scheme, the larger degree of interaction in Ryskamp's has been reduced using its extension control scheme i.e. $(D/(L+D), V/B)$.

## REFERENCES

[1] Agamennoni.O.E and Figueroa.J.L "Advanced Controller Design For Distillation Column" INT.J. Control, Vol.59,N°3,817-839,1994.

[2] Atle. C and Lien. C " Complex Distillation Arrangements "ESCAP, Trondheim 1997.

[3] R. Bendib "Control configurations and assessment of interactions in distillation columns" Master thesis, FHC, Boumerdes University, Algeria, 2004.

[4] Dadhe. K, Engell.S and Gesthuisen.R "Control Structure Selection For a Reactive Distillation Column" IFAC 2002.

[5] Esref .E "Consistency Relations In Process Modeling" IFAC 2002.

[6] Ioannis k. kooks, J.D. Perkins "An Algorithmic Method For The Selection Of Multivariable Process Control Structures" J.O.Process Control, 85-89, 2002

[7] Luyben. W "Practical Distillation Control ",Nostrand Reinhold 1992.

[8] King "Separation Process" Second edition Mc Graw Hill 1980

[9] Luyben.W "Process Modeling, Simulation And Control For Chemical Engineers" 2nd edition, Mc Graw Hill 1990.

[10] Seborg Dalee " Process Dynamics And Control " John Wiley and sons 1989.

[11] Takamatsu.T and Hashimoto. I " Non-Interacting Control System Design Of A Distillation Column By The Partial Model Matching " IFAC 1986

[12] Skogestad.S " Dynamics and control of distillation columns" Trans.IChemE,Vol.75,part A,1997.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Finding Risk Dominance Strategy in Imperfect Game, A Theory Perspective

Seyyed Mohammad Reza Farshchi[1], Irene Kehvazadeh[2]
1,2- Department of Data Mining Advanced Research Center, Mashhad Branch,
Iran. Email: Shiveex@Gmail.Com

**ABSTRACT:**
Game theoretic reasoning pervades economic theory and is used widely in other social and behavioral sciences. It is widely used in decision of economic behavior and other game conditions with risk characters. A model is an abstraction we use to understand our observations and experiences. This paper proposes an imperfect information multi-player system using risk dominance strategy for its decision module. In this article, the Influence of the strategy in different conditions is discussed. A model is unlikely to help us understand a phenomenon if its assumptions are wildly at odds with our observations. The cases that risk dominance worked and the process of the module to exclude risk is illustrated. In the end, basing on the results of experiments, the performance of the strategy is discussed.

**KEYWORDS:** Game theory, Imperfect game, Decision tree.

## 1. INTRODUCTION

Risk dominance and payoff dominance are two related refinements of the Nash Equilibrium (NE) solution concept in game theory, defined by John Harsanyi and Reinhard Selten[2]. Nash equilibrium is considered payoff dominant if it is Pareto superior to all other Nash equilibrium in the game. When faced with a choice among equilibrium, all players would agree on the payoff dominant equilibrium since it offers at least as much payoff as each player's best alternative. Conversely, Nash equilibrium is considered risk dominant if it has the largest basin of attraction [3], meaning the more uncertainty players have about the actions of the other player(s), the more likely they will choose the risk dominant strategy [4].

A game with imperfect information means the players face risks not only come from the uncertain strategies of other players but also from the uncertain game conditions [5]. Thus, risk dominance strategy is imported as the decision module of an imperfect information game called Military Chess.

The main contribution of the system that is presented is the realization of a decision module basing on risk dominance strategy which is used for the selection of optimal step in the Military game. The performance of the strategy is also discussed in the following experiments.

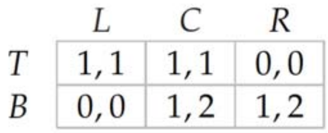
This paper is organized as follows. Section 2 briefly introduces the related theories and works about the risk dominance strategy in game theory. Section 3 gives a primary introduction of Military Chess game. And then, in Section 4, a specific description is presented about the risk dominance strategy and the realization details of the decision module in our game system. Section 5 analyzes the performance of the new system that is tested locally and on the Internet. And finally, section 6 contains our conclusion.

## 2. ITERATED ELIMINATION OF WEAKLY DOMINATED ACTIONS

A strictly dominated action is clearly unattractive to a rational player. Now consider an action a is weakly dominated in the sense that there is another action that yields at least as high a payoff as does $a_i$ whatever the other players choose and yields a higher payoff than does $a_i$ for some choice of the other players. In the game in Figure 1, for example, the action T of player 1 weakly (though not strictly) dominates B.

A weakly dominated action that is not strictly dominated, unlike a strictly dominated one, is not an unambiguously poor choice: by Lemma 356.1 such an action is a best response to some belief. For example, in the game in Figure 1, if player 1 is sure that player 2 will choose R then B is an optimal choice for her. However, the rationale for choosing a weakly dominated action is very weak: there is no advantage to a player’s choosing a weakly dominated action, whatever her belief. For example, if player 1 in the game in Figure 1 has the slightest suspicion that player 2 might choose L then T is better than B, and even if player 2 chooses R, T is no worse than B.

If we argue that it is unreasonable for a player to choose a weakly dominated action then we can argue also that each player should work under the assumption that her opponents will not choose weakly dominated actions, and they will assume that she does not do so, and so on. Thus, as in the case of strictly dominated actions, we can argue that weakly dominated actions should be removed iteratively from the game. That is, first we should mark actions of player 1 that are weakly dominated; then, without removing these actions of player 1, mark actions of player 2 that are weakly dominated, and proceed similarly with the other players. Then we should remove all the marked actions, and again mark weakly dominated actions for every player. Once again, having marked weakly dominated actions for every player, we should remove all the actions and go through the process again. We should repeat the process until no more actions can be eliminated for any player. This procedure, however, is less compelling than the iterative removal of strictly dominated actions since the set of actions that survive may depend on whether we remove all the weakly dominated actions at each round, or only some of them, as the two-player game in Figure 1 shows. The sequence in which we first eliminate L (weakly dominated by C) and then T (weakly dominated by B) leads to an outcome in which player 1 chooses B and the payoff profile is (1, 2). On the other hand, the sequence in which we first eliminate R (weakly dominated by C) and then B (weakly dominated by T) leads to an outcome in which player 1 chooses T and the payoff profile is (1, 1).



|  | *L* | *C* | *R* |
|---|---|---|---|
| *T* | 1,1 | 1,1 | 0,0 |
| *B* | 0,0 | 1,2 | 1,2 |

**Figure 1.** A two-player game in which the set of actions that survive iterated elimination of weakly dominated actions depends on the order in which actions are eliminated.

In the games studied in this chapter, the players are representatives from an evolving population of organisms (humans, animals, plants, bacteria, ... ). Each player's payoffs measure the increments in the player's biological fitness, or reproductive success (e.g. expected number of healthy offspring), associated with the possible outcomes, rather than indicating the player's subjective feelings about the outcomes.

Each player's actions are modes of behavior that the player is programmed to follow.

The players do not make conscious choices. Rather, each player's mode of behavior comes from one of two sources: with high probability it is inherited from the player's parent (or parents), and with low (but positive) probability it is assigned to the player as the result of a mutation. For most of the models in this article, inheritance is conceived very simply: each player has a single parent, and, unless it is a mutant, simply takes the same action as does its parent. This model of inheritance captures the essential features of both genetic inheritance and social inheritance: players either follow the programs encoded in their genes, which come from their parents, or learn how to behave by imitating their parents. The distinction between genetic and social evolution may be significant if we wish to change society, but is insignificant for most of the models considered in this article.

Members of a single large population of organisms are repeatedly randomly matched in pairs. The set of possible modes of behavior of each member of any pair is the same, and the consequence of an interaction for an organism depends only on the actions of the organism and its opponent, not on its name. As an example, think of a population of identical animals, pairs of which periodically are engaged in conflicts (over prey, for example). The actions available to each animal may correspond to various degrees of aggression, and the outcome for each animal depends only on its degree of aggression and that of its opponent. Each organism produces offspring (reproduction is asexual), to each of whom, with high probability, it passes on its mode of behavior; with low probability, each offspring is a mutant that adopts some other mode of behavior.

Harsanyi and Selten's risk dominance is based on what they refer to as a tracing procedure [2], the details of which are beyond the scope of this paper. As an alternative, Selten proposes a measure of risk dominance that is easy to calculate for our games [6], and is indicative of the outcome of the tracing procedure. Let $u_1(X, X)$ be the payoff of player1 with strategy pair *(X, X)*, Selten's [6] weighted average log measure of risk dominance of the equilibrium *(A, A)* over *(B, B)* is given by:

$$R = \log\left[\frac{u_1(A,A)-u_1(B,A)}{u_1(B,B)-u_1(A,B)}\right] \tag{1}$$

If *R* is positive, Harsanyi and Selten's tracing procedure selects *(A, A)* as risk dominance. If *R* is zero, the mixed strategy Nash equilibrium is risk dominant. If *R* is negative, *(B, B)* is risk dominant. Notice that any affine transformation of the payoffs in the game would not change either the sign or the magnitude of *R*.

This criterion can explain the experimental result of Paul Straub. In Game 1, *R* of player 1 can be calculated by (1) that equal to Log (2.5), so *(A, A)* is both payoff and risk dominance. In Game2, *R* of player 1 changes to Log (0.67) which is a negative value, so the strategy *B* has risk dominance. That is why the players picked strategy *B* in the final three rounds of the session utilizing Game 2.

There are two main factors that affect the risk characteristics of the games [7]. One is the trust level between the players as showing in table 2. Generally speaking, the coordinate game will keep a comparative higher trust level which weakens the influence of risk dominance. However, in uncoordinated games and zero sum games, there is almost no trust between players. In another side, the imperfect information game situation will greatly enhance the uncertainty of the game, thus the effect of risk dominance strategy should be carefully considered in decision process.

The notion of Nash equilibrium requires only that each player's strategy be optimal in the whole game, given the other players' strategies; after histories that do not occur if the players follow their strategies, the actions specified by a player's Nash equilibrium strategy may not be optimal. In some cases we can think of the actions prescribed by a strategy for histories that will not occur if the players follow their strategies as "threats"; the notion of Nash equilibrium does not require that it be optimal for a player to carry out these threats if called upon to do so. In the previous chapter we studied the notion of subgame perfect equilibrium, which does impose such a requirement: a strategy profile is a subgame perfect equilibrium if every player's strategy is optimal not only in the whole game, but after every history (including histories that do not occur if the players adhere to their strategies).

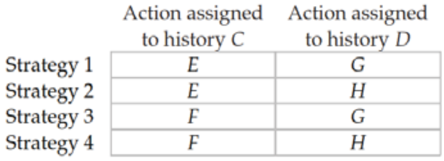
**Table 1.** Payoff and risk dominance in two games condition

| Game1 | Player2 | | | | |
|---|---|---|---|---|---|
| Player 1 | | A | | B | |
| | A | | 8 | | 3 |
| | | 8 | | 4 | |
| | B | | 4 | | 6 |
| | | 3 | | 6 | |
| Game2 | Player 2 | | | | |
| Player 1 | | A | | B | |
| | A | | 8 | | 4 |
| | | 8 | | 0 | |
| | B | | 0 | | 6 |
| | | 4 | | 6 | |

**Table 2.** Games with low trust level, risk dominance will be very important in decision process

| Game | Player2 | | | | |
|---|---|---|---|---|---|
| Player1 | | A(20%) | | B(80%) | |
| | A | | 8 | | 4 |
| | | 8 | | 0 | |
| | B | | 0 | | 6 |
| | | 4 | | 6 | |

**Table 3.** The four strategies of player 2 in the game in Figure 1.



| | Action assigned to history *C* | Action assigned to history *D* |
|---|---|---|
| Strategy 1 | *E* | *G* |
| Strategy 2 | *E* | *H* |
| Strategy 3 | *F* | *G* |
| Strategy 4 | *F* | *H* |

## 3. MILITARY CHESS GAME SYSTEM WITH MONTE-CARLO SAMPLING

In this section, the imperfect information game system basing on a game called Military Chess is presented.

### 3.1. Introduction for Proposed Game

The Military Chess, which is also called Kriegspiel, has been a very popular game for many years. The rules of the Military Chess are quite complex. Simply speaking, the basic process of the game is to move pieces, attack opponent's pieces and finally occupy the position of enemy's flag. Maybe it sounds quite like Chess but there are at least three main differences between them as following.

Opposite and two opponents sits aside, like Bridge. Second, the pieces which have a certain military rank can only destroy those with lower ranks. Last but most important, player cannot see the ranks of other player's pieces, even though it belongs to his associate. As the result, players have to guess the ranks of other pieces. This is the most difficult and complex point of the development of Military Chess game system. A Military Chess game system is developed which is called mygames [11].

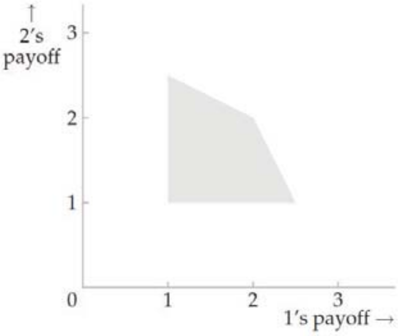


**Figure 2.** The approximate set of Nash equilibrium discounted average payoffs for the infinitely repeated Prisoner's Dilemma with one-shot payoffs to 1.

As in the case of a finite horizon game, if a strategy profile is a subgame perfect equilibrium then certainly it satisfies the one-deviation property, since no player must be able to increase her payoff by any change in her strategy. What we need to show is the converse: if a strategy profile is not a subgame perfect equilibrium then there is some subgame in which the first-mover can increase her payoff by changing only her initial action.

### 3.2. Monte-Carlo Sampling

In imperfect information game, none of the players in the game have the whole knowledge about the current game state. In order to solve such problem in practice, one has to either resort to special cases, or fall back heuristics which is created by Monte-Carlo sampling [5].

Monte-Carlo sampling is an algorithm that deals with imperfect information games. When evaluating the choices of one player, the Monte-Carlo method randomly guesses the unseen pieces of the other players. The game is then searched using min-max game tree as if these pieces are known to each player. As demonstrated in figure 2, the square represents the max node and the circle represents the min node. Only the root node is the random (the diamond) and all the other nodes are board states with perfect information.

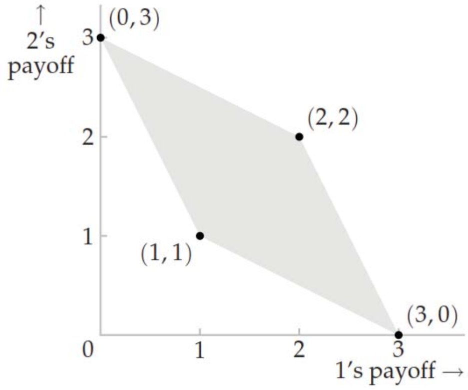


**Figure 3.** Game tree of imperfect information using Monte-Carlo sampling

### 3.3. Realization with Monte-Carlo Sampling

The set of Nash equilibria consists of the action profile (Hare,...,Hare) in which all hunters catch hares, and any action profile in which exactly k hunters pursue the stag and the remaining hunters catch hares.

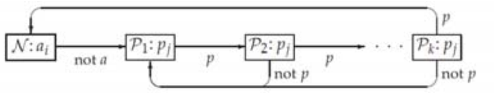


**Figure 4.** A subgame perfect equilibrium strategy for player i in a two-player infinitely repeated game. The outcome p is that in which each player's action is one that holds the other player down to her minmax payoff.

In the process of game, when it is the agent's turn to move, the current state that agent faces is an imperfect information state. Monte-Carlo sampling process transforms the imperfect information state to perfect information state. Next, basing on the sampled perfect information, move generator module generates moves list and build the game tree. And then, tree search algorithm is used searching the optimal move from the game tree. When a move has been chosen, the system check the end condition, which could be set for a limit of running time or a limit times for system's iteration. This process will iterate many times which is set by users. Each iteration generates a candidate move. When the end condition is fulfilled, system will collect all the candidate moves in before iterations and decision module will take an analysis about them and finally choose the optimal move.

## 4. DECISION MODULE IN THE SYSTEM WITH RISK DOMINANCE STRATEGY

The task of decision module is to select one optimal move from tens of candidates. The criterion is the expect payoff of the moves. However, in the case of imperfect information, the uncertainty of the game condition leads the payoff for one move is uncertain. Decision module builds a suppositional game condition to deal with the problem. Just like the table 5 illustrates, player 1 denotes the player for whom the system make decision, and player 2 is set to denotes different Monte-Carlo sample results from the uncertain game circumstance which greatly influences the payoff of the player1's strategies.

For concisely express, set the numbers of candidate move and different conditions are both three. *E(X)* means the expect payoff of move $X$, $P_a$ means the appear probability of condition *a* and $V_{Ab}$ means the payoff when condition *b* appears and player1 adopts move *A*. Let *M* means the cluster of all conditions and the system calculates the *E(X)* as following:

$$E(X)=\sum_{i\in M} V_{Xi}P(i) \qquad (2)$$

And the finial optimal result will be decided by $\max\{E(A),E(B),E(C)\}$ .

The system calculates the *E* value with equation (2) of each move, and then chooses the largest to be the optimal strategy. The process of decision can deal with most of the conditions well except some special cases which is also the reason the risk dominance strategy be adopted.

There are some cases that terrible departure lost appears just like table 6 shows. When condition b happens, candidate move *B* will leads a terrible lost.

The low appearance probability of condition *b* will smooth the lost of *B* when calculates *E (B)*. Just like no one want to take the risk of low probability to eat candy in a bowl which is full of hundreds of candy and only one of them is poison, the task of risk dominance strategy in this case is to exclude *B*-like candidate moves from the final result.

The decision module creates the game condition like table 7 shows. Value *(A, a)* means the average payoff of other conditions and *(B, b)* means the payoff in the dangerous condition. Using equation (1) to calculate the value *R* is Log (5) which is a positive value and it can be concluded that the strategy *A* is risk dominance. That is the process that risk strategy *B* is excluded.

## REFERENCES

[1] Straub, P.G., “Risk dominance and coordination failures in static games”, Quart. Rev. Econ. Finance 35, 1995, pp.339–363.

[2] John C. Harsanyi and Reinhard Selten, “A General Theory of Equilibrium Selection in Games”, MIT Press, 1988, ISBN 0262081733.

[3] David Schmidt, Robert Shupp, James M. Walker, and Elinor Ostrom, “Playing safe in coordination games: the roles of risk dominance, payoff dominance, and history of play”, Games and Economic Behavior, 2003, vol. 42, issue 2, pp.281-299.

[4] Andreoni, J., Croson, R., “Partners versus strangers: random rematching in public goods experiments”. Mimeo. Forthcoming in: Plott, C.R., Smith, V.L. (Eds.), Handbook of Results in Experimental Economics, 1998.Forthcoming in: Plott, C.R., Smith, V.L. (Eds.), Handbook of Results in Experimental Economics, 1998.

[5] Howard James Bampton, “Solving imperfect information games using the Monte Carlo heuristic [Master thesis]”, University of Tennessee, Knoxville, 1994.

[6] Selten, R., “An axiomatic theory of a risk dominance measure for bipolar games with linear incentives”, Games Econ. Behav., 8, 1995, pp.213–263.

[7] LIU Zongqian and WANG Feng, “Risk Dominance in Multi-equilibrium Selection and Mixed Strategy”, Department of Mathematics, Capital Normal University, Beijing 100037, China, 2005.

[8] Xuan Wang, Zhaoyang Xu, and Xiao Ma, “TD(Λ) optimization of imperfect information games evaluation function”, Intelligence Computing Research Center Harbin Institute of Technology Shenzhen Graduate School, WCCGC, Japan, 2007.

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# Robust $H_\infty$ Control of Singular Systems over Networks with Data Packet Dropouts

Yongbo Lai, Guoping Lu
Department of Electrical Engineering
Jiangsu College of Information Technology
Wuxi, China
E-mail: yongbo100@sina.com, Lgp@jsit.edu.cn

*Abstract*—**The problems of stochastic stability and $H_\infty$ control for a class of discrete time singular networked control systems (SNCSs) with data packet dropouts and nonlinear perturbation are investigated in this paper. By modeling the sensor-to-controller and controller-to-actuator with random data packet dropouts as Markov chains, the closed-loop system can be expressed as a jump discrete singular system with four modes. A sufficient condition for the existence of a controller is established in terms of linear matrix inequalities (LMIs), the controller gain can be solvable via the cone complementary linearization method, and the designed controller guarantees the systems to be regular, causal and stochastically stable and satisfies $H_\infty$ performance. In addition, a numerical example is given to illustrate the effectiveness of the proposed approach.**

***Keywords-singular systems; networked control systems; data dropout; $H_\infty$ control ; Markov chain***

## I. INTRODUCTION

Networked control systems (NCSs) have been widely used in various industrial areas, with the rapid development of computer, network and communication technology. The primary advantages of NCSs are simplified system structure, lower cost of system integration, ease of diagnosis, remote distributed control, and increasing system agility. In an NCS, several important issues need to be treated, which include the network induced delay and data packet loss, more recently, the analysis and synthesis problem of networked systems with data dropout and network induced time delays have attracted many research interests [1-7]. Due to bandwidth limited communica -tion channels, packed loss and time delays are the most important and special two issues of NCSs. Generally speaking, there were three main approaches for modeling packed loss and time delays in the NCSs. The first one is to model the data loss and induced delays as a binary switches sequence which obeys the Bernoulli process with certain probability, the NCSs in both continuous time case and discrete time case with packet loss and delays were studied in [5,6]. The second approach is to model the process as discrete time Markov jumping systems, in which transmission times are varying within a time interval or driven by a stochastic process with Markov chain, a class of methods for stabilization analysis for an NCS were proposed [3,4]. The third model is to view data packet loss as a special time delay system, which deals with the stability and controllability [2]. However, it can only be used to treat the systems with sense-to-controller packet loss or controller-to-actuator packet loss cases.

In [15], the stabilization problem for a class of NCSs in the discrete time domain is studied by modeling the sensor-to controller packet loss and controller-to-actuator packet loss as Markov chains. It is worth noting that the mentioned literature results are still very limited for they are all concerned with a nonsingular controlled plant. In real projects, a certain linear NCS model provides just an approximate singular of the real facts. Recently a few results have been reported for the singular NCSs [7, 16].

In actual physical networked transmission field, there coexist data packed loss and nonlinear perturbation, however, the aforementioned articles do not consider the two problems simultaneously. The $H_\infty$ control problem for discrete time singular markov jump systems with data loss and nonlinear perturbation is also an important problem, and it is not simple extension to stability for the systems over networks. Motivated by recent research on singular perturbed systems and networked control systems [8-14, 18], this paper considers the problems of stochastic stability and $H_\infty$ control for a class of discrete time singular networked control systems with data packet dropouts and nonlinear perturbation. Firstly, we consider the case that modeling the sensor-to-controller and controller-to-actuator with random data packet dropouts as Markov chains, and the nonlinear perturbation satisfying Lipschitz condition, the closed-loop system can be expressed as a jump discrete singular system with four modes in Section II; Next, the state feedback controller design and solvable problem are proposed in section III and section IV; Finally, an example is given to illustrate the effectiveness of the proposed approach in section V.

*Notations:* Throughout this paper, for real symmetric matrices $X$ and $Y$, the notation $X \geq Y$ (respectively, $X > Y$) means that the matrix $X - Y$ is semi-positive definite(respectively, positive definite); $\mathbb{R}^n$ and $\mathbb{R}^{n\times m}$ denote the *n*-dimensional Euclidean space and the set of all *n*×*m* real matrices, respectively; $I$ is the identity matrix with appropriate dimension; the superscript $T$ represents the transpose of a matrix; $\|X\|$ refers to Euclidean norm of the vector $X$ ; $\mathbb{Z}$ denotes the set of non-negative integer numbers;

This work was supported by the National Natural Science Foundation of China under Grant No.61174066, also by the Jiangsu Overseas Research & Training Program for University Prominent Young & Middle aged Teachers & Presidents.

$\mathcal{E}\{\bullet\}$ denotes the mathematical expectation, and $*$ denotes the matrix entries implied by symmetry of a matrix.

## II. Problem Statements and Preliminaries

In this paper, considering a typical NCS model with data packet dropouts exist in the communication links from sensor to controller and controller to actuator as shown in Fig. 1.

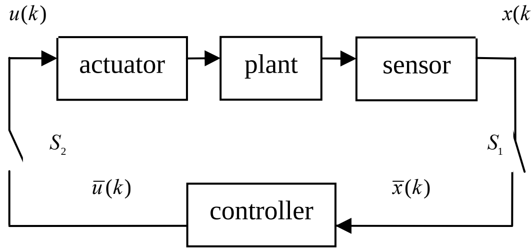


Fig.1 Framework of networked control system

The physical process to be controlled is the following singular discrete-time nonlinear model:

$$\begin{cases} Ex(k+1) = Ax(k) + Bu(k) + B_\omega \omega(k) + f(x,u,k) \\ z(k) = Cx(k) + Du(k) + D_\omega \omega(k) \end{cases} \tag{1}$$

Where $x(k) \in \mathbb{R}^n$ is the system state, the matrix $E \in \mathbb{R}^{n\times n}$ may be singular, we shall assume that $rank(E) = r \le n$, $u(k) \in \mathbb{R}^p$ is the control input, $\omega(k) \in \mathbb{R}^q$ is the disturbance input which belongs to $L_2[0,\infty)$, and $z(k) \in \mathbb{R}^m$ is the system controlled output; $f(\bullet) \in \mathbb{R}^n$ is nonlinear uncertain perturbation, with $f(0,0,k) = 0$ satisfies Lipschitz condition; $A, B, B_\omega, C, D, D_\omega$ are known real constant matrices with appropriate dimensions.

Assume the system state can be measured and the data are transmitted in a single packet at each time step, the data loss information is important for controller design, it is desirable that the state feedback controller is

$$\bar{u}(k) = K\bar{x}(k) \tag{2}$$

which $K$ is the controller gain to be determined.

In Fig.1, $S_1, S_2$ are networked switches, and $\alpha, \beta$ $[\alpha, \beta \in (0,1)]$ are the states, respectively, when $\alpha = \beta = 0$, there is no data packet loss, then $\bar{x}(k) = x(k)$, $\bar{u}(k) = u(k)$; when $\alpha = \beta = 1$ there exist data packet loss, then

$\bar{x}(k) = \bar{x}(k-1)$, $u(k) = u(k-1)$.

Then $\bar{x}(k)$ and $u(k)$ can be written as:

$$\begin{cases} \bar{x}(k) = (1-\alpha)x(k) + \alpha\bar{x}(k-1) \\ u(k) = (1-\beta)\bar{u}(k) + \beta u(k-1) \end{cases} \tag{3}$$

By introducing new state vectors $\xi(k) = [x^{\mathrm{T}}(k), \bar{x}^{\mathrm{T}}(k-1), u^{\mathrm{T}}(k-1)]^{\mathrm{T}}$, then the closed-loop system resulting from equations (1), (2) and (3) can be expressed as ($i \in (1,2,3,4)$)

$$\begin{cases} \bar{E}\xi(k+1) = A_i\xi(k) + \bar{B}_\omega\omega(k) + \bar{f}(k) \\ z(k) = C_i\xi(k) + D_\omega\omega(k) \end{cases} \tag{4}$$

Where

1) $\alpha = \beta = 0$, $A_1 = \begin{pmatrix} A+BK & 0 & 0 \\ I & 0 & 0 \\ K & 0 & 0 \end{pmatrix}, C_1 = [C+DK, 0, 0]$

2) $\alpha = 0, \beta = 1$, $A_2 = \begin{pmatrix} A & 0 & B \\ I & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, C_2 = [C, 0, D]$

3) $\alpha = 1, \beta = 0$, $A_3 = \begin{pmatrix} A & BK & 0 \\ 0 & I & 0 \\ 0 & K & 0 \end{pmatrix}, C_3 = [C, DK, 0]$

4) $\alpha = 1, \beta = 1$, $A_4 = \begin{pmatrix} A & 0 & B \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix}, C_4 = [C, 0, D]$

$$\bar{B}_\omega = [B_\omega^T \quad 0 \quad 0]^T, \ \bar{E} = \begin{bmatrix} E & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \bar{f}(k) = [f^T(\bullet) \quad 0 \quad 0]^T$$

It is noted from above analysis, the closed-loop system (4) is component by four sub-systems, so it can be constructed as Markov jump systems

$$\begin{cases} \bar{E}\xi(k+1) = A_{\theta(k)}\xi(k) + \bar{B}_\omega\omega(k) + \bar{f}(k) \\ z(k) = C_{\theta(k)}\xi(k) + D_\omega\omega(k) \end{cases} \tag{5}$$

where $\{\theta(k), k \in \mathbb{Z}\}$ is discrete Markov chains that takes values in $l = \{1,2,3,4\}$. Its transition probability matrix is $\Pi = \{\lambda_{ij}\}$, which is defined $\lambda_{ij} = P(\theta(k+1) = j \mid \theta(k) = i)$ with $\lambda_{ij} \ge 0$ and $\sum_{j=1}^{4} \lambda_{ij} = 1$ for all $i \in l$.

Assume the nonlinear $\bar{f}(k)$ of system (4) satisfies

$$\bar{f}^{\mathrm{T}}(k)\bar{f}(k) \le \xi^{\mathrm{T}}(k)H\xi(k) \tag{6}$$

Hear $H$ is known real constant matrix.

***Remark 2.1*** When $E = I, f = 0$, system(5) reduces to

$$\begin{cases} \xi(k+1) = A_{\theta(k)}\xi(k) + \bar{B}_\omega\omega(k) \\ z(k) = C_{\theta(k)}\xi(k) + D_\omega\omega(k) \end{cases} \tag{7}$$

is a special case of this paper, which discussed in [18].

The following definitions will be used in the sequel.

**Definition 2.1** [12] The discrete singular system (5) under without disturbance ($\omega(k) \equiv 0$) is said to be stochastically stable if for any $x_0 \in \mathbb{R}^n$, there exist $\delta(x_0, \theta_0) > 0$ and a scalar $\rho$, such that

$$\lim_{N\to\infty} \mathcal{E}\left\{ \sum_{k=0}^{N} \|x\|^2 \middle| x_0, \theta_0 \right\} \le \rho\delta(x_0, \theta_0) \tag{8}$$

**Definition 2.2** [10] For all $i \in l$, the discrete singular system (5) is said to be

i . regular if $\det(z\bar{E} - A_i)$ is not identically zero.

ii. causal if $\deg(\det(z\bar{E} - A_i)) = rank(\bar{E})$.

iii. stochastically admissible if it is regular, causal and stochastically stable.

**Definition 2.3** [11] System (5) with $u(k) \equiv 0$ is said to be robustly mean square quadratic stability, if there exists a scalar $\gamma > 0$ such that $\|z(k)\|_2 \leq \gamma \|\omega(k)\|_2$ , for any nonzero disturbance $\omega(k) \in L_2[0, \infty)$ , where

$$\|z(k)\|_2 = \sum_{k=0}^{\infty} \mathcal{E}\left\{z^T(k)z(k)\right\}, \|\omega(k)\|_2 = \sum_{k=0}^{\infty} \mathcal{E}\left\{\omega^T(k)\omega(k)\right\}$$

The objective of this paper is to establish a sufficient condition such that the closed-loop system (5) is stochastically stable and to design a state-feedback controller in the form of (2), to satisfy $H_\infty$ performance.

## III. Stability Analysis

In this section, we analyze the stochastic stability and $H_\infty$ performance of system (5). Before presenting the main results, we introduce the following lemma.

**Lemma 3.1** [10] The system (5) with $\omega(k) \equiv 0$ is stochastically admissible if and only if there exists $P_i > 0$ such that the following matrix inequalities holds:

$$\begin{cases} \bar{E}^T P_i \bar{E} \geq 0 \\ A_i^T \bar{P}_i A_i - \bar{E}^T P_i \bar{E} < 0 \end{cases} \tag{9}$$

where $\bar{P}_i = \sum_{j=1}^{N} \lambda_{ij} P_i$

Now, we propose the results of this section as follows.

**Theorem 3.1** Given scalar $\gamma > 0$ for each $i \in l$ , if and only if there exist matrices $P_i = P_i^T$ and $\varepsilon > 0$ such that the following matrix inequalities (10) and (11) hold

$$\bar{E}^T P_i \bar{E} \geq 0 \tag{10}$$

$$\begin{bmatrix} A_i^{\mathrm{T}} \bar{P}_i A_i - \bar{E}^{\mathrm{T}} P_i \bar{E} & A_i^{\mathrm{T}} \bar{P}_i \bar{B}_\omega & C_i^{\mathrm{T}} & H^{\mathrm{T}} \\ * & \bar{B}_\omega^{\mathrm{T}} \bar{P}_i \bar{B}_\omega - \gamma^2 I & D^T & 0 \\ * & * & -\gamma I & 0 \\ * & * & * & -\varepsilon^{-1} I \end{bmatrix} < 0 \tag{11}$$

then the SNCs (5) is stochastically admissible and robust asymptotically stable , moreover satisfies $H_\infty$ performance $\gamma$ norm.

Where $\bar{P}_i = \sum_{j=1}^{4} \lambda_{ij} P_i$

***Proof:*** stochastically admissible and asymptotically stable analysis.

construct the following lyapunov function

$$V(\xi(k), \theta(k)) = \xi^T(k) \bar{E}^T P_{\theta(k)} \bar{E} \xi(k) \tag{12}$$

when $\omega(k) \equiv 0$ , for simplicity $\theta(k) = i$ , calculating the difference of $\Delta V(k, \theta(k))$ along the trajectory of system (5) and taking the mathematical expectation, we have

$$\mathcal{E} \Delta V(\xi(k), \theta(k)) = \mathcal{E}[V(\xi(k+1), \theta(k+1))] - V(\xi(k), \theta(k))$$

$$= \mathcal{E}\left\{\xi^T(k+1) \bar{E}^T P_{i+1} \bar{E} \xi(k+1)\right\} - \xi^T(k) \bar{E}^T P_i \bar{E} \xi(k)$$

$$= \begin{bmatrix} \xi(k) \\ \bar{f}(k) \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} A_i^{\mathrm{T}} \bar{P}_i A_i - \bar{E}^{\mathrm{T}} P_i \bar{E} & H^{\mathrm{T}} \\ H & -\varepsilon^{-1} I \end{bmatrix} \begin{bmatrix} \xi(k) \\ \bar{f}(k) \end{bmatrix}$$

$$\leq -\lambda_{\min} \left\{ -\begin{bmatrix} (A_i^{\mathrm{T}} \bar{P}_i A_i - \bar{E}^{\mathrm{T}} P_i \bar{E}) & H^{\mathrm{T}} \\ H & -\varepsilon^{-1} I \end{bmatrix} \right\} \bar{\xi}^{\mathrm{T}}(k) \bar{\xi}(k)$$

$$\leq -\rho x^{\mathrm{T}}(k) x(k) \tag{13}$$

where $\lambda_{\min}(\bullet)$ denotes the minimum eigenvalue of matrix $(\bullet)$ , $\bar{\xi}(k) = \left[ \xi^{\mathrm{T}}(k) \quad \bar{f}(k)^{\mathrm{T}} \right]^{\mathrm{T}}$ and $0 < \rho < \min\left\{\lambda_{\min}(\bullet)\right\}$ .

Inequality (13) implies that

$$\mathcal{E}[V((k+1), \theta(k+1))] - \mathcal{E}[V(0, \theta(0))] \leq -\rho \sum_{k=0}^{N} \mathcal{E}[x^{\mathrm{T}}(k) x(k)] \tag{14}$$

So $\sum_{k=0}^{N} \mathcal{E}[x^{\mathrm{T}}(k) x(k)] \leq \dfrac{1}{\rho} \mathcal{E}[V(\xi(0), \theta(0))]$ ,

let $N \to \infty$ then

$$\sum_{k=0}^{\infty} \mathcal{E}[x^{\mathrm{T}}(k) x(k)] \leq \frac{1}{\rho} \mathcal{E}[V(\xi(0), \theta(0))] < \infty \tag{15}$$

applying definition 2.1 and lemma 3.1, we can obtain the system is stochastically admissible and asymptotically stable.

Then we prove system (5) with zero initial condition, the output $z(k)$ satisfies $H_\infty$ performance.

Definition $J = \sum_{k=0}^{N} \mathcal{E}\left\{\gamma^{-1} z^T(k) z(k) - \gamma \omega^T(k) \omega(k)\right\}$ (16)

Through computation, we have

$$J \leq \mathcal{E} \sum_{k=0}^{N} \left\{\Delta V(k, \theta(k)) + \gamma^{-1} z^T(k) z(k) - \gamma \omega^T(k) \omega(k)\right\}$$

$$= \mathcal{E} \sum_{k=0}^{N} \left\{ \begin{bmatrix} \xi(k) \\ \bar{f}(k) \\ \omega(k) \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \tilde{A} & A_i^{\mathrm{T}} \bar{P}_i \bar{B}_\omega & C_i^{\mathrm{T}} \\ * & \bar{B}_\omega^{\mathrm{T}} \bar{P}_i \bar{B}_\omega - \gamma^2 I & D^T \\ * & * & -\gamma I \end{bmatrix} \begin{bmatrix} \xi(k) \\ \bar{f}(k) \\ \omega(k) \end{bmatrix} \right\} \tag{17}$$

where $\tilde{A} = A_i^{\mathrm{T}} \bar{P}_i A_i - \bar{E}^{\mathrm{T}} P_i \bar{E} + \varepsilon H^{\mathrm{T}} H$

According to the theorem 3.1, applying Schur complements formula

$\lim\limits_{N \to \infty} \mathcal{E} \sum_{k=0}^{N} \left\{Inequality(17)\right\} < 0$ ,

then $\sum_{k=0}^{\infty} \mathcal{E}\left\{z^T(k) z(k)\right\} < \gamma \sum_{k=0}^{\infty} \mathcal{E}\left\{\omega^T(k) \omega(k)\right\}$ ,

so the system has $H_\infty$ performance $\gamma$ norm.

This completes the proof.

***Remark 3.1*** For systems (7), considered the random data packet dropout as Bernoulli process ( $\alpha = \beta = 1$ or $\alpha = \beta = 0$ ) for state feedback controller designing in [18]. Without disturbance ( $\omega(k) \equiv 0$ ) for systems (7), a question of robust mean square stability of networked control systems with packed dropout was investigated in [15], in the theorem 3.1, we consider the system with external and nonlinear disturbance simultaneously, and establish a sufficient stochastic stability condition for state feedback controller designing, which is more generality in actual networked data transmission and application.

## IV. State Feedback $H_\infty$ Controller Design

In theorem 3.1, not given out the controller gain $K$ solve method, in this section, we will present a design method for the state feedback controller (2) in terms of LMIs.

**Theorem 4.1** Given scalars $\gamma > 0$, $\varepsilon > 0$ for each $i \in l$, if there exist matrices $X_i > 0$, $P_i > 0$ and $K$, such that the inequalities (18) and (19) are feasible, then the SNCs (5) is stochastically admissible and asymptotically stable, satisfies $H_\infty$ performance norm $\gamma$.

$$\bar{E}^T X_i \bar{E} \geq 0 \tag{18}$$

$$\begin{bmatrix} \Omega_1 & \Omega_2 & \Omega_3 & \Omega_4 & \Omega_5 \\ * & -\sum_{j\in l} \lambda_{ij} P_i & 0 & C_{0i} + DKE_i & 0 \\ * & * & -\gamma I & D_\omega^T & 0 \\ * & * & * & -\gamma I & 0 \\ * & * & * & * & -\varepsilon H^{-1} \end{bmatrix} < 0 \tag{19}$$

where

$$\Omega_2 = \begin{bmatrix} \sqrt{\lambda_{i1}}(A_{0i} + M_i K N_i) \\ \sqrt{\lambda_{i2}}(A_{0i} + M_i K N_i) \\ \sqrt{\lambda_{i3}}(A_{0i} + M_i K N_i) \\ \sqrt{\lambda_{i4}}(A_{0i} + M_i K N_i) \end{bmatrix} \quad \Omega_3 = \begin{bmatrix} \sqrt{\lambda_{i1}}\bar{B}_\omega \\ \sqrt{\lambda_{i2}}\bar{B}_\omega \\ \sqrt{\lambda_{i3}}\bar{B}_\omega \\ \sqrt{\lambda_{i4}}\bar{B}_\omega \end{bmatrix}$$

$$\Omega_1 = -diag\left[X_i, X_i, X_i, X_i\right] \quad \Omega_4 = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}^{\mathrm{T}}$$

$$\Omega_5 = \begin{bmatrix} X_i & 0 & 0 & 0 \end{bmatrix}^{\mathrm{T}}$$

***Proof:*** First, in theorem 3.1, we rewrite $A_i$ and $C_i$ as

$$A_i = A_{0i} + M_i K N_i, \quad C_i = C_{0i} + DKE_i$$

where

$$A_{01} = \begin{bmatrix} A & 0 & 0 \\ I & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, C_{01} = \begin{bmatrix} C & 0 & 0 \end{bmatrix}$$

$$A_{02} = \begin{bmatrix} A & 0 & B \\ I & 0 & 0 \\ 0 & 0 & I \end{bmatrix}, C_{02} = \begin{bmatrix} C & 0 & D \end{bmatrix}$$

$$A_{03} = \begin{bmatrix} A & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix}, C_{03} = \begin{bmatrix} C & 0 & 0 \end{bmatrix}$$

$$A_{04} = \begin{bmatrix} A & 0 & B \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}, C_{04} = \begin{bmatrix} C & 0 & D \end{bmatrix}$$

$$M_1 = M_2 = \begin{bmatrix} B^T & 0 & I \end{bmatrix}^T, M_3 = M_4 = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^T$$

$$N_1 = \begin{bmatrix} I & 0 & 0 \end{bmatrix}, N_2 = \begin{bmatrix} 0 & I & 0 \end{bmatrix} \quad N_3 = N_4 = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$$

$$E_1 = \begin{bmatrix} I & 0 & 0 \end{bmatrix}, E_3 = \begin{bmatrix} 0 & I & 0 \end{bmatrix} \quad E_2 = E_4 = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$$

Then applying Schur complements formula, inequality (11) can be rewritten as an equivalent form (20)

$$\begin{bmatrix} \bar{\Omega}_1 & \bar{\Omega}_2 & \bar{\Omega}_3 & \bar{\Omega}_4 & \bar{\Omega}_5 \\ * & -\sum_{j\in l} \lambda_{ij} P_i & 0 & C_{0i} + DKE_i & 0 \\ * & * & -\gamma I & D_\omega^T & 0 \\ * & * & * & -\gamma I & 0 \\ * & * & * & * & -\varepsilon H^{-1} \end{bmatrix} < 0 \tag{20}$$

where

$$\bar{\Omega}_1 = -diag\left[P_i^{-1}, P_i^{-1}, P_i^{-1}, P_i^{-1}\right], \ \bar{\Omega}_2 = \Omega_2, \bar{\Omega}_3 = \Omega_3, \bar{\Omega}_4 = \Omega_4$$

$$\bar{\Omega}_5 = \begin{bmatrix} P_i^{-1} & 0 & 0 & 0 \end{bmatrix}^{\mathrm{T}}$$

Let $P_i^{-1} = X_i$ in inequality (20), it yields (19).

This completes the proof.

***Remark 4.1*** It should be noted that inequality (20) are non-convex due to $P_i^{-1}$, which can be solved via the cone complementary linearization method [17], the controller gain solve method converted into the following nonlinear optimization problem involving LMI conditions.

$$\min_{\gamma, P_i, X_i} tr(\sum_{i=1}^{4} P_i X_i)$$

$$\text{s}ubject\,to \quad i)(18) and (19), \ ii)\begin{bmatrix} P_i & I \\ I & X_i \end{bmatrix} > 0 \tag{21}$$

The $\gamma$ sub-optimal $H_\infty$ controller gain $K$ can be found using the following algorithm:

***Algorithm 4.1***

*Step 1* Find an initial feasible set $(P^0, X^0, K^0)$ satisfying (21), let $k = 0$.

*Step 2* Solve the following LMI problem

$$\begin{cases} \min_{\gamma, P_i, X_i} tr(\sum_{i=1}^{4} P_i^k X_i + P_i X_i^k) \\ \text{s}ubject\,to \quad i)(18) and (19), ii)\begin{bmatrix} P_i & I \\ I & X_i \end{bmatrix} > 0 \end{cases}$$

*Step 3* let $P_i^{k+1} = P_i, X_i^{k+1} = X_i, K^{k+1} = K^k$

*Step 4* If $k > N$, where $N$ is the maximum numbers of iterations allowed, give up and stop.

*Step 5* If $(P, X, K)$ satisfy (21), then stop, if not, let $k = k + 1$ go to step2.

## V. Illustrative Example

In this section, we present a numerical example to illustrate the theoretical results developed earlier. Consider the system (1) with the following matrices borrowed from [18] example 3, we have the parameters

$$E = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, A = \begin{bmatrix} 1 & 1 \\ -1 & -0.5 \end{bmatrix}, \ B = \begin{bmatrix} 6 \\ 2 \end{bmatrix}, B_\omega = \begin{bmatrix} -0.03 & 0.01 \\ 0 & 0.01 \end{bmatrix}$$

$$C = \begin{bmatrix} -0.1 & -0.05 \end{bmatrix}, D_\omega = \begin{bmatrix} 0.01 & 0.01 \end{bmatrix}, \ D = 0.5, H = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Applying augmentation matrix method, it can be obtained parameters of system (5)

$$A_{01}=\begin{bmatrix}1 & 1 & 0 & 0 & 0 & 0\\ -1 & -0.5 & 0 & 0 & 0 & 0\\ 1 & 0 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\end{bmatrix}$$

$$A_{02}=\begin{bmatrix}1 & 1 & 0 & 0 & 6 & 0\\ -1 & -0.5 & 0 & 0 & 2 & 0\\ 1 & 0 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0 & 0 & 1\end{bmatrix}$$

$$A_{03}=\begin{bmatrix}1 & 1 & 0 & 0 & 0 & 0\\ -1 & -0.5 & 0 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\end{bmatrix}$$

$$A_{04}=\begin{bmatrix}1 & 1 & 0 & 0 & 6 & 0\\ -1 & -0.5 & 0 & 0 & 2 & 0\\ 0 & 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0 & 0 & 1\end{bmatrix}$$

$$C_{01}=C_{03}=\begin{bmatrix}-0.1 & -0.05 & 0 & 0 & 0 & 0\end{bmatrix}$$

$$C_{02}=C_{04}=\begin{bmatrix}-0.1 & -0.05 & 0 & 0 & 0.5 & 0\end{bmatrix}$$

$$M_1=M_2=\begin{bmatrix}6 & 2 & 0 & 0 & 0 & 0\end{bmatrix}$$

$$M_3=M_4=\begin{bmatrix}0 & 0 & 0 & 0 & 0 & 0\end{bmatrix} \quad N_1=\begin{bmatrix}1 & 0 & 0 & 0 & 0 & 0\end{bmatrix}$$

$$N_2=\begin{bmatrix}0 & 0 & 1 & 0 & 0 & 0\end{bmatrix} \quad N_3=N_4=\begin{bmatrix}0 & 0 & 0 & 0 & 0 & 0\end{bmatrix}$$

$$E_1=\begin{bmatrix}1 & 0 & 0 & 0 & 0 & 0\end{bmatrix}$$

$$E_3=\begin{bmatrix}0 & 0 & 1 & 0 & 0 & 0\end{bmatrix}, E_2=E_4=\begin{bmatrix}0 & 0 & 0 & 0 & 0 & 0\end{bmatrix}$$

The transition probability matrix is given as

$$\Pi=\begin{bmatrix}0.3 & 0.5 & 0.1 & 0.1\\ 0.2 & 0.3 & 0.5 & 0.0\\ 0.1 & 0.7 & 0.0 & 0.2\\ 0.6 & 0.1 & 0.2 & 0.1\end{bmatrix}$$

applying algorithm 4.1 to this example, it can be obtained $\gamma_{\min}=0.175$, the controller gain matrix

$$K=\begin{bmatrix}0.2490 & -0.4263\\ -0.2328 & 0.4256\end{bmatrix}$$

In the initial condition $x(0)=[1 \quad -1]^{\mathrm{T}}$ and $\theta(0)=1$, the external disturbance $\omega(k)$ is assumed to be,

$$\omega(k)=\begin{cases}0.2, & 10\le k\le 15\\ -0.2, & 15\le k\le 25\\ 0, & else\end{cases}$$

the system state response and control output trajectory are given in Fig.2 and Fig.3, respectively.

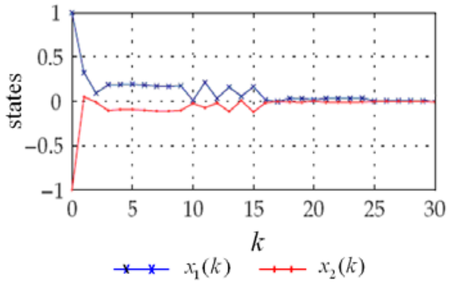


Fig.2 The state response of system

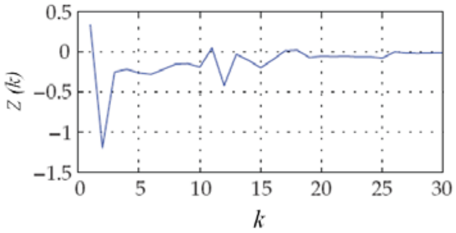


Fig.3 The control output of system

from the simulation, it shows the asymptotical stability of the closed-loop system.

For system (1), when $E=I, f=0$, in [14] only considered the networked control systems are modeled as a Markov jump linear system with two operation modes, compared with references [15, 16, 18], our result has more generality in practical application.

## CONCLUSION

In this paper, the problem of robust $H_\infty$ control of discrete time singular systems over networks has been investigated, both the data packed dropout and nonlinear disturbance are considered, the system communication link failure is modeled as markovian jump singular systems with four models, based on the Lyapunov function method and LMI technique, an approach has been developed to design a state feedback $H_\infty$ controller, such that the closed-loop system is stochastically stable and preserves a guaranteed $H_\infty$ performance. An example is shown that the approach presented in this paper is effective.

## REFERENCES

[1] JUNG E H, LEE H H, SOO Y, "LMI based output feedback control of networked control systems," *IEEE Transactions on Automatic Control*, 2004, 49: 311-314.

[2] HU S S, ZHU Q X, "Stochastic optimal control and analysis of stability of networked systems with long delay," *Automatic*, 2003, 39: 1877-1884.

[3] L.Q. Zhang, Y. Shi, T. W. Chen, and B. Hang, "A new method for stabilization of networked control systems with random delays," *IEEE Transactions on Automatic Control*, 2005, 50(8): 1177-1181.

[4] XIONG J L, LAM J, "Stabilization of linear systems over networks with bounded packed loss," *Automatic*, 2007, 43: 80-87.

[5] YU M, WANG L, CHU T G, “An LMI approach to networked control systems with data packed dropout and transmission delays,” *Proc. IEEE Conf. on Decision and Control*, 2004:14-17.

[6] M. Yu, L. Wang, T. Chu, and F. Hao, “An LMI approach to networked control systems with data packed dropout and transmission delays,” *Proc. 43rd IEEE Conf. Decision and Control*, 2004: 3545-3550.

[7] Xiao Xiaoqing, Zhang Zhenjuan, Zhou lei, Lu Guoping, “Stabilization of descriptor systems over networks with random communication delays,” *Proceedings of the 30th Chinese Control conference* , July, 2011: 4704-4709.

[8] S. XU,C. Yang, “state feedback control for discrete singular systems,” *IEEE Transactions on Automatic Control*, 2000, 45: 1405-1409.

[9] E. Fridman, “ $H_\infty$ control of linear state delay descriptor systems: an LMI approach,” *Linear Algebra Appl.*, 2002, 15: 271-279.

[10] S. Y. Xu, and J. Lam, “Robust control and filtering of singular systems,” *Berlin, Germany: Springer-Verlag*, 2006.

[11] J. Lam, Z. Shu, S. Y. Xu, and E. K. Boukas, “Robust $H_\infty$ control of descriptor discrete-time Markovain jump systems,” *International Journal of Control*, 2007, 83(3): 374-385.

[12] Y. Q. Xia, J. H. Zhang, and E. K. Boukas, “Control for discrete singular hybrid systems,” *Automatic*, 2008, 44(10): 2635-2641.

[13] E. K. BOUKAS, Z. K. LIU, “Robust stability and $H_\infty$ control of discrete-time Jump linear systems with time delay: an LMI approach,” *Proceedings of the 39th IEEE conference on decision and control*, 2000: 1527-1532.

[14] Ma Weiguo, Shao Cheng, “Robust *H*∞ control for networked control systems,” *Journal of Systems Engineering and Electronics*, 2008, 19(5): 1003-1009.

[15] Xie D X, Han X D, Huang H, “Research on robust mean square stability of networked control systems with packed dropout,” *Journal of Systems Engineering and Electronics*, 2010, 21(1): 95-101.

[16] Guoliang Wang, Qingling Zhang and Chunyu Yang, “Exponential filtering for singular systems with markovian jump parameters,” Int. J. Robust. Nonlinear Control, (*2012, Published online in Wiley online library*).

[17] Ghaoui L E, Oustry F, Aitrami M, “A cone complementarity linearization algorithm for static output feedback and related problems,” *IEEE Transactions on Automatic Control*, 1997, 42(8): 1171-1176.

[18] Jinhui Zhang and Yuanqing Xia, “Design of $H_\infty$ fuzzy controllers for nonlinear systems with random data dropouts,” *Optim. Control Appl. Meth*. 2011, 32: 328-349.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# dSpace Based Direct-driven Permanent Magnet Synchronous Wind Power System Modeling and Simulation

**Yan-xia Shen , Fan Li， Dinghui Wu, Ting-long Pan， Xiang-xia Liu**

**Institute of Electrical Automation**
**Jiangnan University**
**Wuxi,China**
**shenyx@jiangnan.edu.cn**

*Abstract*— **When wind speed is below the rated value, the efficiency of captured wind energy must be maximized and the mechanical oscillation be guaranteed to be small. To deal with these problems, This essay describes the advantages of the direct-driven permanent magnet synchronous wind power system, and then introduces the two-frequency-loop model based on frequency separation principle, The LPV model and LPV control method are suggested for the high-frequency part of the system. The output of the high-frequency part is used to compensate the mechanical torque. The mathematical model is built with MATLAB, and the online test is carried out by dSpace. The simulation results show that the controller reduces mechanical oscillation effectively, and enhances the system reliability.**

***Keywords-LPV; direct-driven; permanent magnet synchronous wind turbine; Wind Power Conversion System; dSpace***

## I. INTRODUCTION

Wind power is one of clean renewable resources. After 20 years development, the system performance of wind power generation is gradually improved, and its cost is greatly reduced. Wind power has become an important strategy for national economic development [1, 2].

The wind power system mostly adopts DFIG at present, with a gearbox between generator and wind turbine. The gearbox not only consumes a part of energy, but also has big noises, high failure rate and high maintenance costs. Permanent magnet synchronous generator is a newer type motor without brush and commutator, so it has big power factor. Without gearbox, namely that the wind turbine and generator is connected directly, it can promote the efficiency, improve the reliability of the system and decrease failure rate and maintenance costs.

In the wind power system, the control objective under rated wind speed is to maximize the efficiency of captured wind energy, but the traditional control methods usually cause high mechanical oscillation. Since the wind energy conversion system (WECS) is a typical strong nonlinear system, many control means need to transform nonlinear model to linear model, such as PI or PID control [3], and with some advanced control ways, like LQ and LQG control [4], however, the robustness of all these control methods is low, this becomes an applying restrictions for these control methods. The randomness of wind often results in low resolution linear model, reference [5] proposed LPV (Linear Parameter Varying) control which can solve above problems efficiently.

Firstly, this paper introduces two frequency loop model of the system [6, 7], and then in the low-frequency part, PI control is employed. After build LPV model for the high-frequency part, the LPV control method is adopted to promote the accuracy of the model, also, maximize the efficiency of captured wind energy and restrain oscillation. This paper built simulation model based on MATLAB, then loaded into dSPACE to make on-line experiment, the experiment result s show LPV control can improve the performance of the system, and it proved the feasibility and superiority of the control method.

## II. WIND ENERGY CONVERSION SYSTEM MODEL

It is reasonable to ignore the dynamic process of generator electromagnetic response in this paper, since the electromagnetic time constant is much smaller than the mechanical time constant. The structure of variable speed constant frequency wind power energy conversion system is shown in Fig. 1.

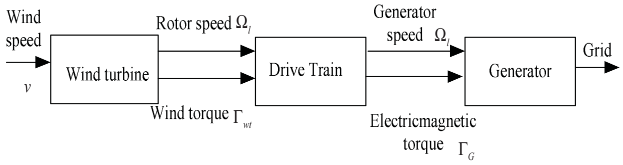

Wind speed
$v$
Wind turbine
Rotor speed $\Omega_l$
Wind torque $\Gamma_{wt}$
Drive Train
Generator speed $\Omega_l$
Electricmagnetic torque $\Gamma_G$
Generator
Grid

Figure. 1 Structure of Variable Speed Constant Frequency WECS

In Fig. 1, the wind turbine and generator are connected directly, drive train represents the shaft, and there is no gearbox between them. For this reason, they have the same speed $\Omega$.

### A. Wind Speed Model

Wind speed $v(t)$ which is a non-statistical random process is decomposed into two components in [8, 9], that is

$$v(t)=\overline{v}(t)+\Delta v(t) \tag{1}$$

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

Where $\bar{v}(t)$ is the low-frequency component, which describes long-time scale and low-frequency changes, it is usual to assume $\bar{v}(t)$ as a Weibull distribution; and $\Delta v(t)$ is high-frequency component, it is made up of Gaussian white noise $e(t)$ as a disturbance signal composed of a first-order filter.

$$\Delta \dot{v}(t) = -\frac{1}{T_w}\Delta v(t) + \frac{1}{T_w}e(t) \tag{2}$$

Where $T_w$ is the time constant of the filter, and $T_w = L_t / \bar{v}$. $L_t$ is the length of wind speed turbulence.

### B. Wind Turbine Model

According to the Bates theory, the mechanical power captured by turbine is:

$$P_{wt} = 0.5\pi\rho R^2 C_p(\lambda) v^3 \tag{3}$$

Where $\rho$ is air density, $R$ is the radius of wind turbine, $v$ is wind speed, $\lambda$ is tip speed ratio, and $\lambda = R \cdot \Omega_l / v$, $\Omega_l$ is the angular velocity of wind turbine rotor, $C_p(\lambda)$ is the power factor of wind turbine, which is defined as:

$$C_p(\lambda, \beta) = 0.22(\frac{116}{\lambda} - 0.4\beta - 5)e^{\frac{-12.5}{\lambda}}$$

Where $\beta$ is the pitch angle of variable-pitch control. For the fixed-pitch control $\beta = 0$.

The torque of wind turbine is:

$$\Gamma_{wt} = \frac{P_{wt}}{\Omega_l} = 0.5\pi\rho R^3 v^2 C_\Gamma(\lambda) \tag{4}$$

Where $C_\Gamma(\lambda)$ is torque coefficient which is defined as $C_\Gamma(\lambda) = C_p(\lambda)/\lambda$.

### C. Drive Train Model

Neglecting the transients, the rigid drive train is expressed as:

$$J_l \dot{\Omega}_l = \Gamma_{wt} - \frac{i}{\eta}\Gamma_G \tag{5}$$

Where $J_l$ is the total inertia of drive train, $i$ is the gear box ratio, for the direct-driven case, its value is 1, $\eta$ is the efficiency of the transmission shaft, $\Gamma_G$ is the electromagnetic torque of the generator. As mentioned above, the formula (4) and (5) constitute the basic low-frequency model of the wind power conversion system.

### D. Model of PMSG

Assuming that,

- Ignore the influence of the core magnetic saturation, excluding the eddy current and hysteresis loss.
- The conductivity of permanent magnetic materials is zero.
- There is no damper winding in rotor.
- The three-phase of stator is symmetrical and the induced EMF (electromotive force) is sinusoidal.

The electromagnetic torque of permanent magnet generator in d,q coordinate system is:

$$\Gamma_G = p(\Phi_d i_q - \Phi_q i_d) = p[\Phi_m i_q + (L_d - L_q) i_d i_q] \tag{6}$$

It is assumed that the load of the generator $R_l$ is independent and symmetric three-segment, and the states and input of the system are defined as:

$$x = [x_1(t) \quad x_2(t)]^T \equiv [i_d(t) \quad i_q(t)]^T$$

$$u \equiv R_l$$

The state model of the generator can be expressed:

$$\begin{cases} \dot{x} = \begin{bmatrix} \frac{1}{L_d + L_s}(-Rx_1 + p(L_q - L_s)x_2\Omega_l) \\ -\frac{1}{L_q + L_s}(-Rx_2 - p(L_d + L_s)x_1\Omega_h + p\Phi_m\Omega_h) \end{bmatrix} + \\ \begin{bmatrix} -\frac{1}{L_d + L_s} & 0 \\ 0 & -\frac{1}{L_q + L_s} \end{bmatrix} \\ y \equiv \Gamma_G = p\Phi_m x_2 \end{cases} \tag{7}$$

Where $R$ is the stator resistance, $L_d$ and $L_q$ are the inductance of the stator in d, q coordinate system, $i_d$ and $i_q$ are the stator current, $L_s$ is the equivalent inductance of grid and converter, $\Phi_m$ is the flux, and is a constant for the permanent material, p is the pole pairs of the generator.

## III. LPV Controller Design

The LPV theory was firstly proposed by Professor Shamma, its dynamic characteristics depend on the adjustable parameters which are measured in real time [5]. Since these parameters can reflect the nonlinearity of the system, LPV system can be applied to describe nonlinear system. The gain scheduling controller is then designed using linear method to make controller gain change with the parameters.

The LPV model of the system can be expressed in [6, 7],

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) + Le(t) \\ y(t) &= Cx(t) \end{aligned} \tag{8}$$

The formula (8) shows that external interference $e(t)$ exists in the LPV model. In order to effectively suppress wind disturbance and improve the system dynamic performance, the controller is designed based on the LPV dynamic model, making the $H_\infty$ norm of the closed-loop transfer function $T_{ez}(s)$ from the disturbance input $e(t)$ to the control output $y(s)$ less than a given performance index, that is,

$$\left\| \boldsymbol{T}_{ez\infty}(s) \right\|_\infty < \gamma_\infty$$

Then the designed state feedback controller is:

$$u(t) = K(\rho(t)) \cdot x(t)$$

The closed-loop system is obtained as:

$$\begin{aligned} \dot{x}(t) &= (A(r(t)) + B(r(t))K(r(t)))x(t) + Le(t) \\ y(t) &= C(\rho(t))x(t) \end{aligned} \tag{9}$$

The affined parameters of this LPV model depend on $\rho(t)$, and the controller matrix is solved according to theorem 1.

Theorem 1[10]:

For the LPV model described by formula(8) and a given positive constant, if there are continuously differentiable symmetric positive definite matrix $X(\rho(t))$, symmetric positive definite matrix $Y$, matrix $V$ and $R(\rho(t))$ satisfying formula(10) for all the parameters, then the parameters of closed-loop (9) are quadratic stability and meet the given $H_\infty$ performance.

$$\begin{bmatrix} -(V+V^T) & * & * & * & * & * \\ M & -X(\rho(t))+Y & * & * & * & * \\ 0 & 0 & -Y & * & * & * \\ L^T(\rho(t)) & 0 & 0 & -\gamma_\infty I & * & * \\ 0 & C(\rho(t))V & 0 & 0 & -\gamma_\infty I & * \\ V^T & 0 & 0 & 0 & 0 & -X(\rho(t)) \end{bmatrix} < 0 \tag{10}$$

where $M = V^T A^T(\rho(t)) + R^T(\rho(t))B^T(\rho(t)) + X(\rho(t))$. If the inequality (10) has feasible solution, then the state feedback controller gain matrix which dependent on the parameters is

$$K(\rho(t)) = R(\rho(t))V^{-1}$$

According to the formula (5),

$$J_T \dot{\overline{\Delta\Omega_l}} = \overline{\Delta\Gamma_{wt}} - \frac{1}{\eta}\frac{\overline{\Gamma_G}}{\overline{\Gamma_{wt}}}\overline{\Delta\Gamma_G} \tag{11}$$

Where $J_T = J_l \overline{\Omega_l}/\overline{\Gamma_{wt}}$, $\dot{\overline{\Delta\Omega_l}} = \frac{\Delta\dot{\Omega}_l}{\overline{\Omega_l}}$, $\dot{\Omega}_l = \dot{\Omega}_l - \dot{\overline{\Omega_l}}$, $\overline{\Omega_l}$ is the stable value of the $\Omega_l$, also for the $\overline{\Omega_l}$, $\overline{\Delta\Omega_l}$ and $\overline{\Delta\Gamma_{wt}}$.

According to the high-frequency pulsation wind speed modeling method in [9] and formula (2), there is

$$\dot{\overline{\Delta v}} = \frac{1}{T_w}\left(e - \overline{\Delta v}\right) \tag{12}$$

From formula (4) and low-frequency sub-model,

$$\dot{\overline{\Delta\Gamma_{wt}}} = \gamma \cdot \dot{\overline{\Delta\Omega_l}} + (2-\gamma)\dot{\overline{\Delta v}} \tag{13}$$

Where $\gamma$ depends on the low-frequency operating point of the system, its value is $\gamma = \frac{\overline{\lambda} C_p'(\overline{\lambda})}{C_p(\overline{\lambda})} - 1$, and $C_p'(\overline{\lambda}) = \frac{dC_p(\overline{\lambda})}{d\lambda}$.

Put the formula (11) and (12) into (13), there is

$$\begin{aligned} \dot{\overline{\Delta\Gamma_{wt}}}(t) &= \left(\frac{\gamma}{J_T} - \frac{1}{T_w}\right)\overline{\Delta\Gamma_{wt}}(t) + \frac{\gamma}{T_w}\overline{\Delta\Omega_l}(t) \\ &- \frac{\gamma}{J_T\eta}\frac{\overline{\Gamma_G}}{\overline{\Gamma_{wt}}}\overline{\Delta\Gamma_G}(t) + \frac{2-\gamma}{T_w}e(t) \end{aligned} \tag{14}$$

Formula (11) and (14) constitute the high-frequency sub-model of the conversion system, for this model and according to the theory above, $u(t) = \overline{\Delta\Gamma_G}$ is chose as the control input, $x(t) = \left[\overline{\Delta\Omega_l} \quad \overline{\Delta\Gamma_{wt}}\right]^T$ is the state vector, $y(t) = \overline{\Delta\lambda}(t) = C(\rho(t))x(t)$ is defined to be the output vector, and the matrixes obtained by formula (14) are

$$A = \begin{bmatrix} 0 & 1/J_T \\ \gamma/T_w & \gamma/J_T - 1/T_w \end{bmatrix},\ B = \begin{bmatrix} \frac{-1}{J_T}\frac{\overline{\Gamma_G}}{\overline{\Gamma_{wt}}} \\ \frac{-\gamma}{J_T}\frac{\overline{\Gamma_G}}{\overline{\Gamma_{wt}}} \end{bmatrix},\ L = \begin{bmatrix} 0 & (2-\gamma)/T_w \end{bmatrix}^T,$$

$$C = \begin{bmatrix} \frac{2}{2-\gamma} & -\frac{1}{2-\gamma} \end{bmatrix}$$

Use the LMI toolbox to describe the matrix of the theorem 1 to obtain the controller K.

## IV. Online Simulation And Results Analysis

According to Fig. 1 and the analysis above, the general structure diagram of the direct-driven permanent magnet synchronous wind power system based on LPV is shown in Fig. 2, and the online experiment is done by dSPACE. dSPACE system is a development and testing platform based on MATLAB/Simulink in real-time environment, it can be connected with MATLAB/Simulink seamlessly, and can realize real-time control and modify for the control system, so it is much easier and overcomes some inconvenience of on-line simulation.

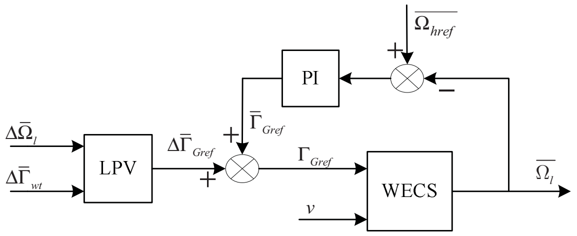

$\overline{\Omega}_{href}$
PI
+
−
$\overline{\Gamma}_{Gref}$
+
$\Delta\overline{\Omega}_l$
$\Delta\overline{\Gamma}_{wt}$
LPV
$\Delta\overline{\Gamma}_{Gref}$
+
$\Gamma_{Gref}$
WECS
$v$
$\overline{\Omega}_l$

Figure. 2 Gain Scheduling Control Structure Based on LPV

In MATLAB simulation environment, find the solution of matrix in theorem 1 with LMI and Simulink toolbox, and build the general simulation diagram of the system, then download the simulation to the dSPACE to do the real-time simulation experiment. The parameters of the experiment are shown in Table 1.

TABLE I. EXPERIMENT PARAMETERS

| Name | Value | Name | Value |
|---|---|---|---|
| R | 2.5 m | ρ | 1.25kg/m$^3$ |
| $T_w$ | 21.4286s | η | 0.95 |
| $J_T$ | 0.563kg*m2 | $C_{pmax}$ | 0.476 |
| $T_{Gmax}$ | 40Nm | $\lambda_{out}$ | 7 |

The experiment results of power factor of the wind turbine $C_p$ and tip speed ratio $\lambda$ (lam) are shown in Fig. 3 and Fig. 4.

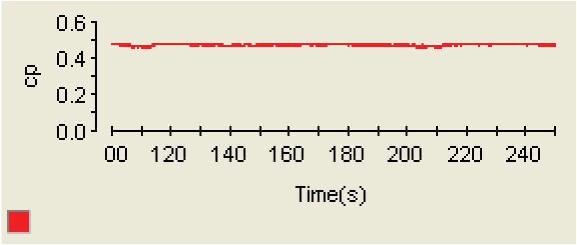


Figure. 3 Power Factor

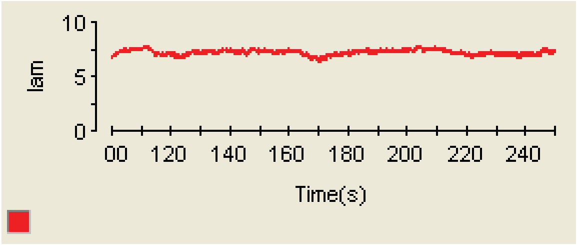


Figure. 4 Tip Ratio

From Fig. 3, it is easy to see that the value of $C_p$ is stable and very closed to the maximum value 0.476; and the Fig. 4 shows that the accuracy of tip speed ratio tracking the best value is high, and the robustness of the system is good, so the controller can capture the wind energy as much as possible, and the oscillation is small, solve the contradictions between the wind energy capturing and big oscillation. The experiment results show the effectiveness and advantages of this control method.

## V. CONCLUSIONS

The situation discussed in this paper is below the rated wind speed of the direct-driven permanent magnetic synchronous wind power system, we need to capture the maximum wind energy and to ensure the smaller mechanical vibration, while increasing the model precision, firstly we build the basic model of the system, then for the high-frequency part, according to the LPV theory to linear the part and design the LPV controller, with PI control of the low-frequency part, the torque can be controlled well. The simulation results based on the dSPACE show that the control method of this paper can fulfill the control objective and efficiently improve the performance, while the real-time experiments are still on research. The research of this paper has broad application prospects in direct-driven permanent magnetic synchronous wind power system.

## ACKNOWLEDGMENT

This work was supported in part by NSFC under Grant Nos. 61104183 and Program for New Century Excellent Talents in University NCET-10-0437.

## REFERENCES

[1] Akpinar, E.K, Akpinar, S, “An investigation for wind power potential required in installation of wind energy conversion system,” Proceedings of the Institution of Mechanical Engineers, vol. 220, pp. 1-13, 2006.

[2] Yazhou Lei, Gordon Lightbody, “Wind power and electricity market,” Automation of Electric Power Systems, vol. 29, pp. 1-5, 2005.

[3] Tapia A, Tapia G, Ostolaza Jx, José Ramón Sáenz, “Modeling and control of a wind turbine driven doubly fed induction generator,” IEEE Trans-actions on Energy Conversion, vol. 18, pp. 194-204, 2005.

[4] Muhando Be, Senjyu T, Urasaki N, Naomitsu Urasaki, Atsushi Yona, Hiroshi Kinjo, Toshihisa Funabashi., “Gain scheduling control of variable speed WTG under widely varying turbulence loading,” Renewable Energy, vol. 32, pp. 2407-2423, 2005.

[5] Shamma,J., Athans,H, “Guaranteed properties of gain scheduled control for linear parameter-varying plants,” Automatica, vol. 27, pp. 559-564, 1991.

[6] Inlian Munteanu, Antoneta Iuliana Brarcu, Nicolaos-Antonic Cutululis, Emil Ceanga, Optimal Control of Wind Energy Systems, London: Springer, 2008.

[7] Inlian Munteanu, Nicolaos Antonio Cutululis, Antoneta Iuliana Bratcu, Emil Ceanga, “Optimization of Variable Speed Wind Power Systems Based on a LQG Approach,” Control Engineering Practice, vol. 13, pp. 903-912, 2005.

[8] Bianchi F, De Battista H, Mantz RJ, Wind turbine control systems principles, modelling and gain scheduling design. London: Springer, 2008.

[9] Nichita C, Luca D, Dakyo B, Ceang E, “Large band simulation of the wind speed for real time wind turbine simulators,” IEEE Transactions on Energy Conversion, vol. 17, pp. 523-529, 2005.

[10] Junling Wang, Delay Linear parameter varying system stability analysis and gain scheduling control. Beijing: science press, 2008.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Study of MA protection based on Homomorphic Encryption and Composite Function Technology

Wu Jiehong[1] Yin Hang[1]
Engineering Training Center
Shenyang Aerospace University
Shenyang , China
wujiehong@sau.edu.cn

Zhang Po[2] Shi Xiangbin[2]
Computer College
Shenyang Aerospace University
Shenyang , China
wujiehong@sau.edu.cn

***Abstract-*** **Mobile agents ( MA ) are autonomous software entities that are able to migrate across heterogeneous network execution environments. Mobility and autonomy compensate the deficiencies of distributed technology pretty well. But the security issues with mobile agents have not been solved and are becoming obstacles for the application of mobile agents. Homomorphic encryption is a technique in which the encrypted mobile codes can be executed directly on different platforms without decryption. This paper presents a protection scheme of mobile agent (AMHCFES) in network management application, which combining protect method of homomorphic encryption and composite function technology. The correctness and security proof of AMHCFES protection scheme are given in this paper, and malicious hosts can be avoided effectively using this scheme**

***Keywords- mobile agent; homomorphic encryption; composite functio; active protection***

## I. Introduction

The original idea of moving cryptography comes from calculating encrypted mobile agents directly, but as the homomorphic encryption scheme which supporting the idea of moving cryptography can't be found, so moving cryptography can't be used in practice.[1-5] This paper gives a practical scheme to realize the ideas of moving cryptography. This is a compound method, organized by composite function and homomorphic encryption scheme. Both codes and data can be encrypted using this method, and the encrypted program can be executed directly without decryption. This method is an extension of moving cryptography put forward by Sander and Tschudin, which preserve many advantages and get rid of many drawbacks of original cryptography[6-9].

## II. Theory Related

The scheme put forward in this paper is based on theory of three address code, homomorphic encryption scheme (HES) and composite function(FnC).

### *2.1 Three Address Code*

Most original programs will be translated into executing objective codes using compiler. There are several phases before creating objective codes. Explicit middle forms will be created after grammatical analysis and semantic analysis, three address codes is one of the middle forms[10]. Three address codes are description of a series of strings, e.g. x:=y op z here, x, y, z are names of constants or variables, op is a random operator. Usually, three addresses will be included in three address codes, two for operands, one for result. So, original expression may changed into following expressions:

$t1 := y * z$

$t2 := x + t1$

t1 and t2 are temporary variables created by compiler.

### *2.2 Addition-multiplication homomorphic(AMH)*

Addition –multiplication homomorphic is a subset of secret homomorphic. It is defined by Sander and Tschudin as following forms: Suppose R and S make a ring, then there is a encryption function $E : R \rightarrow S$ .

(a) Addition homomorphic means there is a valid algorithm PLUS to calculate E(x+y) according to E(x) and E(y), but don't need to know the concrete size of x and y.

(b) Multiplication homomorphic means there is a valid algorithm MULT to calculate E(xy) according to E(x) and E(y), but don't need to know the concrete size of x and y.

Addition homomorphic and multiplication homomorphic keep back addition and multiplication separately[11-15], both secrecy homomorphic and addition-multiplication homomorphic may guarantee the security of arithmetic operation on encrypted data, and needn't to decrypt the data.

### *2.3 Composite Function*

Composite function is defined as follows: it is consisted of output of h(x) and input of g(x), and shown as $f(x) = g . h$ or $f(x) = g(h(x))$ in math, h(x) is the hidden original function. The agent host which owns function must choose a conversion matrix g(x) to create a composite function f(x). Compare f(x) with encrypted function h(x), f(x) is a different function. So, security and integrity of data get guarantee[16-18]. Because the result of composite function f(x) is encrypted, malicious host don't know the result of function. The owner of function(that is the owner of mobile agent) gets the encrypted result through function g(x). Figure 1 is as follows.

This work is supported by natural science foundation of China (61170185), natural science foundation of Liaoning province (20101082) ,and scientific research project of Liaoning education department(L2011030)

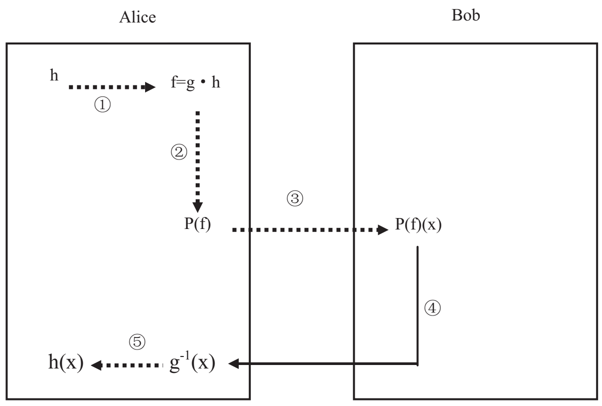


Fig.1 Composite Function

Alice is the owner of agent and have function h(x), she wants to calculate the input x of Bob, but she won't want expose herself function, so she choose a function g(x), and create a function f(x), then send it to Bob. Bob calculates result through f(x) function using his input x, and send result to Alice. Bob can't calculate function h(x), because what he can see is just f(x). Only Alice can get the real result of h(x), through adding f(x) into inverse function, that is h(x) = $g^{-1}$(f(x)).

## III. Addition Multiplication Homomorphic and Composite Function Encryption Scheme(AMHCFES)

First, data and state information of three address codes will be encrypted in this scheme, then three address codes operating code will be encrypted through composite function. Concrete operation steps are as follows:

Three address codes operands will be encrypted using addition-multiplication homomorphic encryption scheme.

Take away the operands dependance problem aroused by addition-multiplication homomorphic encryption scheme(take away encrypted data by addition-multiplication homomorphic)

Three address codes operating code statements will be encrypted using composite function technology.

Solve the operating code dependance problem aroused by composite function technology.

At first, original three address codes will be got from compiler in this scheme, and sensitive data and state information of mobile agents will be encrypted using AMH, then three address codes operating code will be encrypted using composite function technology after first encryption. Double encryption problems of operands and operating code will encounter during the first and second encryption, that will arouse the incorrect encryption of mobile agent. So AMHCFES will find all operands and operating code statements which have been double encrypted, and take away such statements. Double encryption problems will be solved to every encryption process. Encrypted three address codes will be created in AMHCFES, this codes execute the same task as original three address codes and get the same effect. But encrypted three address codes are hard for malicious host to read and modify mobile agents' code, data and state information.

### *3.1 AMHCFES Idea*

The process description of encryption and decryption is given in Fig.2 . When encrypting, AMH is used to encrypt data, a simple function, that is $g(x)=x^3+1$ is used to encrypt operating codes. Then the encrypted mobile agent is sent to other hosts in network directly and results are returned to the original host. When decrypting, the reverse function $g^{-1}(x)=\sqrt[3]{x-1}$ is used first, then AMH scheme is used to get the real return results.

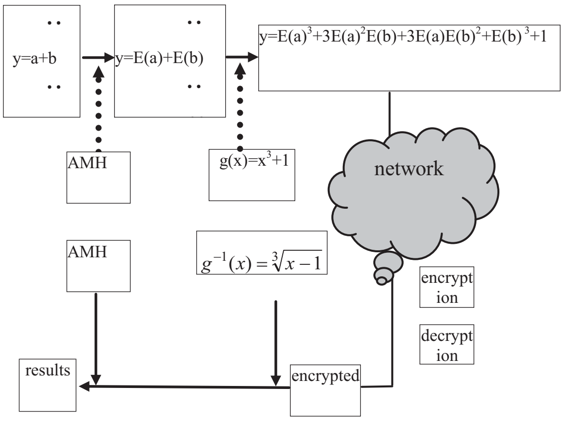


Fig.2 Encryption and Decryption Procedure

### *3.2 AMHCFES Algorithm*

A great number n is used in the algorithm, which makes $n=p\times q$ , *p* and *q* are prime numbers in this expression. Set $Z_p=\{x|x\le p\}$ as original plaintext information set, and set $Z_n=\{x|\ x<n\}$ as ciphertext information set, $Q_p=\{a|\ a\not\in Z_p\}$ is the clue set of encryption.

Addition and Multiplication operating type are defined on $Z_p$.

The encryption and decryption algorithms are as follows:

(1) Encryption algorithm: To given $x\in Z_p$ choosing a random a in $Q_p$, and making x = a mod p. Encrypted result is calculated as y = Ep(x) = a mod n.(This is also can be done by choosing a random r and creating a expression a = x +rp.)

(2) Decryption algorithm: To given $y=E_p(x)\in Z_n$ , using cryptographic key p to recover original value x = Dp(y) = y mod p.

A plain information x can be encrypted into more kinds of ciphertext in this cryptosystem. So, although $E_1(x) \neq E_2(x)$ but D(E1(x)) = D(E2(x)).

### *3.2.1 The Correctness Proof of Algorithm*

Here, theorem 3.1 is given to show the correctness of algorithm, and the proof process will be given.

Theorem 3.1: To all $x \in Z_p$ , existing Dp (Ep(x)) = x.

Proof: let $y = E_p(x)$ and random a is used to encrypt information, exists

a mod n = y (3.1)

Then divide n by p, equality (3.1) meaning

y mod p = (a mod n) mod p = x (3.2)

Done.

The proof of Algorithm AMHCFES's addition and multiplication properties which based on mod n is as follows:

Theorem 3.2: To all s and t in Zp , existing D(E(s)t) = D(E(st)).

Proof: Let's calculate E(s) t and E(s t) first.

(1) E(s) t : In order to encrypt s, let's choose a value $a_1$, make s = $a_1$ mod p, that is obtain the expression $a_1 = k_1 p + s$ , as $y_1 = a_1 \bmod n$ ,so $a_1 = k_2 n + y_1$ ,and then get the expression $k_1 p + s = k_2 n + y_{11}$ .Thus, $y_1$ is obtained.

$$y_1 = k_1 p - k_2 n + s = (k_1 - k_2 q)p + s \qquad (3.3)$$

As $E(s) = y_1, E(s)t = y_1 t$ , get

$$y_1 t = (tk_1 - tk_2 q)p + ts$$

Obtaining the equality through decryption:

$$D(E(s)t) = D(y_1 t) = y_1 t \bmod p = st \qquad (3.4)$$

(2) E(st): In order to encrypt st, let's choose a value $a_2$, make st = $a_2$ mod p, that is $a_2 = k_3 p + st$ , so st is encrypted as $y_2 = a_2 \bmod n$ , that is $a_2 = k_4 n + y_2$ , Obtaining $y_2$ through calculation: $y_2 = (k_3 - k_4 q)p + st$

Obtaining equality through decryption:

$$D(E(st)) = D(y_2) = y_2 \bmod p = st \qquad (3.5)$$

(3) D(E(s)t) = D(E(t)) : Obtaining the results from equality (4) and (5) :

$$y_1 t \bmod p = y_2 \bmod p = st$$

Which means, existing D(E(s)t) = D(E(t)) based on mod p. Done.

### *3.2.2 Examples of Algorithm AMHCFES*

Any input random t can be encrypted automatically using this algorithm[19]. The auto encryption property of algorithm AMHCFES is shown by following example.

E.g. supposing p=101, q=71, then n = pq =7171. For the same reason, supposing host of mobile agent provides E(1) =203. Malicious host hope the input vnumber 8 encrypted, then letthe encrypted number 8 multiply E(1), and obtain ciphertext,that is E(8) =1624. In order to test the equality, choosing A=15966, then 15966 mod 7171 = 1624.

Attention : As A∈Qn..={ A|(A ∉ Zn ) ∩ ( A ≥ n )} , then 1624 mod 101 = 8.

### *3.3 Verification of AMHCFES Security*

This scheme is decrypted by calculating x = y mod p, and the security of this algorithm may be tested by arguments on safety[20-21].

Ciphertext Attack: as y∈ Zq , if cryptanalysts want to find number A∈ Qn , they don't need to get p, but they need p to calculate a mod p = x. So, it is difficult for cryptanalysts to find p in module n, the same as factoring to n, if they know cipher only. So it is hard to find initial values only knowing cipher.

Plaintext Attack: if cryptanalysts know a pair of plaintext-cipher(x, y), then they create a data set of t, that is Ai∈ Qn, i = 1,⋯,t, so Ai = y mod n. To every i, there is Ai = x mod p, so p|(Ai – x), p = gcdti=1(Ai – x). But it is hard to let such thing happens.

Integrity Attack: Since module p is needed by all to execute decryption, any open data such as x < p can be used to decrypt data. So malicious host may choose a value to replace any encrypted data[22-24]. But such choice is blind as module p isn't known. It is hard to find initial values.

## IV. CONCLUSION

This scheme provides a new method to encrypt information without any secret key. MA encrypted by AMHCFES can execute tasks on other hosts of network without decryption. It is effective to defend attacks of malicious hosts.

### *4.1 Shortcoming*

There are some restrictions and assumptions in this scheme, which restrict application scope of the scheme.

Assumption 1: This scheme is used for Integer since AMH is based on loop theory.

Assumption 2: Control structure of MA codes can not be encrypted by composite function, because such operators are included in control structure, e.g. logical expression with other type operators in if statements, other operators here are: logical operator, Boolean operator, assignment operator, etc.

### *4.2 Improvement*

Further study is needed to improve scheme, the work will be done in future is as follows:

Data set handled in this scheme is integer because of loop theory assumption. In order to further expand application scope of MA protection, researching how to expand data set to other data types is the study focus afterwards.

In MA encryption algorithm, type of function call are limited to some basic input output statements, so valid calling method of user defined function and system function will be studied.

## REFERENCES

[1] D. Lange, O. Mitsuru. Seven good reasons for mobile agents [J]. Communications of the ACM, 1999, 42(3): 88-89.

[2] U. Topaloglu, C. Bayrak. Secure mobile agent execution in virtual environment[J]. Autonomous Agents and Multi-Agent Systems,2008,16(1): 1 – 12.

[3] Jason C. Hung, Han-Bin Chang, Hsuan-Pu Chang, Yu-Hsin Cheng, Kuo-Yen Lo. Evolution of ubiquitous autonomous agents[J], International Journal of Ad Hoc and Ubiquitous Computing,2009,4(6):334-343.

[4] Stefan Kraxberger,Peter Danner, Daniel Hein. Secure multi-agent system for multi-hop environments[A], Proceedings of the 5th international conference on Mathematical methods, models and architectures for computer network security, Security of multi-agent systems and software protection[C],LNCS, Sep. 2010,. St. Petersburg, Russia, 270 – 283.

[5] Richard Ssekibuule.Mobile Agent Security Against Malicious Platforms[J],Cybernetics and Systems,2010,41(7):522-534.

[6] T. Sander and C.Tschudin. Protecting Mobile Agents Against Malicious Hosts. Mobile Agents and Security, LNCS 1419, Berlin: Springer-Verlag 1998, 44-60.

[7] Ching Lin, Vijay Varadharajan. MobileTrust: a trust enhanced security architecture for mobile agent systems[J], International Journal of Information Security,2010,9(3): 153 – 178.

[8] Xiaogang Wang, Darren Xu, Junzhou Luo. A Free-Roaming Mobile Agent Security Protocol Based on Anonymous Onion Routing and k Anonymous Hops Backwards[A], Proceedings of the 5th international conference on Autonomic and Trusted Computing. Special Session Papers: Routing and Reliable Systems,LNCS,Jun.2008, Oslo, Norway, 588 – 602.

[9] Woei-Jiunn Tsaur, Chien-Hao Ho. A mobile agent protected scheme using pairing-based cryptosystems[J]. International Journal of Mobile Communications,2005,3(2):183-196.

[10] Ruchuan Wang, Xiaolong Xu. Research of MA security mechanism [J]. Chinese Journal of Computers,2002,25(12) : 1294-1301.

[11] Xiang Tan,Minqing Gu,Congming Bao. Mechanism for Mobile Agent Data Protection[J]. Journal of Software,2005,16(3): 477 - 484.

[12] Xiaoping Wu, Honggen Xing, Zhidong Shen. Research of MA security application model based on distributed confidence level [J]. Computer Engineering and Science,2010,32(6):19-22.

[13] Rossilawati Sulaiman, Xu Huang, Dharmendra Sharma. E-health Services with Secure Mobile Agent[A], Proceedings of the 2009 Seventh Annual Communication Networks and Services Research Conference, Communications Networks and Services Research Conference, IEEE computer society,May.2009,270-277

[14] Monia Loulou, Mohamed Jmaiel, Mohamed Mosbah.Dynamic security framework for mobile agent systems: specification, verification and enforcement[J], International Journal of Information and Computer Security,2009,3(3/4):321-336.

[15] Christopher Colby, Karl Crary, Robert Harper, Peter Lee, Frank Pfenning.Automated techniques for provably safe mobile code[J], Theoretical Computer Science,2003,290(2):1175-1199.

[16] Carles Garrigues, Nikos Migas. Protecting mobile agents from external replay attacks[J]. Journal of Systems and Software. 2009, 82(2):197-206.

[17] D M Hein, R Toegl. An Autonomous Attestation Token to Secure Mobile Agents in Disaster Response. LNICST 17,2009, pp:46-57.

[18] Joan Tomas-Buliart, Marcel Fernandez. Protection of Mobile Agents Execution Using a Modified Self-Validating Branch-based Software Watermarking with External Sentinel. Critical Information Infrastructures Security[C]. LNCS 5508, Berlin: Springer- Heidelberg, 2009, 287-294.

[19] S Venkatesan, C Chellappan, P Dhavachelvan. Advanced mobile agent security models for code integrity and malicious availability check[J]. Journal of Network and Computer Applications.2010,33(6):661-671.

[20] G. Vigna. Cryptographic traces for mobile agents[G]//LNCS 1419: Proceedings of Mobile Agents and Security .Berlin:Springer,1998:137-153

[21] F. Hohl. Time limited blackbox security: Protecting mobile agents from malicious host [A]. Mobile agent and security [C]. LNCS 1419, Springer-Verlag, 1998, 92-113.

[22] Weiwei Song, Zhen Ye, Lei Yue. Dynamic trust model and application in mobile agent environment [J]. Journal of Hefei University of Technology(Natural Science),2009,32(1):73-77.

[23] Dengguo Feng, Yu Qin. Research of proving method in credible calculating environment [J]. Chinese Journal of Computers,2008,31(9):1640-1652

[24] Apostolos P. Fournaris .Trust Ensuring Crisis Management Hardware Module[J], Information Security Journal: A Global Perspective,2010,19(2):74-83..

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# Adaptive Control Design of Uncertain Piecewise-Linear Systems

Nanzhu Lin
School of Electrical Engineering
Jilin University
Changchun, China, 130000
Email: nzl.jlu@gmail.com

Xin Wang
and Yanan Zhang
Changchun University of Technology
Changchun, China, 130000
Email: {ynzhang,xwang}@ccut.edu.cn

Chunze Wang and Shu Diao
School of Electrical Engineering
Jilin University
Changchun, China, 130000
Email: {wangcz6509,diaoshu}@jlu.edu.cn

***Abstract*—In this paper, we study adaptive control design problem of piecewise-linear systems with matching parametric uncertainties. The basic idea is to construct a piecewise control law and a piecewise parameter adaptation law in such a way that a piecewise quadric Lyapunov function can be used to establish the global stability. All of the synthesis conditions are formulated as Linear Matrix Inequities and can therefore be efficiently solved. Moreover, the possibility of sliding motion at the boundary of polytopic regions is considered, and the results are demonstrated by application to control a linear motor.**

## I. Introduction

Piecewise linear(PWL) systems are switched linear systems with state space-partition-based switching. PWL systems are very important in representing many practical systems including power electronics [1], [2], robots [3], [4] and biology regulatory networks [5]. In additional, PWL systems can approximate nonlinear systems to any degree of accuracy and therefore provide a useful framework for the analysis and synthesis of a large class of nonlinear systems.

Remarkable progress on the PWL systems have been achieved thanks to the dedication of many researchers over the last decades. Important achievements include stability and stabilization [6]–[11], state and output tracking control [12]–[14] and robust control of PWL systems [15]–[20] among others.

Although the proposed robust control methods [15]–[20] can be applied to deal with the uncertain PWL systems, model uncertainties coming from parametric uncertainties can not be reduced. In order to achieve the required performance, the feedback gains must be increased, resulting in high-gain feedback. On the other hand, there are always the constrain scope of control input in the practical engineering systems. To avoid the high-gain feedback, an Lyapunov-based adaptive control approach was proposed in [21]. However, the synthesis conditions were formulated as Bilinear Matrix Inequities(BMIs), the computation problem of BMIs is still a common challenge.

In this paper, we study the adaptive control design problem of PWL systems with matching parametric uncertainties. The basic idea is to construct a piecewise control law and a piecewise parameter adaptation law in such a way that a piecewise quadric Lyapunov function[PQLF] can be used to establish the global stability. Comparing with [21], all of the synthesis conditions are formulated as Linear Matrix Inequities(LMIs) instead of BMIs, thus are much more simple to solve with existing software such as MATLAB. Moreover, we consider the possibility of sliding motion at the boundary of polytopic regions, which guarantees the rigidity of the proposed approach.

## II. Problem Formulation and Preliminary

Consider the following uncertain PWL systems,

$$\dot{x} = \begin{cases} A_1 x + B_1 u + \varphi_1(x)\theta & \text{if } x \in R_1, \\ A_2 x + B_2 u + \varphi_2(x)\theta & \text{if } x \in R_2, \\ \cdots & \\ A_m x + B_m u + \varphi_m(x)\theta & \text{if } x \in \mathcal{R}_m, \end{cases} \tag{1}$$

where $\mathbb{R}^n = \cup_{i \in \{1,2,\cdots,m\}} R_i$ denotes a partition of the state space into a number of closed polytopic regions; $x \in \mathbb{R}^{n_x}$, $u \in \mathbb{R}^{n_u}$ and $\theta \in \mathbb{R}^{n_\theta}$, denotes the state, input and unknown parameter vector, respectively. Moreover, $A_i, B_i$ are all constant matrices with appropriate dimensions; $\varphi_i(x) : \mathbb{R}^{n_x} \rightarrow \mathbb{R}^{n_x \times n_\theta}$ is a continuous linear or nonlinear function with $\varphi_i(0) = 0$.

As shown in [6], we can construct matrices $E_i$, $F_i$ for each region $R_i$ such that

$$\begin{cases} E_i x \geq 0, & x \in R_i \\ F_i x = F_j x, & x \in R_i \cap R_j \end{cases} \tag{2}$$

*Assumption 1:* The proposed PWL systems satisfy the matching condition, i.e., there exist function $\psi_i(x)$ for each region, such that

$$B_i \psi_i(x) = \varphi_i(x), \quad x \in R_i. \tag{3}$$

Let $\hat{\theta}$ be the estimation of $\theta$, our objectives are to design a piecewise control law $u(t) = u_i(x, \hat{\theta}), x \in R_i$ and a piecewise parameter adaptation law $\dot{\hat{\theta}} = \upsilon_i(x, \hat{\theta}), x \in \mathcal{R}_i$, such that the closed-loop PWL system is asymptotically stable, i.e, all the possible state trajectories will converge to origin.

## III. Main Results

In this section, we provide the main contribution of the paper. An LMI-based adaptive control approach will be devel-

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

oped for the PWL system (1) based on the PQLF theory [6] and the following two lemmas.

*Lemma 1:* [22] For all positive definite matrices $P$, the following inequality holds:

$$G^TP^{-1}G \geq G^T + G - P. \tag{4}$$

*Lemma 2:* [23] Let $\Phi$ be any given positive definite matrix. The following statements are equivalent:

$$\begin{array}{l} (1).\ \Psi + \Xi + \Xi^T < 0; \\ (2).\ \text{There exists a matrix } W \text{ such that} \\ \begin{bmatrix} \Psi + \Phi - (W + W^T) & \Xi^T + W^T \\ * & -\Phi \end{bmatrix} < 0. \end{array} \tag{5}$$

Now, we give the main result of this paper.

*Theorem 1:* If there exist symmetric matrices $T$, $U_i$, $W_i$ and general matrix $V_i$, $R_i$, such that $U_i$, $W_i$ have nonnegative entries and the following LMIs are satisfied with $P_i = F_i^TTF_i$,

$$P_i - E_i^TW_iE_i > 0, \tag{6}$$

$$\begin{bmatrix} -(V_i + V_i^T) & V_i^TA_i^T + R_i^TB_i^T + P_i & V_i^T \\ * & -P_i & 0 \\ * & * & E_i^TU_iE_i - P_i \end{bmatrix} < 0. \tag{7}$$

Then, using the piecewise control law

$$\begin{cases} u = -\psi_i(x)\hat{\theta} + K_ix \\ K_i = -R_iV_i^{-1} \end{cases} \quad x \in \mathcal{R}_i \tag{8}$$

and the piecewise parameter adaptation law

$$\dot{\hat{\theta}} = \varphi_i(x)^TP_ix, \quad x \in \mathcal{R}_i, \tag{9}$$

the resulting closed-loop PWL system is asymptotically stable.

**Proof:** We choose a piecewise quadratic Lyapunov function

$$V(x, \hat{\theta}) = \sum_i \beta_i[x^TP_ix + \tilde{\theta}^T\tilde{\theta}],\ \beta_i = \begin{cases} 1 & x \in R_i \\ 0 & \text{others} \end{cases} \tag{10}$$

where $\tilde{\theta} = \hat{\theta} - \theta$ denotes the estimation error of unknown parameter $\theta$.

Then, the sufficient conditions for the asymptotical stability are

$$\begin{cases} V(x, \hat{\theta}) \geq 0 \\ \dot{V}(x, \hat{\theta}) < 0 \end{cases} \tag{11}$$

That is because if considering the augmented system (1),(8),(9), and let

$$E = \{(x, \hat{\theta}) \in (X, \mathbb{R}^{n_\theta}) | \dot{V}(x, \hat{\theta}) = 0\} \tag{12}$$

then

$$E = \{(0, \theta_1) | \theta_1 \in \mathbb{R}^{n_\theta}\} \tag{13}$$

and $E$ is a invariant set. Therefore,using LaSalle's theorem[25], the solution of the augmented system $(x(t), \hat{\theta}(t))$ converges to the set E, i.e.,

$$\lim_{t \to \infty} x(t) = 0 \tag{14}$$

Therefore, we just need to show the conditions (11) hold. With the help of (10), we learn that the conditions (11) can be implied by

$$\begin{cases} x^TP_ix > 0 \\ 2\dot{x}^TP_ix + 2\tilde{\theta}^T\dot{\hat{\theta}} < 0 \end{cases} \quad x \in R_i \setminus 0\ . \tag{15}$$

Replacing the piecewise control law and adaptation law by (8–9), the sufficient conditions become

$$\begin{cases} x^TP_ix > 0 \\ 2(\bar{A}_ix + \varphi_i(x)\theta - B_i\psi_i(x)\hat{\theta})^TP_ix + 2\tilde{\theta}^T\varphi_i(x)^TP_ix < 0 \end{cases} \tag{16}$$

where $\bar{A}_i = A_i + B_iK_i$.

By eliminating the terms containing $\tilde{\theta}$, the sufficient conditions can be reformulated as the following inequalities

$$\begin{cases} x^TP_ix > 0 \\ x^T(\bar{A}_i^TP_i + P_i\bar{A}_i)x < 0 \end{cases} \quad x \in R_i\ . \tag{17}$$

These conditions can be relaxed using S-procedure and the polyopic region description (2), yielding the sufficient conditions as

$$\begin{cases} P_i - E_i^TW_iE_i > 0 \\ \bar{A}_i^TP_i + P_i\bar{A}_i + E_i^TU_iE_i < 0 \end{cases} \quad x \in R_i\ . \tag{18}$$

It is noted that the first inequality of the above conditions is same with (6), so our rest work is to show the proof of second inequality using condition (7).

Let $Q_i = P_i^{-1}$, $S_i = E_i^TU_iE_i$, then the second inequality can be written as

$$Q_i\bar{A}_i^T + \bar{A}_iQ_i + Q_iS_iQ_i < 0 \quad x \in R_i. \tag{19}$$

The use of Lemma 2 with $\Psi = Q_iS_iQ_i$ and $\Xi = Q_i\bar{A}_i^T$ yields,

$$\begin{bmatrix} Q_iS_iQ_i + \Phi_i - (W_i + W_i^T) & \bar{A}_iQ_i + W_i^T \\ * & -\Phi_i \end{bmatrix} < 0. \tag{20}$$

By the congruence transformation $\begin{bmatrix} V_i & 0 \\ * & P_i \end{bmatrix}$ with $V_i = W_i^{-1}$, the inequality (20) becomes

$$\begin{bmatrix} V_i^T(P_i^{-1}S_iP_i^{-1} + \Phi_i)V_i - (V_i + V_i^T) & V_i^T\bar{A}_i + P_i \\ * & -P_i\Phi_iP_i \end{bmatrix} < 0. \tag{21}$$

A Schur complement argument on the term $V_i^T(P_i^{-1}S_iP_i^{-1} + \Phi_i)V_i$ shows that, the sufficient condition is equivalent to the following inequality

$$\begin{bmatrix} -(V_i + V_i^T) & V_i^T\bar{A}_i + P_i & V_i^T \\ * & -P_i\Phi_iP_i & 0 \\ * & * & -(P_i^{-1}S_iP_i^{-1} + \Phi_i)^{-1} \end{bmatrix} < 0. \tag{22}$$

The use of Lemma 1 with $\Phi_i = P_i^{-1}$ yields,

$$(P_i^{-1}S_iP_i^{-1} + \Phi_i)^{-1} = P_i(S_i + P_i)^{-1}P_i \geq P_i - S_i. \tag{23}$$

This inequality show that the following condition implies the inequality (22),

$$\begin{bmatrix} -(V_i + V_i^T) & V_i^T\bar{A}_i + P_i & V_i^T \\ * & -P_i & 0 \\ * & * & E_i^TU_iE_i - P_i \end{bmatrix} < 0. \tag{24}$$

Replacing $\bar{A}_i$ by $\bar{A}_i^T$, we can obtain the dual of (25) as

$$\begin{bmatrix} -(V_i+V_i^T) & V_i^T\bar{A}_i^T+P_i & V_i^T \\ * & -P_i & 0 \\ * & * & E_i^TU_iE_i-P_i \end{bmatrix} < 0. \quad (25)$$

substituting $\bar{A}_i = \bar{A}_i + B_iK_i$, $R_i = K_iV_i$ into (25), we obtain the inequality 7. In summary, (6–7) are the sufficient conditions for the the asymptotical stability of the resulting closed-loop PWL SYSTEM, which completes the proof.

*Remark 1:* Although the proposed piecewise adaption law can only ensure the boundedness of $\hat{\theta}$ rather than asymptotical convergence to the true parameter value $\theta$, the state trajectory will still converge to the origin as the disturbance converges to zero.

## IV. Simulation Results

In this section, we consider an epoxy core linear motor drive system with negligible electrical dynamics described in the following [24].

$$\dot{x_1} = x_2 \quad (26)$$
$$M\dot{x_2} = u - \theta x_2 + [f_c - f_c e^{-|\dot{x_1}|}]sgn(\dot{x_1}) \quad (27)$$

where $x_1$, $x_2$ are the position and speed of the inertia load, respectively. The control input is $u$, which ensures the system stable. All the parameter values and their physical meanings are illustrated in Table 1, where $\theta \in [0.2, 0.28]$ is the uncertain parameter.

TABLE I: Simulation parameters

| | |
|---|---|
| $M$=0.55 | Motor mass (Kg) |
| $f_c$=1 | the Coulomb friction force |
| $\theta \in [0.2, 0.28]$ | an equivalent viscous friction coefficient |

The objective is to design a state feedback controller with the control gain constrain $\|K\|_\infty \leq 2$ that forces the motor to stop at a certain point.

Given the possible initial velocity $x_2 \in [-2, 2]$. The nonlinear function (27) – (28) can be approximated by PWA functions yielding a PWL system with four regions as below.

$$R_1 = \{x|x \in \mathbb{R}^2|x_2 \in [-2, -1]\},$$
$$R_2 = \{x|x \in \mathbb{R}^2|x_2 \in [-1, 0]\},$$
$$R_3 = \{x|x \in \mathbb{R}^2|x_2 \in [0, 1]\},$$
$$R_4 = \{x|x \in \mathbb{R}^2|x_2 \in [1, 2]\}.$$

Moreover, the system matrix $A_i$ is obtained by its PWA approximation, and

$$A_1 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0.6645 & -0.2690 \\ 0 & 0 & 0 \end{bmatrix} \quad (28)$$
$$A_2 = A_3 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1.8182 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$
$$A_4 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0.6645 & 0.2690 \\ 0 & 0 & 0 \end{bmatrix}$$
$$B_i = \begin{bmatrix} 0 \\ 1.8182 \\ 0 \end{bmatrix}, \quad i \in \{1, 2, 3, 4\}$$

such as

$$\dot{\overline{x}} = A_i\overline{x} + B_iu + \varphi_i\theta \quad (29)$$
$$where\ \overline{x} = \begin{bmatrix} x_1 \\ x_2 \\ 1 \end{bmatrix}\ and\ \varphi_i(x) = \begin{bmatrix} 0 \\ -x_2 \\ 0 \end{bmatrix}. \quad (30)$$

Thus, the Assumption 1 is satisfied, where $\psi_i(x)$=$-0.55x_2$. Therefore, the following piecewise control law and parameter adaptation law are employed,

$$\begin{cases} u = 2\overline{x}_2\hat{\theta} + K_i\overline{x} \\ \dot{\hat{\theta}} = \varphi_i(\overline{x})^TP_i\overline{x} \end{cases} \quad x \in R_i\ , \quad (31)$$

where $K_i$ and $P_i$ can be obtained by solving the LMIs proposed in Theorem 1. The solutions are obtained by MATLAB as below.

$$K_1 = \begin{bmatrix} -0.7003 & -1.1917 & 0.2382 \end{bmatrix} \quad (32)$$
$$K_2 = \begin{bmatrix} -0.6106 & -1.9195 & 0.0118 \end{bmatrix}$$

$$K_3 = \begin{bmatrix} -0.5024 & -1.8053 & -0.0158 \end{bmatrix} \quad (33)$$
$$K_4 = \begin{bmatrix} -0.4098 & -0.8887 & -0.1969 \end{bmatrix}$$

Simulations have been carried out with the initial condition $[x_1, x_2]$=$[-2, -1.5]$ and $\hat{\theta}(0)$=0.24. Fig.1 shows that, the state trajectory of closed-loop PWL system converges to the origin, which illustrates the efficacy of the proposed LMI-based adaptive control method.

For comparison, we try to design the state feedback controller using the BMI-based adaptive control method [21] for the same PWL system, no feasible solution can be found using YALMIP, which illustrates the advantage of the proposed LMI-based adaptive control approach.

## V. A Remark on Sliding Motion

Actually, by extending the idea of [6], if there exists a quadratic Lyapunov function $V_l(x, \tilde{\theta}) \geq 0$ for each boundary

$l = R_i \cap R_j$, such that for all $x \in l$,

$$V_l(x, \tilde{\theta}) \geq 0 \tag{34}$$

$$\frac{\partial V_l}{\partial x}\bar{co}\{A_{\tau k}x + B_\tau u + \varphi_\tau(x)\theta\} + \frac{\partial V_l}{\partial \tilde{\theta}}\dot{\tilde{\theta}} < 0, \tau = i, j \tag{35}$$

the asymptotical stability can be guaranteed, although the sliding motion may happen. Note that for all $x \in l$, $E_l x = 0$, then using similar procedure in the proof of Theorem 1, the constraints (34-35) can also be expressed as LMIs.

## VI. Conclusion

In this paper, a LMI-based adaptive control method is developed for uncertain PWL systems. By constructing proper piecewise model compensator and online parameter adaptation law, the adaptive control design problem can be converted to the control design problem for a nominal PWL system without parameter uncertainties. Furthermore, the stabilization problem for the obtained nominal PWL system are formulated as optimization problems subject to a set of LMIs, which are numerical solvable by the existing software. However, the proposed method still has a limitation that the parametric uncertainty of the PWL systems must satisfy matching condition, which will restrict the application of the proposed synthesis method. Therefore, our future work will focus on how to get rid of this assumption to achieve wider applications.

## References

[1] J. Deane and C. Hamill, "Instability, subharmonics and chaos in power electronics systems," *IEEE Transactions on Power Systems*, vol. 5(3), pp. 260–268, 1990.

[2] G. Guzelis and I. Goknar, "A canonical representation for piecewise linear affine maps and its application to circuit analysis," *IEEE Transactions on Circuits and Systems*, vol. 38, pp. 1342–1354, 1991.

[3] K. S. Narendra, "Multiple model based adaptive control of robotic manipulators," in *Proc. IEEE Conf. on Decision and Control*, (Florida, USA), pp. 1305–1310, 1994.

[4] P. J. Zufiria and R. Riaza, "Neural adaptive control of nonlinear plants via a multiple inverse model approach," *International Journal of Adaptive Control and Signal Processing*, vol. 13, pp. 219–239, 1999.

[5] H. D. Jong, J. L. Hernandez, C. Page, M. Sari, and J. Geiselmann, "Qualitative simulation of genetic regulatory networks using piecewise-linear models," *Bulletin of Mathematical Biology*, vol. 66, pp. 301–340, 2004.

[6] M. Johansson and A. Ranter, "Computation of piecewise quadratic lyapunov functions for hybrid systems," *IEEE Trans. Automatic Control*, vol. 43(4), pp. 555–559, 1998.

[7] Y. Iwatani and S. Hara, "Stability tests and stabilization for piecewise linear systems based on poles and zeros of subsystem," *Automatica*, vol. 42, pp. 1685–1695, 2006.

[8] D. Ding and G. Yang, "State-feedback control design for continuous-time piecewise-linear systems: An LMI approach," in *Proc. American Control Conf.*, (Seattle, Washington), pp. 1104–1108, 2008.

[9] M. Lazar and W. P. M. H. Heemels, "Global input-to-state stability and stabilization of discrete-time piecewise affine systems," *Nonlinear Analysis: Hybrid Systems*, vol. 2, pp. 721–734, 2008.

[10] S. M. Barijough and J. W. Lee, "A computational stability analysis of discrete-time piecewise linear systems," in *Proc. IEEE Conf. on Decision and Control*, (Shanghai, China), pp. 1106–1111, 2009.

[11] Z. Sun, "Stability and contractivity of conewise linear systems," in *Proc. IEEE Multi-Conference on Systems and Control*, (Yokohama, Japan), pp. 2094–2098, 2010.

[12] N. V. D. Wouw, A. Pavlov, K. Pettersen, and H. Nijmeijer, "Output tracking control of PWA systems," in *Proc. IEEE Conf. on Decision and Control*, (San Diego, USA), pp. 2637–2642, 2006.

[13] N. V. D. Wouw and A. Pavlov, "Tracking and synchronisation for a class of PWA systems," *Automatica*, vol. 44, pp. 2909–2915, 2008.

[14] K. Sakuramaa and T. Sugieb, "Trajectory tracking control of bimodal piecewise affine systems," vol. 78(16), pp. 1314–1326, 2005.

[15] Y. Zhu, D. Li, and G. Feng, "H-infinity controller synthesis of uncertain piecewise continuous-time linear systems," *IEE Proc., Control Theory Appl*, vol. 152, pp. 513–519, 2005.

[16] J. Zhang and W. Tang, "Output feedback $H_\infty$ control for uncertain piecewise linear systems," *Journal of Dynamical and Control Systems*, vol. 14(1), pp. 121–144, 2008.

[17] X. Song, S. Zhou, and B. Zhang, "A cone complementarity linearization approach to robust $H_\infty$ controller design for continuous-time piecewise linear systems with linear fractional uncertainties," *Nonlinear Analysis: Hybrid Systems*, vol. 2, pp. 1264–1274, 2008.

[18] Z. D. Wang, G. Wei, and G. Feng, "Reliable $\mathrm{H}_\infty$ control for discrete-time piecewise-linear systems with infinite distributed delays," *Automatica*, vol. 45(12), pp. 2991–2994, 2009.

[19] A. BenAbdallah, M. A. Hammami, and J. Kallel, "Robust stability of uncertain piecewise-linear systems: Lmi approach," *Nonlinear Analysis: Hybrid Systems*, vol. 63, pp. 183–192, 2010.

[20] J. Qiu, G. Feng, and H. Gao, "Approaches to robust $\mathrm{H}_\infty$ static output feedback control of discrete-time piecewise-affine systems with norm-bounded uncertainties," *International Journal of Robust and Nonlinear Control*, vol. 21, pp. 790–814, 2011.

[21] K. Liu, Y. Yao, D. Sun, and V. Balakrishnan, "Adaptive control design for piecewise-linear systems with paremeter uncertainties," in *Proc. IEEE Conf. on Decision and Control*, (Orlando, USA), pp. 3980–3985, 2011.

[22] M. C. D. Oliveira, J. Bernussou, and J. C. Geromel, "A new discrete-time robust stability condition," *Systems & Control Letters*, vol. 37, pp. 261–265, 1999.

[23] P. Apkarian, H. Tuan, and J. Bernussou, "Continuous-time analysis, eigenstructure assignment, and $H_2$ synthesis with enhanced linear matrix inequalities characterizations," *IEEE Trans. Automatic Control*, vol. 46(12), pp. 1941–1946, 2001.

[24] W. Shi and D. Zhang, "Adaptive robust control of linear motor with ripple force compensation," in *3rd Pacific-Asia Conference on Circuits, Communications and System*, (Wuhan, China), pp. 1–4, 2011.

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# Polynomial Accelerated Algorithm Base on Minimizing TV for Computerized Tomographic Image Reconstruction

Hui Kang,Hongxia Gao
Engineering Research Center for
Precision Electronic Manufacturing Equipments
of Ministry of Education
College of Automation Science and Engineering
South China University of Technology
GuangZhou, China
Email: hxgao@scut.edu.cn ,spiritcherry@126.com

Yueming Hu
Engineering Research Center for
Precision Electronic Manufacturing Equipments
of Ministry of Education
College of Automation Science and Engineering
South China University of Technology
GuangZhou, China
Email: auymhu@scut.edu.cn

***Abstract*—Based on total variation minimization (TV) algorithm and the theory of polynomial acceleration, we prompted a modified algorithm, which is called P-TV in this paper. This new algorithm has a better noise immunity and fewer iteration numbers than the traditional TV algorithm. P-TV algorithm can be divided into four steps. Compared to the traditional TV algorithm, the P-TV algorithm applies the theory of polynomials to ART iteration step to enforce data consistency with the projection data and applies the positivity constraint after the GRAD-step too, which accelerate convergence of image reconstruction. Simulation results prove that when the iteration number is the same, the quality of reconstructed image from P-TV algorithm is better than traditional TV algorithm. When the quality of image reconstructed by both algorithms is nearly the same, the iteration number of P-TV algorithm is much less than the number of traditional TV algorithm. At last we applied the two algorithms to CT image reconstruction from the noisy data. All results show that the P-TV algorithm is effective.**

***Index Terms*—CT; Few-views; Image reconstruction; TV algorithm; Polynomial acceleration algorithm**

## I. Introduction

Computed Tomography (CT) image reconstruction is widely used in medicine, industrial inspection and so on. It is an imaging technique to get the cross-sectional information of the image according to projection measurement data which are in different angles. In various forms of computerized tomographic image reconstruction, one of the main issues centers on how to estimate an accurate image from few views data. In two dimensional (2D) CT, the most common problems are the non-sufficient data reconstruction problems[1][2]. Here, we focus on the few-views problem, which obtains the projection data only in some specific angles. Two widely used iterative algorithms for tomographic reconstruction are the algebraic reconstruction technique (ART) [3]and the expectation-maximization (EM) algorithm[4]. For the case where the data are not sufficient to determine a unique solution to the imaging model, the ART algorithm will find the imagethat is consistent with the data and minimizes the sum-of-squares of the image pixel values. But this algorithm will lead to conspicuous artifacts in reconstructed images. Although there are many improvements based on ART, such as SIRT and MART, the effect on elimination of artifacts is not so obvious. In the last few years, total variation minimization (TV) methods, which originates in the field of compressing sensing (CS)[5][6], for CT reconstruction from sparse and noisy data were developed by Sidky et al [7]. This method is effective on artifacts elimination; we can get a high quality reconstruction image with it. Li Yi of North University of China and Liu Baodong of Chongqing University of China both make some modifications about TV algorithm [8][9]. But traditional TV algorithm is time consuming. In this paper, we develop an iterative image reconstruction algorithm, which is combined TV algorithm with theory of polynomial acceleration[10]. We call it P-TV algorithm which accelerates convergence of image reconstruction. Each iteration of P-TV algorithm can be divided into four steps: the POCS-step (Projection onto Convex Sets), which applies the theory of polynomials to ART iteration to enforce data consistency with the projection data; the positivity constraint step, which constrains the values of all reconstructed points within [0, 1]; the GRAD-step, which reduces the TV of the image estimate; the positivity constraint step again, this step is added to the conditional TV algorithm additionally, which can accelerate convergence of image reconstruction.

This paper is organized as follows. In Section2, we describe the CT reconstruction problem and the TV algorithm. In Section 3, we introduce the theory of polynomial acceleration and promote the modified TV algorithm. In Section4, we give the numerical simulation results. In Section5, we conclude this paper.

## II. TV ALGORITHM

### *A. Mathematic Model of CT reconstruction*

The image is described by a vector $\vec{f}$ of length $N_{image}$ with elements$f_j, j = 1, 2N_{image}$, $N_{image} = n \times n$ where the

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

integer n is the width and height of the 2D image array. The projection-data is described by a vector $\vec{p}$ of length $N_{data}$with elements $p_i$

$$p_i = \sum_{j=1}^{N_{image}} w_{ij} f_j, i = 1, 2, \ldots, N_{data} \tag{1}$$

$N_{data}$ is the total number of projection-data. $w_{ij}$ is the element of the system matrixrepresent the impact of pixel $f_j$ to projection-data $p_i$ and the value of $w_{ij}$ is equal to the length of ith projection through the jth pixel[11]. The system matrix W is composed of $N_{data}$ row vectors. The general expression for the algorithm discussed here involves inversion of a discrete-to-discrete linear transform

$$\vec{p} = W\vec{f} \tag{2}$$

### B. TV algorithm

CT reconstruction always is an ill-posed problem, especially when it is few-views condition. The typical method to solve this problem is the regularization. Recently, regularization based on minimizing TV of the image becomes a hot topic[7][12-15].In fact, minimizing TV of the image is a progress to minimize the $l_1$-norm of the gradient image. So, the most elements of gradient image will be zeros and the original image will tend to constant in most pixels [16]. Each iteration of TV algorithm consists of two phases: POCS and gradient descent. The POCS phase is further broken down into two steps that enforce data consistency and positivity. As a result, the steps comprising each loop are: DATA-step, which enforces data consistency with the projection data; the POS-step, which ensures a non-negative image; and the GRAD-step, which reduces the TV of the image estimate. We can refer to references[7] for concrete steps of TV algorithm. During the process of DATA-step, TV algorithm employs the traditional ART iteration formula to enforce data consistency. And in the process of GRAD-step, parameter  is a constant, which is used to control the convergence speed. So, in the beginning of iteration, $f^{(TV\_GRAD)}$ will fall off rapidly to certain value. But with the increase of iteration, $f^{(TV\_GRAD)}$ will fall off slowly. Therefore, we prompted a modified TV algorithm, called it P-TV based on TV algorithm and the theory of polynomial acceleration, which will introduce below. P-TV algorithm has a better performance in convergence speed than TV algorithm. This performance will be discussed in Section 4.

## III. THEORY OF POLYNOMIAL ACCELERATION

The traditional iteration method just used the information of image in last step, such as ART. The theory of polynomial acceleration will use all the information in each iteration[10]. The general formula is as follows:

$$F^k = u_1 F^{k-1} + u_2 F^{k-2} + \ldots + u_k F^0 + w_k r^{k-1} \tag{3}$$

$$k = 1, 2, \ldots, u_1 + u_2 + \ldots + u_k = 1, w_k \neq 0$$

$r^{k-1}$ is remainder term. We can obtain the following formula from (3)

$$F^k - F^0 = \sum_{j=0}^{k-1} \varepsilon_j (W^* W)^j r_0 = q^{k-1}(W^* W) r_0 \tag{4}$$

$$q^{k-1}(\gamma) = \varepsilon_0 + \varepsilon_1 \gamma + \ldots + \varepsilon_{k-1} \gamma^{k-1} \tag{5}$$

$q^{k-1}(\gamma)$ is a polynomial of K-1 order and

$$q^{k-1}(\gamma) = \sum_{j=0}^{k-1} (1-\gamma)^j = \frac{1-(1-\gamma)^k}{\gamma} \tag{6}$$

In order to use polynomial iteration effectively, promoting an algorithm just used the information of $F^{k-1}$ and$F^{k-2}$.

$$F^k = F^{k-1} + u_k (F^{k-1} - F^{k-2}) + w_k W^* (P - W F^{k-1}) \tag{7}$$

$$u_k = \frac{(k-1)(2k-3)(2k+2v+1)}{(k+2v-1)(2k+4v-1)(2k+2v-3)} \tag{8}$$

$$w_k = \frac{(k+v-1)(2k+2v-1)}{(k+2v-1)(2k+4v-1)} \tag{9}$$

We combined this theory with TV algorithm, the improved algorithm of TV are as follows:

*1) Initialization:*

$$f^{(0)} = 0, \alpha = 0.2, N_{grad} = 20, v = 0.8$$

*2) Data projection iteration, for $k = 1, 2, \ldots, N_{data}$:*

$$f^{(k)} = f^{(k-1)} + u_k (f^{(k-1)} - f^{(k-2)}) + w_k \frac{p_i - w_{ij} f^{(k-1)}}{\sum_{j=1}^{N_{image}} w_{ij}^2} w_{ij} \tag{10}$$

$$i = 1, 2 \ldots, N_{data} \; j = 1, 2 \ldots, N_{image}$$

$$u_k = \frac{(k-1)(2k-3)(2k+2v+1)}{(k+2v-1)(2k+4v-1)(2k+2v-3)} \tag{11}$$

$$w_k = \frac{(k+v-1)(2k+2v-1)}{(k+2v-1)(2k+4v-1)} \tag{12}$$

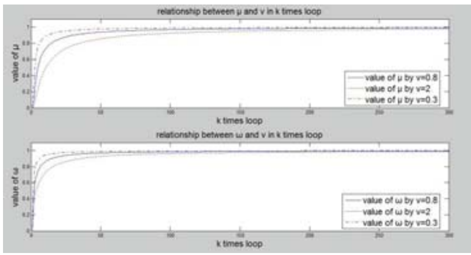


Fig. 1. tendency of $u_k$ and $w_k$ with the addition of iteration number

Fig.1. show the tendency of $u_k$ and $w_k$ with the addition of iteration number. v is a constant which controls the convergence rate of iteration. We can see from Figure, different value of v will lead to different convergence rate about $u_k$ and $w_k$ .Furthermore, different $u_k$ and $w_k$ will lead to different convergence rate in reconstruction iteration. Here, we set v=0.8 according to lots of experiments.

*3) Positivity constraint:*

$$f_{pose}^{(k)} = \begin{cases} 1, & f_j^{(TV\_GRAD)} > 1 \\ f_j^{(k)}, & 0 < f_j^{(TV\_GRAD)} < 1 \\ 0, & else \end{cases} \tag{13}$$

$$j = 1, 2 \ldots, N_{image}$$

*4) TV gradient descent initialization:*

$$d_A(n) = \|f^{(k-1)} - f_{pose}^{(k)}\| \tag{14}$$

$$f^{(TV\_GRAD)} = f_{pose}^{(k)} \tag{15}$$

*5) TV gradient descent for $m = 1, 2, \ldots N_{grad}$:*

$$v_j^{(m)} = \frac{\partial \|f^{(TV\_GRAD)}\|_{TV}}{\partial f_j^{(TV\_GRAD)}} \tag{16}$$

$$f^{(TV\_GRAD)} = f^{(TV\_GRAD)} - \alpha d_A \frac{v_j^{(m)}}{\|v_j^{(m)}\|} \tag{17}$$

*6) Positivity constraint:*

$$f_{pose}^{(TV\_GRAD)} = \begin{cases} 1, & f_j^{(TV\_GRAD)} > 1 \\ f_j^{(TV\_GRAD)}, & 0 < f_j^{(TV\_GRAD)} < 1 \\ 0, & else \end{cases} \tag{18}$$

*7) Stop condition:* The iteration will stop when there is no appreciable change in following formula, else return to 2).

$$|f_{pose}^{(k)} - f_{pose}^{(k-1)}| < \varepsilon \tag{19}$$

## IV. NUMERICAL SIMULATION RESULTS

The true image solution is taken to be the Shepp-Logan image shown in below discredited on a 128*128 pixel grid. This phantom is often used in evaluating tomographic reconstruction algorithms. We take parallel beam in experiment and scanning parameters are given in table 1.

TABLE I
SCANNING PARAMETERS

| Parameters | Values |
|---|---|
| Image size(pixel) | 128*128 |
| Projection group number | 18(every $10^o$) |
| Detector number | 185 |
| Angular range | $[0, \pi]$ |

For comparison, the images are reconstructed by use of TV algorithm and P-TV algorithm. Fig.2. Show the reconstructions results from noiseless data. It is shown clearly when the iteration number is same (both are 39), the reconstruct image quality of P-TV is better than TV.

For further comparison, we compare the image profiles [8] of line 50 and line 103 as fig.3. We can see image reconstructed by P-TV algorithm is closer to the original image than image reconstructed by TV algorithmespecially in some peak value. In order to show the P-TV algorithm's good performance on convergence speed, we show a group of pictures as below. Picture (b) is obtained by P-TV algorithm, picture (c) is reconstructed by traditional TV algorithm. We can see that the quality of picture (b) and picture (c) is nearly the same, but the iteration number of picture (c) is 305 while the iteration number of picture (b) is only 39. That is to say, P-TV algorithm's convergence rate is nearly ten times as much as TV algorithm.

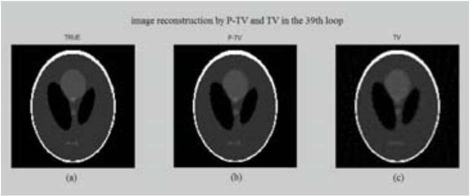


Fig. 2. Reconstruct Image (a) original image; (b) reconstructed image by P-TV algorithm; (c) reconstructed image by TV algorithm

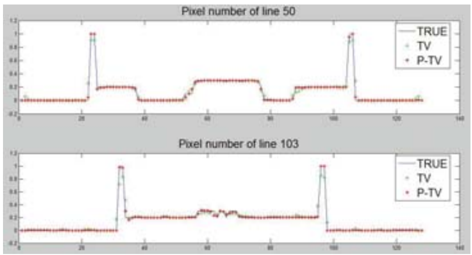


Fig. 3. Profiles (a) original image; (b) reconstructed image by P-TV algorithm

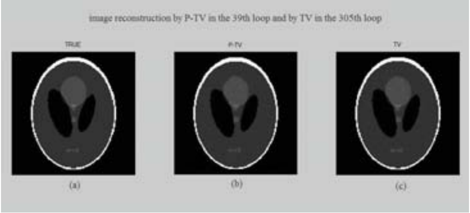


image reconstruction by P-TV in the 39th loop and by TV in the 305th loop
TRUE
P-TV
TV
(a)
(b)
(c)

Fig. 4. Reconstruct Image (a) original image; (b) reconstructed image by P-TV algorithm; (c) reconstructed image by TV algorithm

To verify the algorithm's stability for noise, we have also applied the two algorithms to noisy data generated by adding Gaussian noise to the projection data. The standard deviation of the Gaussian noise is $0.1\%$ of the maximum value of the projection data. The results are shown in Fig.5. It indicates that P-TV algorithm has better noise immunity than TV algorithm.

We also show the mean square error (MSE) of the reconstructed images used two algorithms and give them in Fig.6. We can see P-TV algorithm converge faster than TV algorithm and MSE of P-TV is smaller.

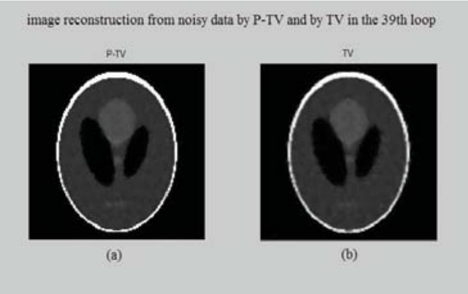


image reconstruction from noisy data by P-TV and by TV in the 39th loop
P-TV
TV
(a)
(b)

Fig. 5. Reconstruct Image with Gaussian noise: reconstructed image by P-TV algorithm; (b) reconstructed image by TV algorithm

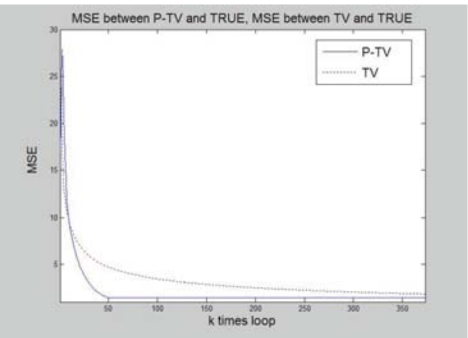


Fig. 6. Reconstruct Image with Gaussian noise: reconstructed image by P-TV algorithm; (b) reconstructed image by TV algorithm

## V. Conclusion

In this paper, a P-TV algorithm based on TV algorithm and the theory of polynomial acceleration is prompted. The experimental results showed evidence that the algorithm is effective and stable. And the most prominent advantages of this algorithm are the rapid convergence speed and the less time consuming. This algorithm can be applied to 3D CT configurations without or with small modifications.

## Acknowledgment

This work supported by the Natural Science Foundation of China under Grant No. 61040011 and 60835001, also funded by the Fundamental Research Funds for the Central Universities under Project No. 2012ZZ0107.

## References

[1] L.Li, K. Kang, Z. Chen, L. Zhang and Y. Xing, A general region-of-interest image reconstruction approach with truncated Hilbert transform, Journal of X-Ray Science and Technology 17 (2009), 135-152.

[2] H.Yu, Y. Ye and G. Wang, Interior reconstruction using the truncated Hilbert transform via singular value decomposition, Journal of X-Ray Science and Technology 16 (2008), 243-251.

[3] G.T. Herman, Image Reconstruction from Projection: the Fundamentals of Computerized Tomography, Academic Press, New York, 1980.

[4] H.H. Barrett and K.J. Myers, Foundations of Image Science, John Wiley Sons, Inc. Hoboken, New Jersey, 2004.

[5] E. Candes, J. Romberg and T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete Frequency Information, IEEE Trans Inf Theory 52(2)(2006),489-509.

[6] E. Candes, J. Romberg and T. Tao, Stable signal recovery from incomplete and inaccurate measurements, Commun Pure Appl Math 59(7) (2006), 1207-1223.

[7] E.Y. Sidky, C. Kao, and X. Pan. Accurate image reconstruction from few-views and limited-angle data in divergent-beam CT[J].Journal of X-Ray Science and Technology,(2006),14:119-139.

[8] Baodong Liu. A Thesis Submitted to Chongqing University in Partial Fulfillment of the Requirement for the Degree of Doctor of Engineering. [D](2010). Chongqing University of China.

[9] Li Yi. The study of limited angle three-dimensional CT image reconstruction algorithm. (2011), North University of China.

[10] Yu Chen. Research on Inverse Problems Solving and Image Reconstruction Algorithm For Electrical Capacitance Tomography System. (2010), Harbin University of Science and Technology of China.

[11] TianGe Zhuang. Principle and Algorithm of CT. Shanghai Jiaotong University press, 1992.

[12] Sidky E Y and Pan X. Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization [J].Phys.Med.Biol. 2008, 53:4777-4807.

[13] Sidky E Y and Pan X. Accurate image reconstruction in circular cone-beam computed tomography by total variation minimization: a preliminary investigation [J].IEEE Nucl. Sci. Conf.Rec.2006, 5:2904-7.

[14] Sidky E Y,Chartrand R and Pan X. Image reconstruction from few views by non-convex optimization [J].IEEE Nucl. Sci. Conf. Rec., 2007,5:3526-30.

[15] LaRoque S J,Sidky E Y and Pan X. Accurate image reconstruction from few-view and limited-angle data in diffraction tomography [J].Opt.Soc.Am.A,2008,25:1772-82.

[16] Gengsheng Zeng. Medical Image Reconstruction [M].Beijing: Higher Education Press, 2010

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# CAN Based Unified Customizable Diagnostic Measure Research And Realization

Yan Song[1,2], Tianran Wang[1], Aidong Xu[1],
1.Shenyang Institute of Automation,
Chinese Academy of Sciences.
Shenyang 110016,China.

Kai Wang[1] and Zhijia Yang[1]
2.Graduate University of
Chinese Academy of Sciences.
Beijing 100039, China

*Abstract*— **To diagnostic vary CAN bus based network, an unified customizable diagnostic measure is introduced. The measure is based on customizable RPN format protocol expression and CAN UDS, support most of CAN bus based network. Not only decoding or other normal diagnostics are supported, but also network activity is analyzed, too. Furthermore, the design of customizable diagnostic software is given.**

***Keywords-UDS;RPN;CAN bus Diagnstics;ISO 14229***

## I. INTRODUCTION

CAN based fieldbuses have been used for years and different upper application protocols were produced, such as CANOpen, ISO Bus, SAEJ1939 and DeviceNet, and so on. To apply and maintain those fieldbuses, protocol analysis tool are overwhelming needed. Early detection of network failures and performance degradations is a key to rapid fault recovery and robust networking [1]. Though fieldbus mentioned above are based on CAN bus, the great gap between their application layer make it was very difficult to diagnostic all of them in one application platform. Automotive network ECU manufactured by different vendors is different [2], so a unified customizable diagnostic tool is important to auto diagnostic, too.

CANOE or other software can do this job but the huge invest and not very clearly decode explains are the fatal weakness. There are vary diagnostic tools exist in the CAN bus application. Different diagnostic tools lead to more development and maintenance cost, and for some scenes it is too inconvenient.

In road vehicle application scope, a series of Unified Diagnostic Services (UDS) standards have been made by ISO, its purpose is to make diagnostic more convenient and less development repetitious. ISO 15765 and ISO 14229 have achieve CAN based UDS[3], there should be only one diagnostic protocol theoretically, but it is applied almost only in vehicle area in fact. The CAN based UDS most likely can be used in vehicle-related application scenes.

A CAN based customizable diagnostic method (CCD-Measure) is introduced in this article, furthermore, a customizable diagnostic tool aim on CAN based fieldbus is realized. This CAN bus based customizable diagnostic tool (CCD-Platform) support ISObus, SEAJ1939[4][5], CAN UDS, etc., the most significant is customizable application layer analysis function. Any CAN based protocol nevertheless normal or private can be decoded, explained and analyzed well. Only if a clear definition is provided.

## II. UNIFIED DIAGNOSTIC MEATURE

CCD-Measure introduced in this article is customizable and support CAN UDS, the former make it is possible that CCD-Platform can be used in non-vehicle area such as CANOpen, DeviceNet, meanwhile the latter make sure that most newly designed road vehicle ECU and network can be analyzed and diagnosed by this platform.

### *A. CAN Bus based Unified Diagnostic Services*

CAN UDS is proposed in two related standards, ISO 15765 and ISO 14229. It is designed for road vehicle diagnostics, but it can be used in other area such as farm machinery, construction machinery, etc. Some ECU (electronic control unit, a control device on vehicle bus) manufactures has support the CAN UDS such as Bosch, Perkins and so forth. CAN UDS is an independent upper protocol, support 25 services including “ECURese”t, “TesterPresent” and “ReadDataByIdentifier”, etc. ISO 14229 is an “application layer” in the UDS architecture and ISO15765 part 3 is a simplify version of ISO 14229, both of them define the diagnostic services.

TABLE I. CAN UDS & RELATED OBD[3]

| OSI model | CAN UDS | OBD |
|---|---|---|
| Application Layer | ISO15765-3/ISO14229 | ISO15031-5 |
| Presentation Layer | N/A | N/A |
| Session Layer | N/A | N/A |
| Transport Layer | N/A | N/A |
| Network Layer | ISO 15765-2 | ISO 15765-4 |
| Data Link Layer | ISO 11898-1 | ISO 15765-4 |
| Physical Layer | ISO 11898-1 | ISO 15765-4 |

As Shown in table 1, ISO 15765 support both CAN UDS and old OBD diagnostic method, OBD mean “On Board

Diagnostic", it was a legislated measure for road vehicle diagnostic. In this article we concern about CAN UDS only. The CAN UDS has no Presentation layer, Session layer and Transport layer, both data link layer and physical layer make use of ISO 11898 which is CAN bus international standards.

ISO 15765 part 1 is a general specification for whole CAN UDS, it introduces remain parts of CAN UDS and gives some important concept in CAN UDS such as architecture and related standards. Part1 shows the overview and relationship of different parts. ISO 15765 part 2 gives a general specification of UDS network layer, which was used to diagnose device across different network, such as from power train bus to implement bus. Both single frame (non-segment frame) and multiple frames (segment frame) are introduced. SF is used to report some immediate error code and if one frame is not sufficient for the message, it will be divided several different frame and send under the Flow Control Mechanism's control. To the residual error codes it needs to be report in multiple frames, as shown in figure 1.

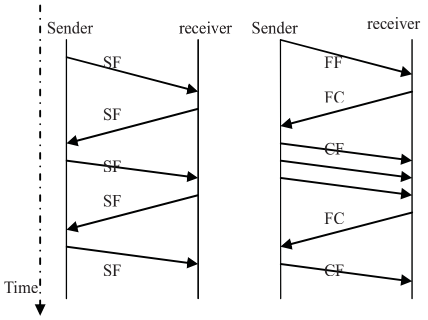


Figure 1. Single Frame and MultiFrame

ISO 15765- part 3 is a simplified ISO 14229, more detailed information is described more clearly in latter. It is a core part of CAN UDS, most application services used by diagnostic are produced in this part. As mentioned early, there are mainly 25 (in ISO 14229 25, ISO 15765 24) services of diagnostic, can be divided into diagnostic and communication management functional unit, data transmission functional unit, stored data transmission functional unit, I/O control functional unit and upload/download function unit. ISO 15765 part 4 describes the OBD diagnostic requirement and it is useless in CAN UDS.

### *B. Choosen Essential Parameters of Customizable Measure*

As mentioned above, a customizable diagnostic measure is introduced and not only for road and non-road vehicle but also for automation application such as DeviceNet, CANOpen, etc. CAN based customizable diagnostic measure (CCD Measure) make use of XML to customizable information. The customizability is based on formal method[7][8]. Although fingerprinting[7] is an useful information, but in this article it wouldn't be concerned.

A protocol is a series of statutes to transfer data correctly and efficiently. Protocol means "A set of standardized procedures for transmitting or storing data, especially those used in regulating data transmission between computers or peripherals."[7] From the view of analysis and diagnostic, an application protocol can be described in four metrics, as shown in formula (1), a protocol can be defined by Format, Error message, Timing and Action. To perform diagnostic, extra information helpful for analysis is Statistic parameters and Other diagnostic information, as shown in formula (2), where F mean Format, T mean timing, S mean Statistics and OD mean Other Diagnostic parameters.

$$P = \{F, E, T, A\} \tag{1}$$

$$D = \{F, T, S, OD\} \tag{2}$$

Format parameters describe structure information of protocol, such as how much fields there are and how many bytes every fields occupied. Vary protocols may have same identifier but different format or same format but different identifier. To CAN bus (2.0B) based network, there are always 29 ID and 8 bytes data, but the explanation were very different from vary protocols.

An application protocol may have various application protocol data unit (A-PDU), but all of them may have same basic format named Root Format, and other kinds of PUD inherent from the Root Format PDU. The hierarchy diagram is shown as figure 2, the root of all format is named Root Format, which may include some important and basic coding rule, such as CAN ID's explanation, it describes an application protocol's basic structure. Level1 Format inherent from root format and can be treated as a "primitive" level, L2 format is service level. The hierarchy between different level of PDU format may achieved by XML.

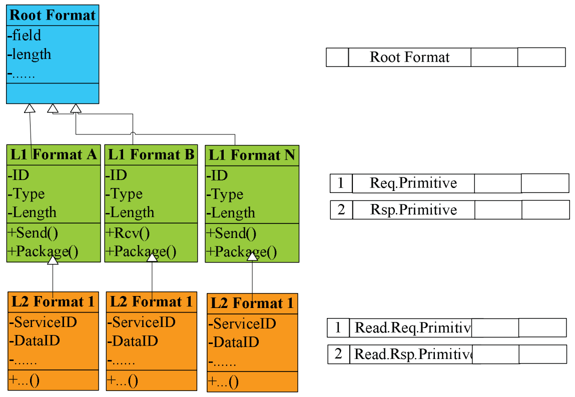


Figure 2. PDU Format Hierarchy

Timing parameters is a time line profile of communication,. When a protocol working, timing parameters describe all the outer activities between different communication peers, which services PDU to be sent, what time and which one response PDU should be sent are defined by timing parameters.

Statistics Parameters are important attributes to diagnostic and analysis of protocol, but not core attributes of protocol definition. Typical statistics parameters are bus load rate, error rate, certain application layer service occurs count, retry count, etc.

Other diagnostics parameters are group of diagnostics parameters which user interested or needed. Those parameters depend on the application protocol to be analyzed. For example, ISO bus has a "Parameter Group Number" (PGN), different

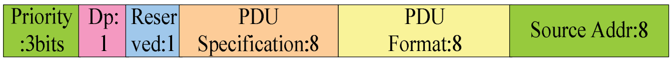
PGN respect some physical value, such as ground speed, wheel speed. Some PGN's occur ratio may be very important for analysis of ISO bus.

### C. Customizable Diagnostic Measure Design

According to (1), and because F, E, T and A are a type of set also, there are the following formulas:

$$\begin{cases} F=\{f_1,f_2,\ldots,f_m\} \\ E=\{e_1,e_2,\ldots,e_n\} \\ T=\{t_1,t_2,\ldots,t_n\} \\ A=\{a_1,a_2,\ldots,a_n\} \end{cases} \quad (3)$$

To illustrate what are "F, E, T, A" look like exactly, a SAE J1939 basic profile is given following as an example. SAE J1939 is a supported protocol by this tool and basic PDU format of SAE J1939 is given as follow picture.



| Priority :3bits | Dp: 1 | Reserved:1 | PDU Specification:8 | PDU Format:8 | Source Addr:8 |
|---|---|---|---|---|---|

Figure 3. Basic pdu format of SAE J1939

As figure.3 shown, a pdu format concerned field type, field name, field length, and so on. Filed name can be treated as variables, such as pri (priority), dp (data page), length can be handled as coefficients, such as 3, 1, etc. According mentioned above, figure3 can be translated an algebraic format as $f_1$, formula (4). Formula (4) $e_1$ is an "error" element of E formula (1), which means negative acknowledgement with error code "-1" and address "error_addr". Formula (4) $t_1$ means a request primitive can be responded by a response primitive.

$$\begin{cases} f_1 = 8*pri+1*dp+1*r+8*pf+8*ps+8*sa \\ e_1 = -1*ack*error\_addr \\ t_1 = 100 \\ a_1 = req+rsp(0) \end{cases} \quad (4)$$

This formula will be converted into RPN (Reverse Polish Notation) to deal with computer conveniently. The transform will be done by computer program, thus the xml input would look like formula (4), but latter will be changed into RPN format.

$$\begin{cases} f_1 = 8pri*1dp*1r*8pf*8ps*8sa*+++++ \\ e_1 = -1ack*error\_addr* \\ t_1 = 100 \\ a_1 = reqrsp0*+ \end{cases} \quad (5)$$

Formula (4) and (5) shows a little part of prime definition of protocol SAEJ1939. The $f_1$ is an algebraic form of figure3 PDU format.

Protocol system fault detection is often conducted by active testing [9] and action analysis. To diagnose the network and each node more clearly, three sets are made, they are normal set, abnormal set and error set. Normal set includes protocol information provided by formula (1), at last it will turn into RPN format like formula (5). Normal set can be predefined by user, it is easy to develop vary protocol's RPN definition according mentioned above. When RPN format protocol definition has been fulfilled, empty abnormal set and error set should be established meanwhile. When it is first run, all traffic on the net work should been captured and analyzed according to RPN protocol definition in normal set. If the traffic is consistent with normal set $P_1$,$P_2$, or $P_i$, it should be a "right " or "normal " action, otherwise, it is an abnormal or error action and the action will be put into abnormal set. An abnormal action will be treated as error action only when user made it. Both abnormal set and error set can be preserved for future. Normal, abnormal and error sets' relationship is shown in following picture.

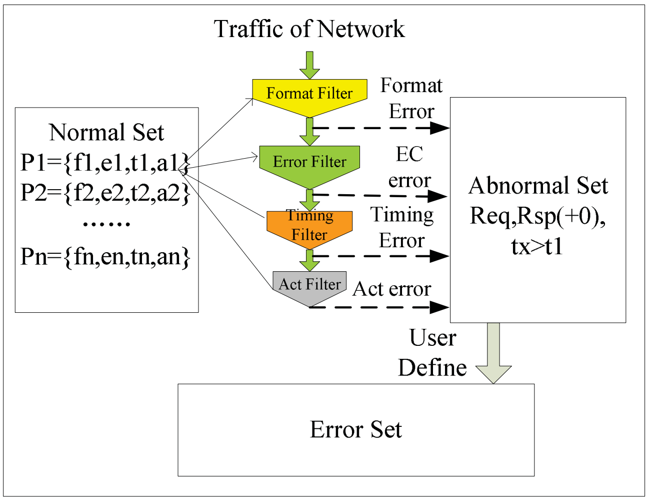


Figure 4. Normal , abnormal and error set

As figure.4 shown, traffic of network will be analyzed according to rules of normal set, compare with $P_i$ (i=1…n) until match or finish, if $f_i$, $e_i$, $t_i$ or $a_i$ is not matched, compare with $P_{i+1}$ , if three of four elements are matched but still not all matched with $P_i$, put it in abnormal set with $P_i'$ (an abnormal $P_i$) otherwise put it as a new element. If user picks up an element of abnormal set and let it belong to error set, then it will be moved into error set.

Customizable characters are reflected in translate protocol definition to RPN formula approach, by this way any protocol can be defined clearly and easily.

### D. Customizable diagnostic meature model

As mentioned above a protocol's analysis is defined by 4 metric essential parameters named F,T, S and OD. A customizable diagnostic measure model is given in this part. Formal Protocol Rule Interpreter (FPRI) is a core concept in CCD measure, it should take 4 metric parameters namely F, T, S and OD.

Customizability is achieved by the FPRI and some Formal Protocol Describe File (FPDF), according to protocol 4 metrics attributes, there are 4 FPDF configuration files. FPRI translate protocol definition to RPN format like formula (5). The four files and other needed files can describe a certain protocol as well as pre-programmed one. FPRI read and interpret the former files to parameters in four metrics, the interpreted parameters are sent to runtime general protocol analysis. The runtime general analysis handle unpackaged raw data stream and explain PDU to physical information. Explained PDU and corresponding physical information are sent to display model, by which list physical information of PDU, unpackaged PDU data, timing information, statistics information and diagnostic information, as shown in figure 5.

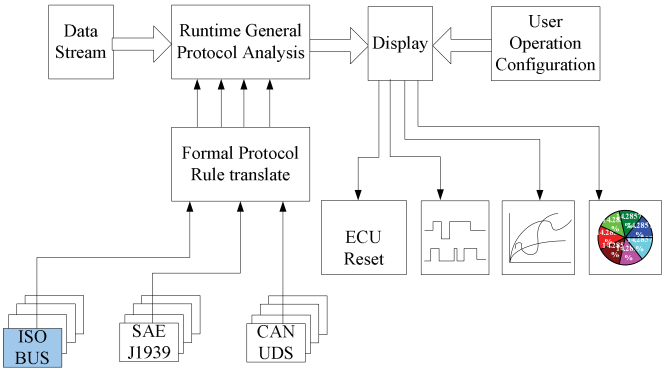

Data Stream
Runtime General Protocol Analysis
Display
User Operation Configuration
Formal Protocol Rule translate
ECU Reset
ISO BUS
SAE J1939
CAN UDS

Figure 5. Customizable diagnostic measure mechanism flaw

Runtime general protocol analysis module is responsible for semantic retrive, which means data mapping to semantic. The physical and diagnostic information are created online according to runtime created semantic interpretive rules (according to configure file FPDF). User operation configuration module provides a configuration interface to user, by which operator may decide what kind of data should be displayed and how to do it.

## III. The Design of Unified Diagnostic Platform

A customizable unified diagnostic platform consist of definitions, specifications, data structures, program modules and inter relationship between them.

### A. *The Architecture Design and Analysis*

This platform can be divided into the following parts: CAN bus access device, basic software module, application software module. KVASER leaf light acts as CAN bus access device, cause of it is minimal in physical size and has sufficient supports for developer. Basic software is consisted of driver module, formal protocol rule interpreter, the application module include runtime protocol analysis module, display module and user configure module. As shown in figure 4, the platform is divided into several parts.

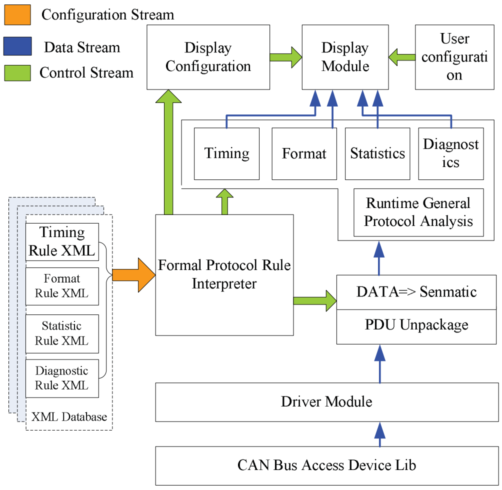


Figure 6. Unified Customizable Diagnostic platform's model structure

As figure 4 shown that blue lines means "data stream", data stream comes from CAN bus and process step by step in our platform and become into some analysis/diagnostic data finally. Salmon pink lines means "configuration stream", those configuration come from user defined FPDF described by XML, in which timing, format, statistics and diagnostics restrict are given. Meanwhile the green lines mean "control stream", the control information define related module's activities.

CAN bus access device lib and driver module provide an interface to access real CAN bus, by which data from CAN bus were captured in time. They are the glue between CAN bus access hardware and the upper application.

Formal protocol rule interpreter is respond to interpret FPDF XML to computer program, the source is formatted nature language or diagnostic formula, and the destination is computer program or corresponding parameters. Formal protocol rule interpreter take one input and produce three output streams.

XML data base including four kinds of FPDF, it provide rule staff to formal protocol rule interpreter.

### B. *Platform Software and Its Deployment*

This software platform is developed by C++ builder and EVC, the CAN bus access tool is KVASAR leaf light. The software's interface is shown as Figure 7.

The platform software can decode and diagnose different CAN based protocol concurrently, as figure 7 shown there are three kinds of protocol named SAEJ1939, ISO Bus and UDS are decoded, analyzed and diagnosed. These three protocols have same root format, meanwhile there are some differences at format hierarchy level 1 and level 2 (figure2). The platform software has decoded J1939 TP.DT, UDS SF and UDS FF frame. The left top area gives a detail diagnostic of UDS, the left middle area decoded SAEJ1939-73's diagnostic information. Because of CAN based protocols have same root PDU format, the decoding and diagnostic program can be integrated together.

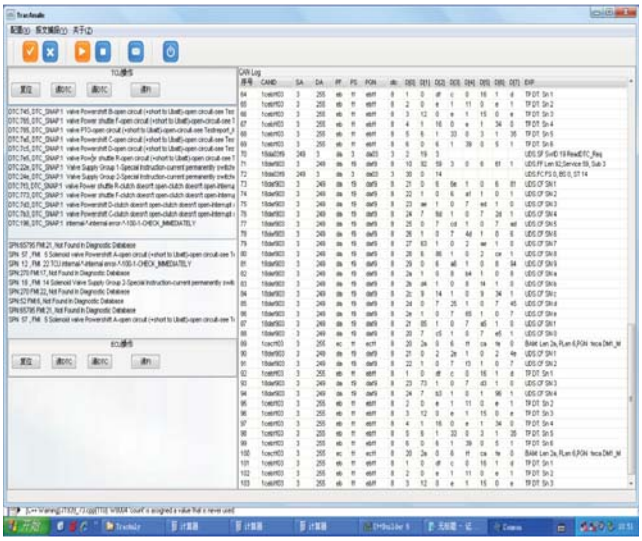


Figure 7. CAN based Customizable diagnostic Platform

The Platform software can be easily applied in Tractor to diagnostic, the application approach is shown as following. Both PC and Wince embedded devices can be make use of.

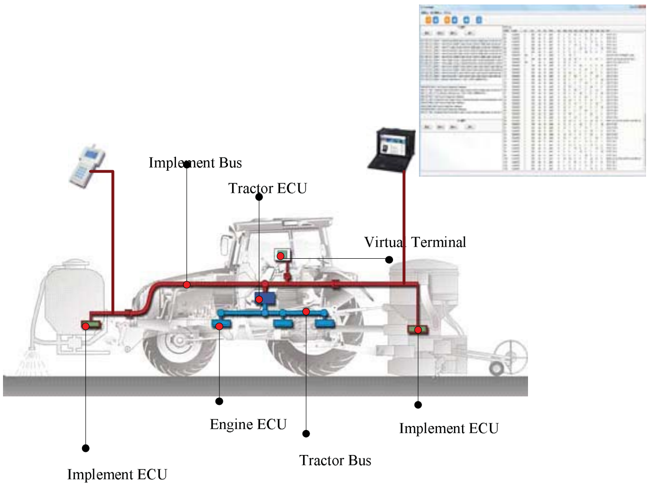


Figure 8. CAN based Customizable diagnostic Platform

The Platform software is applied in 200 horsepower wheeled tractors and 300 horsepower wheeled tractors of YTO Group Corporation, which was biggest tractor and argiculture machine manufacture in China.

## IV. The Design of Unified Diagnostic Platform

A customizable diagnostic of CAN bus based network analysis and diagnostic platform software is designed and realized. Customizable diagnostic parameters are given, furthermore, a customizable diagnostic measure based on protocol RPN definition is introduced.

This platform software can be defined easily and clearly by user, such as SAEJ1939 and ISO bus or other CAN bus based networks. The platform software not only decode or analysis every frame on network, but also spy every challenge-response communication action. And based on the traffic action statistics parameters, a network status can be defined and provide more detail information to user. This platform software was used at YTO Group Corporation.

## Acknowledgment

This work is supported by the Natural Science Foundation of China under contract 61004068.

## References

[1] Jun Jiang, “A network fault diagnostic approach based on a statistical traffic normality prediction algorithm”, Global Telecommunications Conference, 2003. GLOBECOM '03. IEEE,Vol: 5, pp: 2918 - 2922,2003.

[2] Suwatthikul, J.,"Automotive Network Diagnostic Systems", Industrial Embedded Systems, 2006. IES '06. International Symposium on, pp: 1 - 4,Oct. 2006.

[3] ISO, ISO 14229 "Road vehicles — Unified diagnostic services (UDS) — Specification and requirements",ver 2nd,Switzerland,2006.

[4] SAE, "Data link layer", SAE, J1939-21, rev ARP 2001, 2001.

[5] SAE, "Network layer", SAE, J1939-31, rev ARP 2004, April 2004.

[6] Tianlong Gu, Guoyong Cai, “Formal analysis and design of network protocols” 1st ed, Publishing House of Electronics Industry, Beijing, 2003.

[7] Rudin, H. ,” An informal overview of formal protocol specification”, Communications Magazine, IEEE, Switzerland, Vol: 23, pp: 46 – 52, 1985.

[8] Guoqiang Shu ,”A Formal Methodology for Network Protocol Fingerprinting”, Parallel and Distributed Systems, IEEE Transactions on, USA, Vol: 22, pp: 1813 – 1825, Nov. 2011.

[9] Lee, D. ,"Network protocol system monitoring-a formal approach with passive testing", Networking, IEEE/ACM Transactions on, USA, Vol: 14, pp: 424 - 437,April 2006.

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# *Implementation of Network Control Systems for Improvement of Traditional Dyeing Processes*

Jidai Wang，Aiqin Sun，Dongyue Zhang
School of Mechanical and Electronic Engineering
Shandong University of Science and Technology
Qing Dao 266590, P.R. China
Djdwang8911@sina.com

Yali Xue
Department of Thermal Engineering
Tsinghua University
Beijing 100084, P.R. China
xueyali@tsinghua.edu.cn

*Abstract*—**In China, many traditional industrial processes need to be maintained and upgraded to incorporate new technology development. The paper reports our work in improvement of the traditional dyeing process using advanced network control technology. Based on a dyeing industry, we designed the network system of dyeing equipment using industrial Ethernet & Multi-bus technology. The system takes three-level structure: factory management level, workshop control level and field equipment level. Field equipments comprise some PLC and frequency conversion control systems; workshop level adopts two Siemens PLC to centrally control field equipments; HMI, monitoring room PC and remote office PC monitors and manages workshop control level. The network system has multi-stage monitoring, access control and remote alarm functions, realized management - control integration.**

***Keywords-dyeing equipment; industrial Ethernet; Multi-bus technology; network system***

## I. INTRODUCTION

With IT network technology flying, "one network to the end" industrial system constructed by field bus and Ethernet makes factory senior management personnel can obtain the process information of production directly, realizing highly integration of factory management and production[1]. Dyeing industry is one of traditional industries in China; though the automation level of dyeing equipment has greatly increased, but only be single automation equipment, not be network system. In actual production, dyeing equipments of different types and brands are used; it's difficult to realize centralized management and control, adding cost burden to factory unquestionably. Based on the dyeing equipment's present situation of a factory in Qingdao, the paper designed a network system to realize the centralized management and control of production equipments.

## II. NETWORK SYSTEM STRUCTURE

The network system structure comprises field equipment level, workshop control level and factory management level. 1) Field equipment level consists of all original dyeing automation equipments in the factory, including a set of batching equipment for dye ingredients, a set of wax printing machine, three sets of rotary screen printing machine and a set of variable frequency ventilation equipment. Siemens S7-200 CPU226 is the controller of batching equipment; rotary screen printing machines' controllers all are Siemens S7-200 CPU224; the wax printing machine uses Delta PLC of DVP - ES series as controller; and ventilation equipment's controller is Siemens MicroMaster420 inverter. 2) Workshop control level consists of two Siemens PLC, realizing the configuration control of field devices. 3) Factory management level uses monitoring room PC, office remote PC and on-site HMI to achieve data acquisition, management and monitoring of control units in the workshop level. The hardware structure of network is shown in Fig. 1.

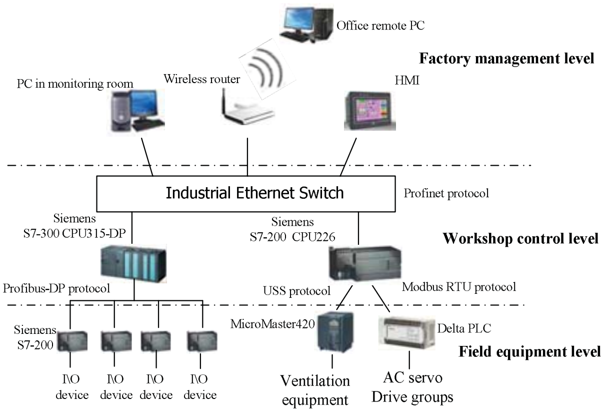


Figure 1. Structure of network system

This network system uses industrial Ethernet and Multi-field bus technology, as shown in Fig. 1. Industrial Ethernet is used for communication of factory management level and workshop control level, and fieldbus is used for communication of workshop control level and field equipment level. For the different control units adopted by dyeing equipments, meanwhile considering equipments' placement, the two sets of S7 PLC are used to realize centralized configuration control of all dyeing equipments in field level. One controller is Siemens S7-300 CPU315-2 DP, which communicates with the batching equipment and rotary screen printing machines using Profibus-dp protocol, form master-slave communication mode. Another controller is Siemens S7-200 CPU226, which communicates with the controller of wax printing machine using the USS protocol, communicates with the controller of variable frequency ventilation equipment using the Modbus RTU protocol.

## III. THE INDUSTRIAL ETHERNET AND FIELDBUS

In the system, factory management level communicates with workshop control level via the PROFINET protocol, which has 100Mbit/s high ethernet transmission speed of IEEE802.3u standard, real-time response time of 10ms, meanwhile can communicate with computer using TCP/IP protocol, the transmission distance reaching 1.5 km [3].

According to different communication functions of control units of dyeing equipments, in the communication of workshop level and field level, the network system adopts Profibus-DP protocol, USS protocol and Modbus RTU protocol. 1) Profibus-DP protocol is based on open Profibus industrial field-bus and an international standard bus used for workshop level and field level. Its transmission speed is 12Mbit/s; its response time is about 1ms. And the interface uses RS-485 mode and shielding twisted-pair cable. 2) USS protocol, which based on serial bus, uses master-slave communication mode, has low hardware cost and reliable message format, and meanwhile includes flexible data transmission. The protocol is suitable for communication of Siemens S7-200 PLC and MicroMaster420 inverter. 3) Modbus RTU protocol, which has become a general standard in the field of industrial control [4], also uses RS485 twisted-pair cable as communication medium. Siemens and Delta PLC both support Modbus RTU protocol, so it is chosen as the data transmission protocol between the S7-200 CPU226 and the DVP-ES series PLC.

## IV. NETWORK SYSTEM COMMUNICATION

### A. *Industrial Ethernet Communication Network*

The system has five industrial Ethernet stations, which use the Ethernet switch's RJ45 interface to communicate with each other. Every station of the system has its own IP address according to local LAN segments. Hardware configuration and IP address of every station are shown as in table I.

TABLE I. HARDWARE CONFIGURATION OF ETHERNET NETWORKS

| **Ethernet station** | **Connecting hardware** | **IP address** |
|---|---|---|
| Office remote PC | Wireless router | 192.168.68.10 |
| Monitoring PC | Ethernet card | 192.168.68.11 |
| HMI | Ethernet interface | 192.168.68.12 |
| Siemens S7-300 | CP341 IT module | 192.168.68.13 |
| Siemens S7-200 | CP243-1 module | 192.168.68.14 |

The Ethernet switch adopts Siemens SCALANCE X208, which has eight RJ45 interfaces. According to the IP address in the data frame from the sending station, the switch determines which station should accept data. Data transmissions of each group are mutual shielding. In online communication, the switch firstly sorts data, then gives the mapping of orders, and finally data exchanges through the internal mechanism.

In factory management level, remote PC and monitoring PC use Wincc6.2 configuration software as platform, which set different access to operate the two PLC in workshop control level. The HMI adopts MT8070iH touch screen, and using touch screen to communicate with PLC in workshop level, the operators can get the dyeing equipment production information easily and achieve three levels management of equipments.

### B. *Profibus-DP Communication Network*

S7-300 PLC communicates with the four sets of S7-200 PLC, forming the structure of a master and four slaves. In the system S7-200 CPUs are all mounted EM277 intelligent communication module, communication rate adapting by itself. As slave station, EM277 accepts data from I/O configuration of master station S7-300. Master station can directly read or write data from S7-200 storage area.

The design process of profibus-DP network configuration is shown in Fig. 2.

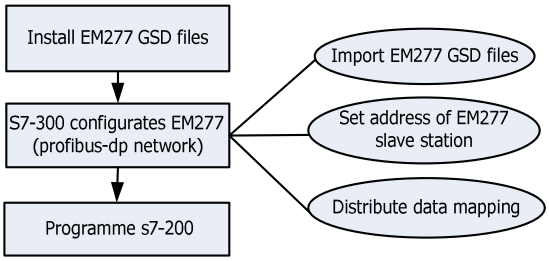
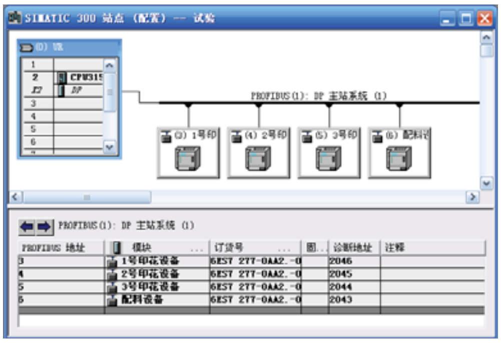


Figure 2. Flow chart of profibus-DP network configuration

After hardware are configured as shown in Fig. 3, the configuration information should be downloaded into S7-300 CPU module. When the four sets of S7-200 slave stations communicate with S7-300, both sides read or write data from the memory of master-slave data mapping (as shown in table II), and realize two-way communication.



PROFIBUS(1): DP 主站系统 (1)

| PROFIBUS 地址 | 模块 ... | 订货号 ... | 固... | 诊断地址 | 注释 |
|---|---|---|---|---|---|
| 3 | 1号印花设备 | 6ES7 277-0AA2.-0 | | 2046 | |
| 4 | 2号印花设备 | 6ES7 277-0AA2.-0 | | 2045 | |
| 5 | 3号印花设备 | 6ES7 277-0AA2.-0 | | 2044 | |
| 6 | 配料设备 | 6ES7 277-0AA2.-0 | | 2043 | |

Figure 3. Hardware configuration of profibus-DP network

TABLE II. DATA MAPPING OF MAIN-SLAVE STATIONS

| **Slave station name** | **Add** | **Slave station memory** | **Master station memory** |
|---|---|---|---|
| Printing machine 1 | 3 | VB500～VB515 | IB0～IB7 QB0～QB7 |
| Printing machine 2 | 4 | VB520～VB535 | IB8～IB15 QB8～QB15 |
| Printing machine 3 | 5 | VB540～VB555 | IB16～IB23 QB16～QB23 |
| Batching equipment | 6 | VB560～VB575 | IB24～IB41 QB24～QB41 |

### C. *Communication program of Modbus RTU protocol*

Modbus protocol is located in the second floor of the OSI model and a main-slave protocol, which has only one master station in serial bus, adopts the way of request-response communication. The main station sends a string data involving slave station's address. Only the main station can start the Modbus communication. Slave stations don't send data before receiving the main station's request, also slave stations won't communication between each others.

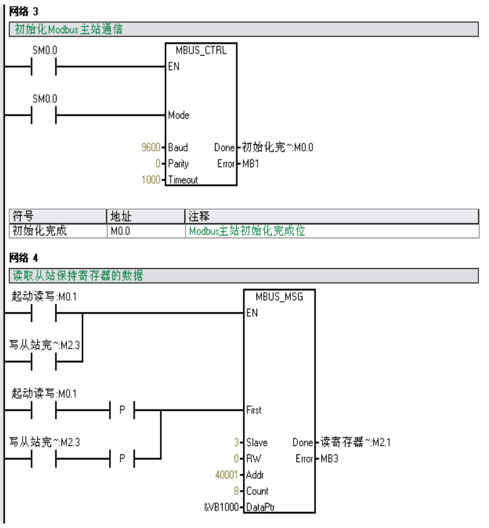
As the master station, S7-200 CPU226 communicates with Delta PLC of the wax printing machine through own Port0 on Modbus RTU protocol. Useing Modbus RTU master instruction in STEP 7-Micro/WIN V4.0 SP5, the digital quantity, analog I/O and holding registers from Delta PLC can be read and written. Before programming, order library should be added and the instruction for calling "Modbus Master Port0" should be chosen. Because the Port0 is taken up by the master station of Modbus, it is set as free port communication mode.

The basic structure of Modbus RTU communication frame is shown in Fig. 4. The slave station address is from 0 to 247 in the frame. The slave address and function code both stand one byte, the starting address and CRC both stand one word in command frame. Data uses word as the unit, high byte in the front, low byte at last.

| Station Address | Function Code | Data 1 | … | Data n | CRC Low Byte | CRC Hight Byte |
|---|---|---|---|---|---|---|

Figure 4. The basic structure of Modbus RTU communication frame

Part of the Modbus RTU communication program code is shown as in Fig. 5:



| 符号 | 地址 | 注释 |
|---|---|---|
| 初始化完成 | M0.0 | Modbus主站初始化完成位 |

Figure 5.  Part of communication program of main station of Modbus RTU

In the Modbus RTU communication, the most important design is the calculation of CRC yards. The CRC data in message are regarded as a continuous binary number, and the most significant bit is sent first. The message data move left by 16 bits, and then are divided by generated polynomial $x^{16}+x^{15}+x^{2}+1$ (the corresponding binary digits is 2 # 1000 0000 0000 0101), the remainder which has 16 bits is the CRC code. It will be attached to the back of message and be sent together. If the process of transmission has no mistakes, the receiving message should be able to be divided with no remainder by the generated polynomial. If happens CRC mistake in the process of transmission, PLC will give up the message data and transmit again. The CRC calculation flow is shown as in Fig. 6.

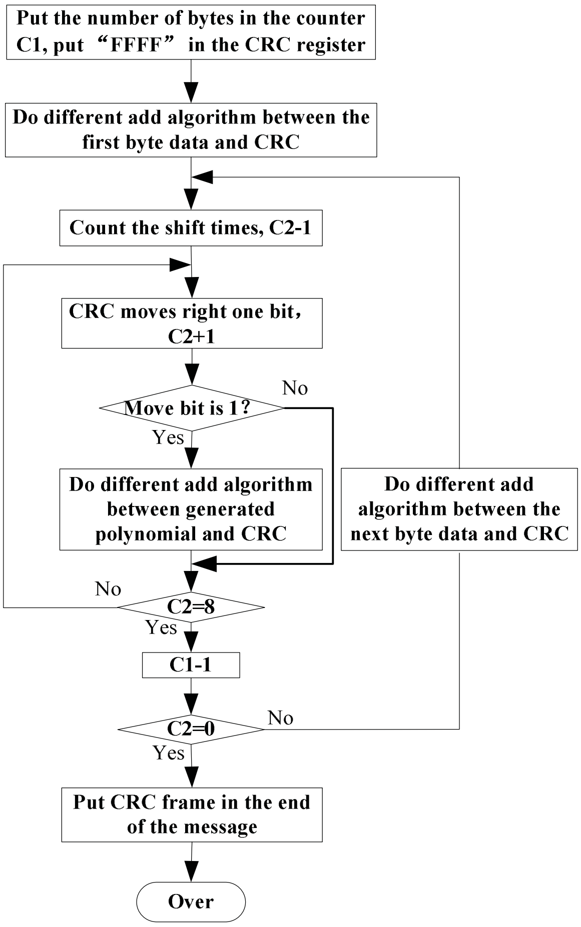

Put the number of bytes in the counter C1, put "FFFF" in the CRC register
Do different add algorithm between the first byte data and CRC
Count the shift times, C2-1
CRC moves right one bit, C2+1
Move bit is 1?
No
Yes
Do different add algorithm between generated polynomial and CRC
Do different add algorithm between the next byte data and CRC
No
C2=8
Yes
C1-1
C2=0
No
Yes
Put CRC frame in the end of the message
Over

Figure 6. Calculation flow chart of the CRC

### *D. Communication program of USS protocol*

The master station of S7-200 CPU226 uses the USS protocol to communicate with Siemens frequency converter of ventilation equipment through Port1

In the process of controlling frequency converter MicroMaster420, the main program calls communication procedure in S7-200 CPU226 to communicate with the MicroMaster420. Communication hardware consists of twisted pair and built-in RS485 communication interfaces of PLC and frequency inverter. The instructions of USS protocol are used ccording to the following steps: 1) Code user program. After installing Siemens order library, call control order in "USS Protocol Port 1"; 2) Distribute storage area for order library of USS protocol. In the frame of "storage area of library", set starting address for storage area; 3) Set communication parameter. Set communication parameter for the frequency inverter through the MicroMaster420 operation panel, to keep baud rate and address of slave station consistent with that in the user program. Part of the USS communication program code is shown as in Fig. 7.

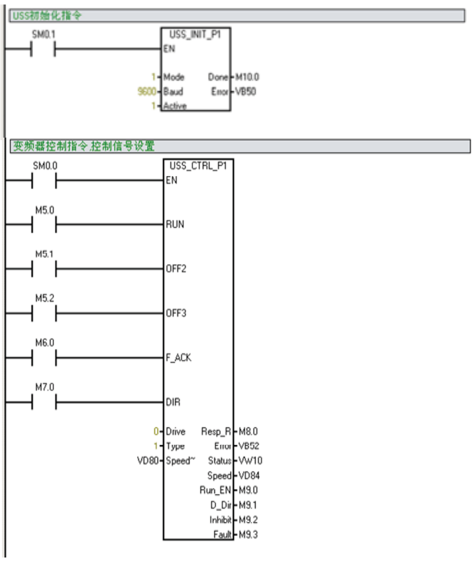


Figure 7. Part of USS protocol communication program

In the software program, most attraction should be put on the design of initial instruction USS-INIT and frequency inverter control instruction USS-CTRL. The initial instruction is used to allow, initialize or prohibit the communication of Micro-Master frequency inverter. The directive can set up communication protocol type and baud rate, activate the corresponding inverter. USS-INIT instruction must be first carried out successfully before other instructions, set “Done” after the instruction completion, and then can continue to carry out other instructions. Inverter control instruction USS-CTRL is used to control the Micro-Master frequency converter which is active. Each converter can only use one such instruction. This instruction puts user commands in a communication buffer. If one frequency converter appointed by "DRIVE" is just right selected by "ACTIVE" parameter in the USS-INIT instructions, the user commands in buffer will be sent to the inverter.

## V. CONCLUDING REMARKS

This article mainly aims at present situation of automation dyeing equipment in factory, designs the network system of three layers communication structure with industrial Ethernet and multi-field bus technology, realizing the integration of equipment management and control. The experiment proves that this system provides a good platform of manufacturing management and monitoring for enterprise management personnel, promotes the management level and production efficiency, and reduces the cost of factory management. The paper provides a feasible technology route for network design of production equipment in industrial control field.

## REFERENCES

[1] Li Zheng-jun, Design of field bus and industrial Ethernet application system [M], Beijing: People's Post and Telecommunication Press, 2006.

[2] Zheng Xiao-qian, Real-time monitoring system design based on industrial Ethernet communication [J]. Micro computer Information, 2010, 26 (3), pp.82-84.

[3] Liao Chang-chu, PLC program and application [M], Beijing: Mechanical Industry Press, 2008, 1.

[4] Siemens Automation Company, The manual of SIMATIC configuration hardware and communications connection, 2007, 2.

[5] Tan Ming, Wei Zhen etc. Analysis for Switched Ethernet Real-time Informations [J]. Microelectronics and Computer, 2011, pp. 51-54.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Design and Implementation of Real-Time Control System Using RTAI and Matlab/RTW

An Baoran, Guoping Liu
Center for Control Theory and Guidance Technology
Harbin Institute of Technology
Harbin, China
baoranan@gmail.com, gpliu@hit.edu.cn

Chai Senchun
School of Automation
Beijing Institute of Technology
Beijing, China
chaisc97@hotmail.com

*Abstract*—**In this paper, Linux-RTAI and Matlab/RTW is used to develop real-time control system for fast real-time simulation and implementation of the control algorithm. The hardware of the controller is based on PC/104 cards under Linux-RTAI operating system. The main characteristic of Matlab/RTW is that it can automatically generate the real-time simulation code for many target processors. The combination of Linux-RTAI as a controller platform and Matlab/RTW as control algorithm development environment brings the two main advantages: hard real-time capability and reduced time for control development. Moreover, online signal monitor and parameter modification is realized on the whole structure system. At last, the minimum execution time of real-time control task is determined and the application of real-time control system into flight vehicle is achieved in the work.**

***Keywords-RTAI; real-time; Matlab/RTW; PC/104; controller***

## I. INTRODUCTION

With unceasing consummation and development of control theory and computer technology in the field of automation, real-time capability is the important guarantee to realize the designed control algorithm. Real-time systems are those that can support the execution of applications with time constraints [1]. Basically, real-time systems are classified into two categories: hard real-time and soft real-time. In a hard real-time system, deadlines cannot be missed and the success depends on the task execution within deadlines [2][3][4]. Whereas in the soft real-time systems, there is a certain degree of tolerance for missing deadlines[5]. The importance of the deadlines depends on the type of control system.

Hard real-time control requires a software platform that guarantees certain time constraints. Real-Time Operating System(RTOS) should meet the basic requirements including predictability, pre-emptability, support for multi-threaded scheduling, concept of priorities, and resource sharing mechanisms avoiding priority inversion. Today, there are many available RTOS applied on the market, such as VxWorks[6],uCos[7],QNX[8] and Windows CE[9],etc., while the RTOS based on Linux are gaining attractions due to its robustness and open source code. And researchers, developers, and programmers prefer Linux as the open platform when developing applications. Based on the Linux kernel, RTAI was developed as a real-time operating environment solution at Dipartimento di Ingeneria Aerospaziale Politecnico di Milano (DIAPM) [10].It extends the Linux kernel with hard real-time functionality in addition to allowing the use of all standard Linux drivers and applications. Compared to the commercially available real-time OSs, The RTAIs performance is very competitive with the best commercial RTOSs. RTAI is open source and free under the terms of the GNU. So all needed software can be freely downloaded on the web during the building of the real-time control system.

Moreover, the diversity of control algorithms and development cycle of the projects are very important for engineers and researchers to develop and design every control project. Matlab, which stands for Matrix Laboratory, is an easy-to-use and powerful software for simulation aid design and technical computing. It not only provides Matlab language but also other application program interface to programming languages such as C, C++ and Fortran. Simulink is a graphical toolkit for modeling and simulation which can carry out many simulations of many aspects such as electronics, control engineering and signal processing. The Real-Time Workshop (RTW) provides a C code generator environment for rapid prototyping and development [11][12][13]. It generates source code from Simulink models to create real-time applications, which can operate in PC, ARM, DSP, PC/104, PLC and other hardware systems[14].

Therefore, combining Linux/RTAI and Matlab/RTW, one real control system is established in our project in order to realize hard real-time capability and a rapid development.

The remainder of this paper is organized as follows. The next section presents structural description of the whole real system. Section 3 presents the process of designing real-time controller. Section 4 presents software configuration and development on Windows PC which can implement automatic code generation and transplant into control application. Experimental results and conclusion are given in last two sections.

## II. SYSTEM ARCHITECTURE

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

The whole system structure includes mainly two parts, one is the real-time control system comprised by hardware and software which can provide the run-environment for real-time task execution, the other is the virtual simulation environment composed of computers and related software. The system structure is shown in Figure 1. PC is the core of the design of control block diagrams and automatic generation and compilation for C code, which have Matlab, Simulink, virtual machine and other related software environment. The objective of PC focuses on how to produce and download executable program applicable to the target computer (real-time control system). The target platform is the RTAI target run-time environment. It uses embedded industrial controlling computer PC/104 as real-time controller, which is responsible for receiving executable code from PC by TCP/IP protocol and data processing of various parameters under real-time environment Linux/RTAI. According to data processing results, it can make decision and produce control information to control the running state of controlled object.

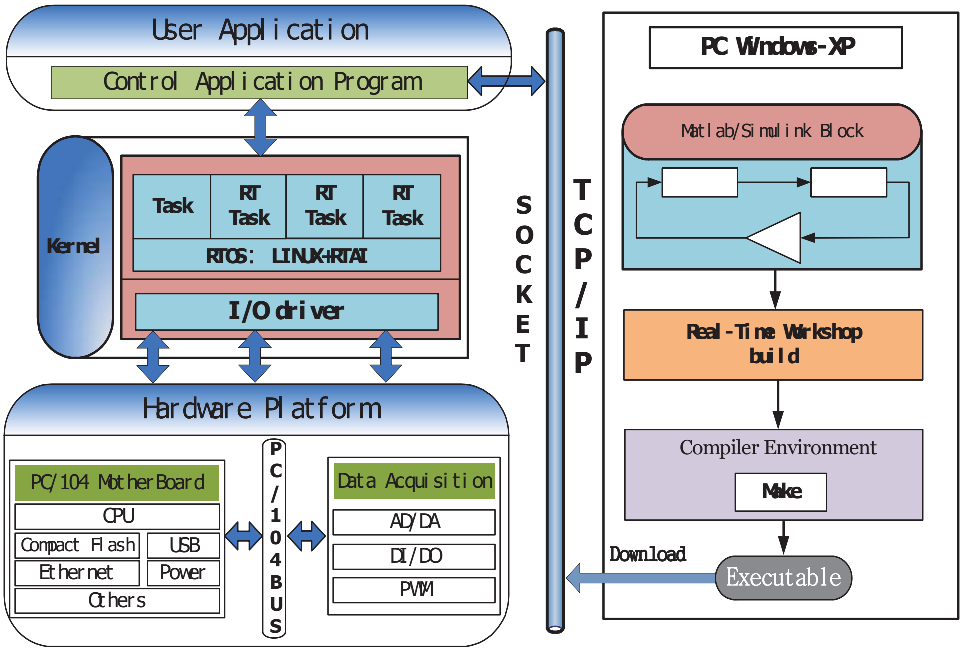


Figure 1. Overall structure

## III. Real-Time Control Target Platform

According to complexity and real-time demands of various kinds of controlled plants, the real controller is developed based on industrial modular design methodology in the following.

### A. *Hardware System*

In order to meet the requirement of a hard real-time environment, boards compliant to the PC/104 standard are chosen as the basis for real-time control system. PC/104 is an industrial control bus, which has some unique features such as small size (96 mm × 90 mm), stack-type connections, low bus drive current and so on[15]. The following two cards utilized in the Linux/RTAI system incorporate the PC/104 bus.

The heart of the hardware system is the CPU card. A PMI2 [16] card manufactured by SBS Science & Technology Co.,Ltd. was selected. The highly integrated CPU board is equipped with:

- Intel Pentium M running at 2 GHz;
- 256 MB of RAM;
- four USB 2.0 ports;
- two serial ports and one parallel port;
- industrial Ethernet interface;
- a CF card socket onboard;
- PC/104 bus.

A PC/104 ADT652 I/O module [17]is added to the stack to implement data acquisition and signal output. The hardware configuration of the card is equipped with the following components:

- 16 analog inputs, 12-bit A/D resolution;
- 100KHz A/D sampling rate;
- 4 analog outputs, 12-bit resolution;
- 24 digital I/O with programmable direction;
- 4 16-bit general purpose PWM ports.

### B. *Software Design*

An important item that allows the use of PC/104 cards in the application of various control algorithms is a real time operating system Linux/RTAI. In the design of the software system, it should include root filesystem, kernel image, TCP

/IP network support, necessary tools and I/O interface drivers. The design process of system software is given below.

*1) Installation GRUB bootloader.*

GRUB is the first software program that runs when a computer starts. It is responsible for loading and transferring control to the operating system kernel software. By using the Grub loader, it is very convenient for booting a GNU/Linux system directly from a CF disk.

*2) Linux 2.6 kernel configuration and filesystem transplantation.*

The new Linux v2.6 offers a host of improvement in the configuration and real-time features compared with the earlier Linux kernel. Therefore, the design chooses Linux 2.6.23 kernel version.

In order to make Linux real-time, the kernel source was patched with the correspondent RTAI patch. In the paper, RTAI version is 3.8 which includes hal-linux-2.6.23-i386-1.12-03.patch for making Linux 2.6.23 real-time. "HOWTO install RTAI"[18] describes instruction on how to configure the Linux kernel and RTAI target in detail.

Building the root filesystem with Busybox involves selecting files necessary for the system to run. Here is a reasonable minimum set of directories for the root filesystem(ext3): /bin, /boot, /dev, /etc, /home, /lib, /mnt, /proc, /sbin, /usr. These directories and files are copied into CF card in reference to “Building a root filesystem” [19].

*3) ADT652 I/O card driver modules including AD, DA, DI, DO, PWM.*

A device driver’s function is to provide access to a piece of hardware and the Linux Kernel provides a stable user space interface or applications. In Linux, device drivers are created as modules of the Linux Kernel that hide the details of how the device works. Each module can be used or removed from the kernel at runtime depending on what hardware is available.

In order to realize data collection control, device drivers for ADT652 card are responsible for analog input signals collection, digital input signals collection, analog output, digital output and PWM signal output. Five modules for ADT652 I/O signals control are designed in C programming, which can be added into kernel dynamically.

*4) Networked receive programs*

In networked control environment, network configuration and programming for receiving executable files is the important guarantee for target platform to realize communication with the host computer (PC side).

There are three important executable files in the /bin directory for network communication: *inetd*, *telnetd* and *recvn*. The *inetd* daemon is a super-sever daemon that manages Internet services on Linux system and the *telnetd* daemon is a server that supports the DARPA-standard TELNET virtual-terminal protocol. They are started automatically as the system starts up so that it is more convenient for users to login Linux/RTAI system and have remote control and on-line modifying from PC side to target computer.

Networked receive process named *recvn* is designed for receiving executable files generated from Simulink block diagrams by RTW in PC computer via TCP/IP using Socket. After successful receipt, the process creates a child process which executes the generated executable.

Figure 2 below illustrates the program execution flow after the target system starts.

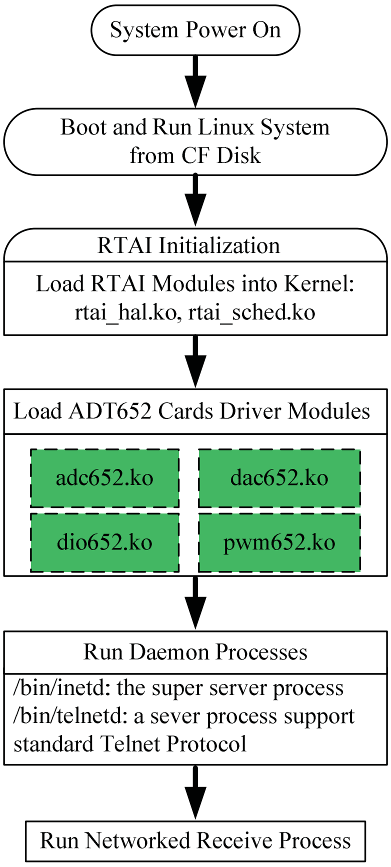

System Power On
Boot and Run Linux System from CF Disk
RTAI Initialization
Load RTAI Modules into Kernel: rtai_hal.ko, rtai_sched.ko
Load ADT652 Cards Driver Modules
adc652.ko
dac652.ko
dio652.ko
pwm652.ko
Run Daemon Processes
/bin/inetd: the super server process
/bin/telnetd: a sever process support standard Telnet Protocol
Run Networked Receive Process

Figure 2. The execution of system program

## IV. SOFTWARE CONFIGURATION AND DEVELOPMENT OF HOST SYSTEM

In this section, host PC’s software configuration and development is covered in details. Figure 3 describes the functions of software on host PC when a control project is built. Matlab and VMware Workstation are two necessary software components for the host PC (Windows XP Computer). As mentioned above, Matlab is responsible for constructing the control block diagrams in Simulink and generating C code through RTW. VMware Workstation, which is a kind of virtual machine software, can run various operating system images at the same time. A real-time Linux operating system is installed in VMware Workstation so that it can provide compile environment for converting generated C code into executable file. In host PC, another software named ‘NetConTop’ is developed to have online monitoring of parameters and signals of executable program running on PC/104 target. For details, see [20].Here is the main development process of Matlab/Simulink/RTW and VMware RT-Linux-OS.

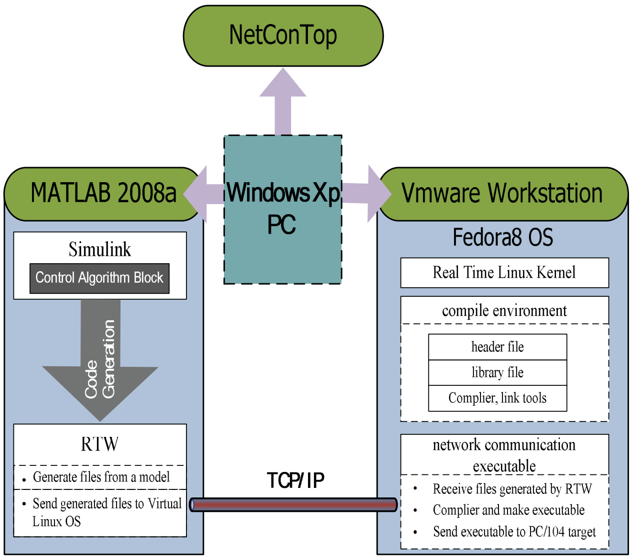


Figure 3. The structure of system software

### A. Matlab/Simulink/RTW

#### 1) I/O Blocks Development

I/O blocks are necessary elements from which real-time models can be built. Simulink enables designers to create custom blocks with C-MEX S-function. C-MEX S-function, which is written in C programming language and compiled as MEX-files, defines block properties and allows users to add desired blocks to Simulink models. In the work, I/O blocks are configured through custom blocks and listed below (Table 1).

TABLE I. I/O BLOCKS

| **Block Name** | **Parameter** | **Port Number** | |
|---|---|---|---|
| | | ***Input*** | ***Output*** |
| ADC | ● Input Channel<br>● Sample Time | 1 | 0 |
| DAC | ● Output Channel<br>● Sample Time | 0 | 1 |
| DI | ● Input Channel<br>● Sample Time | 8 | 0 |
| DO | ● Output Channel<br>● Sample Time | 0 | 8 |
| PWM | ● Output Channel<br>● Frequency<br>● Sample Time | 0 | 1 |

#### 2) Library Files for Code Generation

The process of generating target-specific code is mainly controlled by a target language compiler file and a template makefile. To support Linux/RTAI environment, the target language compiler file named '*grt_pc104.tlc*' is developed to control the way code is generated by RTW and the template makefile named '*grt_pc104.tmf*' is created to generate one right makefile for compile environment on Linux platform.

#### 3) Main Program of Model Files

To obtain real-time model executable, the generated code should be compiled and linked with specific library to one main program. At this point, a main file named '*rtmain.c*' is developed. It includes a main thread and two slave threads. When the executable starts, the main thread initiates the other two slave threads: one starts the real-time control task, the other is responsible for the communication between the real-time control task and the external GUI monitoring software (NetConTop).

#### 4) Files Sending Program

The generated code should be compiled on Linux platform, but Windows is much more popular with most PC users. So Virtual Machine is introduced to create virtual compile environment (Section B will give a detailed account of the process). In this case, a file-sending program named '*sendfilewin.exe*' is designed to send the generated code to Virtual Machine based on Windows Sockets. The command of calling this program is added into '*grt104.tmf*' so that the generated code can be sent automatically by a mouse click.

### B. VMware/RT-Linux-OS

#### 1) Build Compile Environment

VMware Workstation, which is a kind of virtual machine software, can run various operating system images at the same time and get connections to its host or other computers on the network after the right network configurations. Fedora 8, one of the most popular free-as-in-freedom Linux distributions, is installed on VMware Workstation 7.1 running on Windows XP PC host. In order to build complete compile environment, RT-Linux OS is created by using the same methodology as that used in constructing a real-time Linux kernel based on the above Fedora-VMware platform. As necessary files including header files, library files are added into the virtual RT-Linux environment by the way of file sharing, Fedora-VMware system has the ability of compiling and linking the generated code to the final executable file which has the same name as the model file.

#### 2) Network Communication Program

Figure 4 shows data communications among the three components: Matlab, Fedora-VMware and PC/104 target. In Fedora-VMware environment, one network communication program named '*recvfdr*' is designed by using Linux sockets. It has functions in the following:

*a) receive the generated files sent by sendfilewin.exe;*

*b) compile and link main file, generated files and library files together into a single executable file model;*

*c) download the executable model to PC/104 target.*

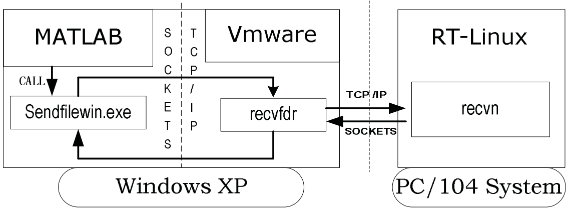


Figure 4. The structure of network communication

## V. APPLICATION EXAMPLE

The following examples illustrate the process during the implementation of a PID closed loop control for a four-wing flight simulator in order to have a hardware-in-the-loop test of one designed real-time control system.

### *A. Matlab Toolbox Testing*

The Matlab version in test is 2008a, and contains totally 804 blocks including Simulink, Control System Toolbox, Fuzzy Logic Toolbox and so on. After the test, there are only 96 blocks failed in the test, and the rest of which can be used in the system well.

### *B. Real-Time Performance Evaluation*

In order to get the minimum execution time of real-time control task, the blocks of DAC and PWM are tested to evaluate the system's real-time performance, and finally the minimum sample period is determined as 18us. As for the medium scale programs, the control task can be finished within one minimum period of 500us.

### *C. Fight-Attitude-Control Experiment*

Figure 5 shows the construction of four-fixed-wing flight simulator system. Two couples of DC Motors as well as gears are fixed on both ends of the beam to drive respective propellers. The motion of propellers can rotate the flight simulator in pitch or yaw. For example, a couple of horizontal propellers can generate a vertical force, causing the flight simulator to pitch up or down. Pitch angle and yaw angle are two important flight dynamics parameters that can be detected by two angle displacement sensors mounted on the pivot and sent to controller unit as the feedback information. Based on the feedback of pitch and yaw angles, the desired controller perform a control signal to motors through a power-expansion circuit in order to realize the attitude control of the flight simulator.

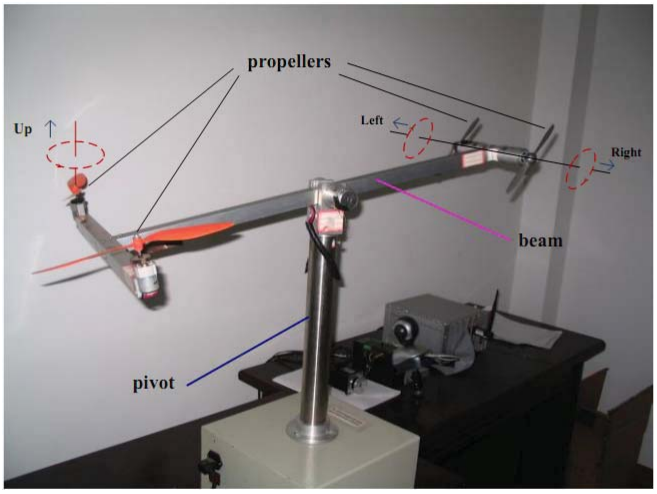


Figure 5. Four-fixed-wing flight simulator

The design and implementation fight-attitude-control system is shown in the following steps.

Step 1 Design a model file including ADC blocks, DAC blocks and other system blocks for both controller and plant in Simulink. In the paper, a PID controller with Loop-up compensator is designed for flight attitude control implementation as shown in Figure 6.

Step 2 Configure IP addresses of VMware and PC/104 target in Matlab/RTW dialog box and start automatic compile and download process by clicking build button.

Step 3 Use NetConTop software to have a monitor of PC/104 controller. In the project of NetConTop, the signals and parameters of the control model can be obtained and tuned with GUI application by users.

Finally, experimental results showed that the entire controller could complete all its tasks within the sampling period of 400 μs. Figure 6 shows the control block diagram and the monitoring interface of flight simulator system.

## VI. CONCLUSION

In order to satisfy the requirement of fast development and hard real-time capability, the paper takes Matlab/RTW and RTAI as software combination, and it realizes the modules imaging, modularization and configuration. The process of hardware and software design for real-time control system is described in details. At last, experimental results show that one control block diagram can complete all of its tasks within the sampling period of 500us and the application into flight vehicle is achieved.

## ACKNOWLEDGMENT

The authors would like to thank the referees for their valuable and helpful comments which have improved the presentation.

## REFERENCES

[1] Ramamritham Krithi, Stankovic John, “Scheduling Algorithms and Operating Systems Support for Real-time Systems,” Proceedings of the IEEE,1994,82(01) doi:10.1109/5.259426

[2] L. Sha, T. Abdelzaher, K. E. Arzen, A. Cervin, T. P. Baker, A. Burns, G. Buttazzo, M. Caccamo,J. Lehoczky, and A. K. Mok, “Real time scheduling theory: A historical perspective,” Real-Time Systems,vol.28, no.4, pp.101-155, Nov. 2004.

[3] Giorgio Buttazzo, “Hard Real-Time Computing Systems: Predictable Scheduling Algorithms And Applications,” Second Edition, Springer, 2005.

[4] E. Douglas Jensen. “Hard and soft real-time,” http://www.real-time.org/,2011.

[5] C. Lin and S. A. Brandt, “Improving Soft Real-Time Performance Through Better Slack Management,” Proceedings of the IEEE Real-Time Systems Symposium (RTSS 2005), pp.314, Miami, Florida, December 58, 2005.

[6] VxWorks, Available: http://www.windriver.com/vxworks

[7] uCos, Available: http://www.micrium.com/

[8] QNX, Available: http://www.gnx.com/

[9] Windows CE, Available: http://www.microsoft.com/windows/embedded/wince

[10] Dipartimento di Ingegneria Aerospaziable Politecnico di Milano.RTAI Homepage. http://www.rtai.org/,2010.

[11] Mathworks Inc. Real-Time Workshop. http://www.mathworks.com/products/rtw/,2010.

[12] Bucher, R., Balemi, S., “Rapid Controller Prototype with Matlab/Simulink and Linux,”Control Engineering Pracrice 14, 2003,pp.185–192.

[13] PC/104 Specification [M]. Version 2.5. PC/104 Embedded Consortium, 2003.

[14] Gherasim C.,Van den Keybus, J.,Driesen, J.,Belmans, R.,"DSP implementation of power measurements according to the IEEE trial-use standard 1459," IEEE Trans on Instrumentation and Measurement Vol. 52,no.4,pp. 1086-1092,2004.

[15] http://www.sbs.com.cn/en/productshow.asp?id=740

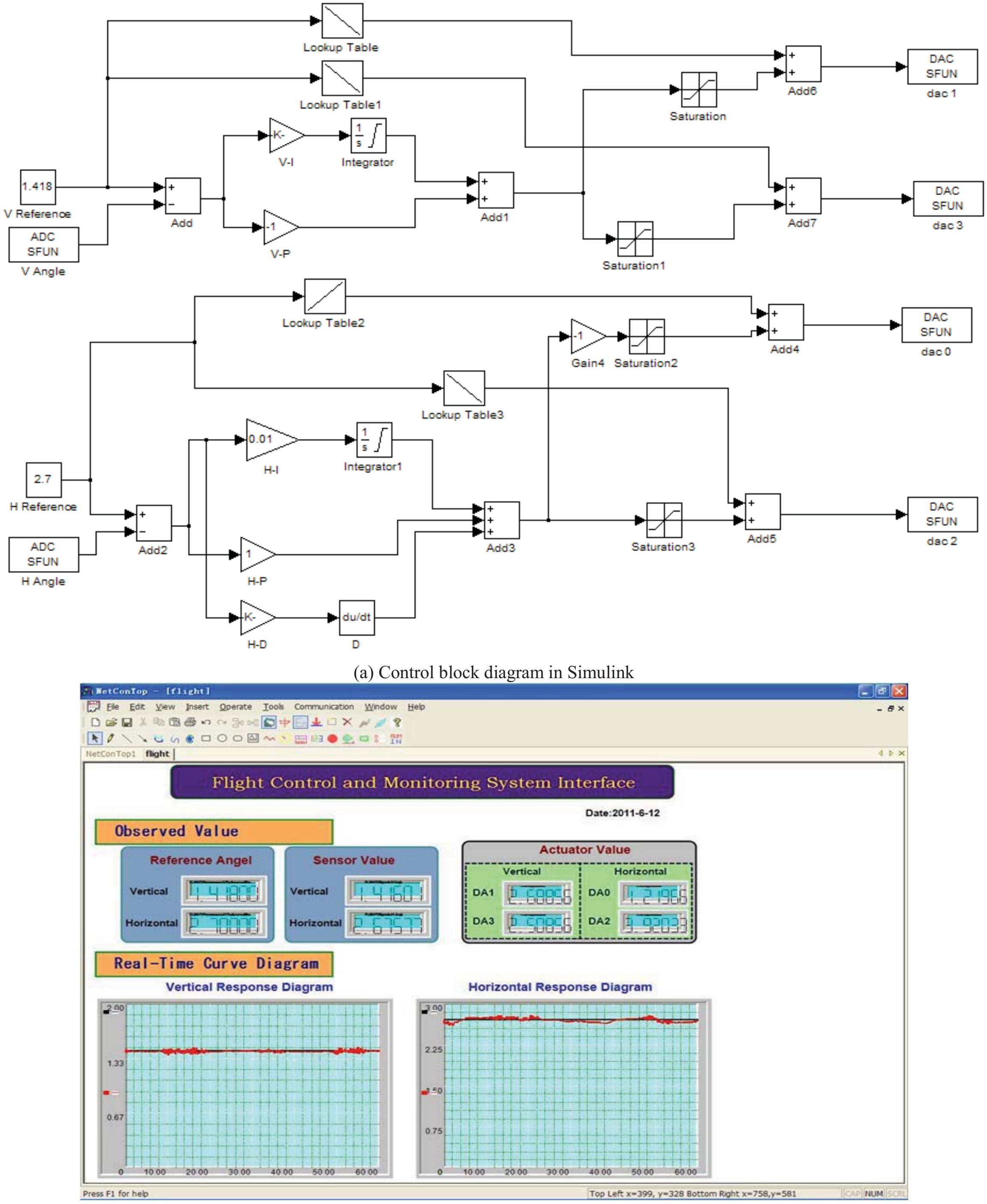
[16] http://www.sbs.com.cn/en/productshow.asp?id=692

[17] Dozio, L.,Mantegazza, P., "Real Time Distributed Control Systems Using RTAI," Proc. IEEE Symp. Object-Oriented Real-Time Distributed Computing, IEEE Press, May 2003, pp.11-18, doi: 10.1109/ISORC.2003.1199229

[18] http://www.isr.uc.pt/~rui/str/howto_install_rtai.html

[19] http://tldp.org/HOWTO/Bootdisk-HOWTO/buildroot.html

[20] PANG Zhonghua, LIU Guoping, Zheng Geng, Dong Zhe. Rapid Realization of Networked Control Systems Based on NetCon[J].Chinese Journal of Control and Instruments in Chemical Industry.2009,36(5):79～83.



(a) Control block diagram in Simulink

NetConTop - [flight]
File Edit View Insert Operate Tools Communication Window Help
NetConTop1 flight

Flight Control and Monitoring System Interface

Date:2011-6-12

Observed Value

Reference Angel
Vertical 1.41800
Horizontal 2.70000

Sensor Value
Vertical 1.41601
Horizontal 2.67577

Actuator Value
Vertical
DA1 2.60056
DA3 2.50856
Horizontal
DA0 1.21966
DA2 3.92033

Real-Time Curve Diagram

(b) Monitoring interface based on NetConTop

Figure 6. Experimental result of the fight-attitude-control system

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Video transmission on JXTA

Zhan-Yu Wang
School of Astronautics
Harbin Institute of Technology
Harbin, China
wangzhanyu@hit.edu.cn

Guo-Ping Liu[1,2]
1.School of Astronautics
Harbin Institute of Technology
Harbin, China
2.Faculty of Advanced Technology
University of Glamorgan
Pontypridd, U.K.
gpliu@hit.edu.cn

Zhen-fu Cui
School of Computer Science and Technology
Beijing Institute of Technology
Beijing, China

***Abstract*—By analyzing the peer classification, structure, and configuration of juxtapose framework (JXTA), a mixed topology in a campus network environment is designed and implemented for peer-to-peer instant communication. A novel method of transmitting video data on JXTA overlap network is proposed, by which video can be included in JXTA applications. The feasibility is proved, furthermore, the performance is tested.**

***Keywords- P2P; video transmision; JXTA***

## I. INTRODUCTION

In recent decades, P2P technology is increasingly used in instant message applications and file sharing systems based on internet. There has appeared many open source or private P2P frameworks provided for developers, by which users can implement their own communication softwares, such as JXTA[1], Self-Chord[2] and [3]. All above frameworks facilitate our developing of transmitting text messages, but they are not convenient to transmitting video. However, Red5[4], another powerful P2P framework, provides both text and video transmission interfaces, it is not widely used restricted by its size of hundreds Mbs.

Considering the raising performance and increasing bandwidth of internet devices, video transmission is not problem any longer for common local PC or even remote PC on wide area network scale. So, a cross-platform P2P instant messaging system with video transmission is quite significant for users.

JXTA, pronounced 'juxta', comes from the word juxtapose, that reflects the operations by which peers establish temporary associations to form a P2P network. Reference [5] shows a method which sends a stream media file via JXTA pipe, it denotes that we can also receive real time video package on JXTA overlap network.

In this paper, we describe how we design and implement the transmission with JXTA version 2.6 comprehensively. It is also applied for version 2.7.

It is quite noticable for JXTA beginners that the URI which is offered by official containing Relay list and Rendezvous list have not worked any more after version 2.5[1]. We should establish our own Rendezvous seed peer and Relay seed peer (if neccessary) on a known IP address, so that our JXTA peer can contact each other via the services provided by them.

## II. PEERS OF JXTA

There are three types of nodes in JXTA framework, edge peer, relay peer and rendezvous peer, they play different roles respectively when JXTA is running.

### *A. Edge Peer*

An Edge peer is nearly the most frequently used peer in JXTA. As a basic JXTA peer, it tries to connect and to remain the connection to one and only one Rendezvous peer. An Edge peer can use both unicasting and multicasting to contact with other JXTA peers. Each communicate mode can be implemented as TCP or HTTP by modifying the configuration file before Edge peer starts. If multicasting is enabled, an Edge peer can communicate with other Edge peers directly on LAN.

### *B. Relay Peer*

A Relay peer is a special Edge peer who has routing ability. We can compare a Relay peer to a router on network. It is needed when a peer with private IP attempts to contact peers from WAN. Typically, a remote peer wants to establish a connection to NATed peer or peer behind firewall, it should try to find a route which can go "through" NAT server or firewall device under the help of Relay peer. It provides means for peers having private IP on a LAN to become reachable. A Relay peer must have a public IP, namely, a private IP is useless for Relay peer.

### *C. Rendezvous Peer*

Rendezvous peers are Edge peers those offer services of JXTA rendezvous for a peer group. A peer group can assimilate to a JXTA LAN, and Rendezvous peer can assimilate to a gateway of this JXTA LAN. So, a peer group should have at least one Rendezvous peer. Edge peers can know each other via Rendezvous peer who performances as the leader of the group. Rendezvous peers accept lease requests from Edge peers and propagate these messages through the peer group to Edge peers.

## III. PEER CONFIRGURATION AND DEPLOYMENT

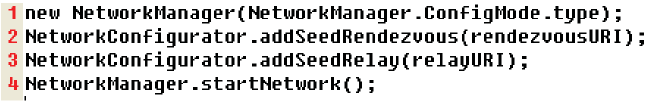
JXTA beginners are always confused for constructing their first JXTA environment, because of their lack understanding of the roles those JXTA peers play. Similarly, we can see different

This work was supported in part by the National Natural Science Foundation of China under Grant 61021002.

type of peers as different network devices with the help of network knowledge. Edge peers work as the common PC in LAN, rendezvous peer can be considered as the gateway PC of this LAN, and we can regard relay peer as the router that makes local peers contact other peers out of the LAN.

According version 2.5 of JXTA and former versions, there was URI http://rdv.jxtahosts.net/cgi-bin/rendezvous.cgi?3 plays the role of default Rendezvous seed. People can use it as a static Rendezvous peer in their own systems. But now, it is not valid any longer. Users need to configure different kinds of peers and deploy them on internet.

If you want to start a peer, at first you should set type of peer, such as Rendezvous or Relay peer, even Edge peer, these codes are given in Figure1.



```
1 new NetworkManager(NetworkManager.ConfigMode.type);
2 NetworkConfigurator.addSeedRendezvous(rendezvousURI);
3 NetworkConfigurator.addSeedRelay(relayURI);
4 NetworkManager.startNetwork();
```

Figure 1. Code of peer's configuration

After the codes on line 1, the seed should be added to the peer you implement, then your peer can communicate with other LAN peers via Rendezvous, even WAN peers via Relay.

These configure operations must be done before you start the JXTA. Once *startNetwork()* method runs, the JXTA framework works. It will cost few seconds to initiate the low layer program.

As mentioned before, different type peers play different roles respectively on JXTA network. For a JXTA application running on LAN, the organization is as simple as shows in Figure 2. All peers on LAN compose the overlap network of JXTA. At least one Rendezvous peer is required to this peer group, and other peers can communicate with each other, for example, using pipes.

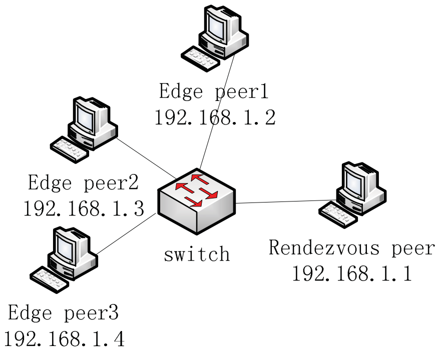


Figure 2. Topology of JXTA overlap network on LAN

But in actual case, only LAN scope is not enough for a more complex application, which peers are required to send and receive data through NAT devices or firewall devices even both of them. In this case, Relay peer works as router to help peers become visible to others, and Rendezvous peer plays a role like gateway to help Edge peers, which connect to it directly, to communicate with other any peers those connect indirectly.

Figure 3 demonstrates a typical topology of JXTA wide area network. A classic mixed structure is adopted to fulfill our design, which contains both peers and server that runs as the register server to let peers logon the JXTA network. Meanwhile, the server is a special configured peer that is not only the Rendezvous, but also a Relay peer.

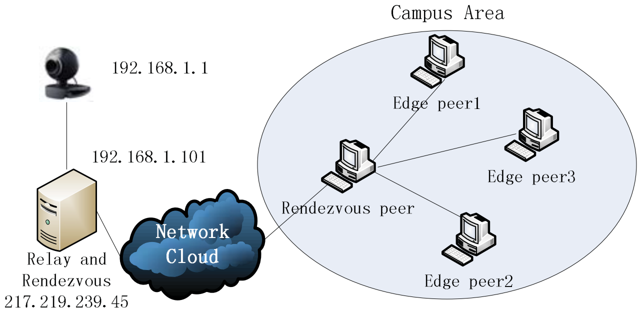


Figure 3. Topology of JXTA overlap network on WAN

## IV. Communication Among Peers

JXTA provides three kinds of data communication methods, singlepipe, bidirectionpipe and multicast socket, whose services are implemented based on TCP socket, and these functions are the most important of JXTA communication.

A singlepipe is a single direction socket pipe. Two participants of singlepipe should possess the same PipeAdvertisement, which is the unique identifier of JXTA pipe in a JXTA network. It means that the later joined peer must know the PipeAdvertisement created by the former peer, namely the PipeAdvertisement needs to be hardcoded in the later joined peer, and data can only be sent from the former to the later.It is quite inconvenient to develop a temporary pipe in the practical programs.

Unlike singlepipe, a bidirectionpipe can send and receive data from either side of the pipe. Contributed by bidirectionpipe, we do not need to create two simple pipes for both side of a data interaction.

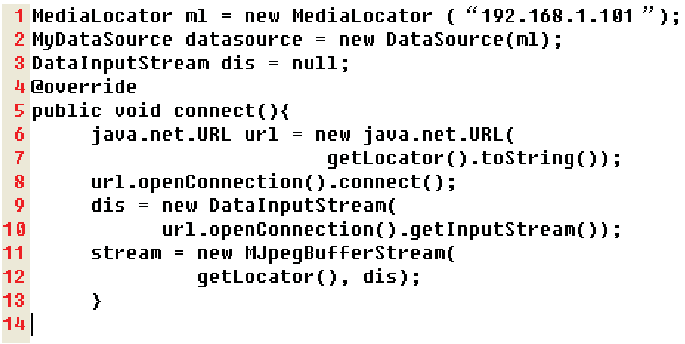
Multicastsocket plays important role in our software, because in many cases data needs to be sent from one peer to many other peers who have the advertisement of this multicastsocket. It is low-performing and inflexible to create many pipes for the data transmission, Instead, multicastsocket is quite appropriate for handling these jobs.

## V. Video Transmission

The working principle of IP camera that we use is to create 30 or less jpeg pictures and to send them each per second by a self-contained web server to browsers of clients via TCP/IP or HTTP links. And our purpose is to transmit these jpeg pictures on JXTA network by sender module below. So, jpeg pictures must be captured first.

### A. Video capture

According to the CGI document provided by IP camera producer, we establish a HTTP connection with IP camera and get the byte stream of jpeg pictures. The code below, in Figure 4 shows how we get data from IP camera.



```
MediaLocator ml = new MediaLocator ( “192.168.1.101 ”);
MyDataSource datasource = new DataSource(ml);
DataInputStream dis = null;
@override
public void connect(){
     java.net.URL url = new java.net.URL(
                         getLocator().toString());
     url.openConnection().connect();
     dis = new DataInputStream(
           url.openConnection().getInputStream());
     stream = new MJpegBufferStream(
              getLocator(), dis);
     }
```

Figure 4. Code of capturing video data from IP camera

As shown above, *ml*, the instance of class MediaLocator, will be set as the media locator in constructed function of class MyDataSource which extends PushBufferDataSource. In the overrided *connect()* method, *getLocator()* method will get the media locator and attach it to url, an instance of class URL. After connecting with url, variant *stream* is filled with class MJpegBufferStream object.

An instance of class MyDataSource is used in class MJpegBufferStream which implements PushBufferStream interface, and *read(Buffer buffer)* method is overrided in MJpegBufferStream class to push data into the parameter—buffer which is used be sent out by JXTA pipe.

### B. Sending and Receiving

After capturing the Jpeg pictures from IP Camera, each Jpeg picture, namely the data in buffer, will be encapsulated as a DatagramPacket object, and sent out via JXTA pipe or JXTAMultiSocket. This sending process is implemented by a single class HttpSender. Then, the peer creates an advertisement that announces its resource, publishes and remote publishes this advertisement on JXTA network, so that other peers can discover it.

Instance of class HttpReceiver takes charge of receiving data of DatagramPacket type one by one from JXTA pipe or JXTA multicastsocket.

### C. Display

One received DatagramPacket data is a MJpeg picture, it will be conducted by JPanelPlayer class which extends JPanel and responses to create a JPanel style player. This player is initiated in class CameraTest and once a picture is retrieved *updateScreen()* method in CameraTest would recall *repaint()* method to repaint the current picture on the JPanel instance. The relationship among all the classes is illustrated as UML diagram below, in Figure 5.

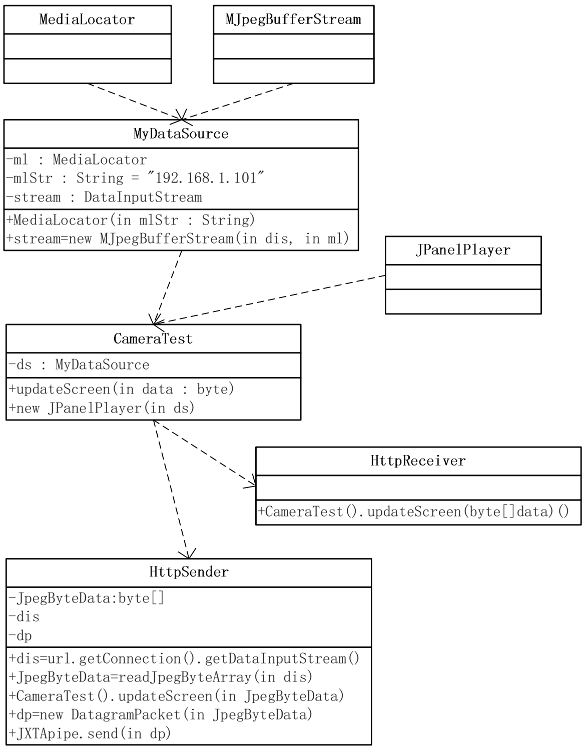

MediaLocator
MJpegBufferStream
MyDataSource
-ml : MediaLocator
-mlStr : String = ″192.168.1.101″
-stream : DataInputStream
+MediaLocator(in mlStr : String)
+stream=new MJpegBufferStream(in dis, in ml)
JPanelPlayer
CameraTest
-ds : MyDataSource
+updateScreen(in data : byte)
+new JPanelPlayer(in ds)
HttpReceiver
+CameraTest().updateScreen(byte[]data)()
HttpSender
-JpegByteData:byte[]
-dis
-dp
+dis=url.getConnection().getDataInputStream()
+JpegByteData=readJpegByteArray(in dis)
+CameraTest().updateScreen(in JpegByteData)
+dp=new DatagramPacket(in JpegByteData)
+JXTApipe.send(in dp)

Figure 5. UML diagram of classes

## VI. Example

### A. Enviroment

Table I shows both system and hardware devices information those participate in our test.

TABLE I. Environment of test

| ***Device*** | ***Info*** | ***Peer mode*** | ***IP Address*** |
|---|---|---|---|
| Server | CPU: i7 930 2.8Ghz<br>Memmory:3.25G<br>System:Win XP 32bit<br>BandWidth of Enthernet:100M | RDV and Rendez-vous | 217.219.118.45 |
| IP camera | | none | 192.168.1.101 |
| PC1 | CPU: AMD 640x4 3.0GHz<br>Memmory:3.25G<br>System:Win XP 32bit<br>BandWidth of Enthernet:100M | Edge | 219.217.242.18 |
| PC2 | CPU: AMD 640x4 3.0GHz<br>Memmory:3.25G<br>System:Win7 32bit<br>BandWidth of Enthernet:100M | Edge | 172.17.52.83 |
| PC3 | CPU: Pentium D 3.0GHz<br>Memmory:1G<br>System:Win XP 32bit<br>BandWidth of Enthernet:100M | Edge | 202.118.233.57 |

Environment of test

A server accesses in internet with a public IP 217.219.239.45, and connects to an IP camera with private IP 192.168.1.101 directly. PC1, PC2 and PC3 locate on different network environments that we do not know clearly, and they are all the receivers which are used to test the performance of this video transmission based on JXTA framework. The topology of this example is illustrated by Figure 6.

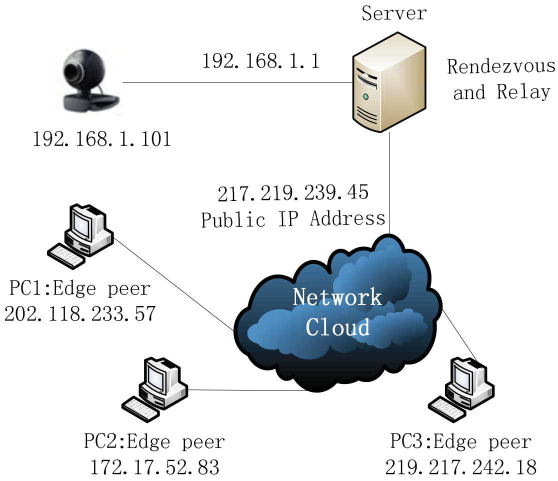


Figure 6. Test topology

## B. *Result*

This software is tested in three different places. It firstly runs on an IP 219.217.242.18 that connects to server directly, After connection established, video is captured by the recerver, and displayed in a JPanel style window that shows in Figure 7.

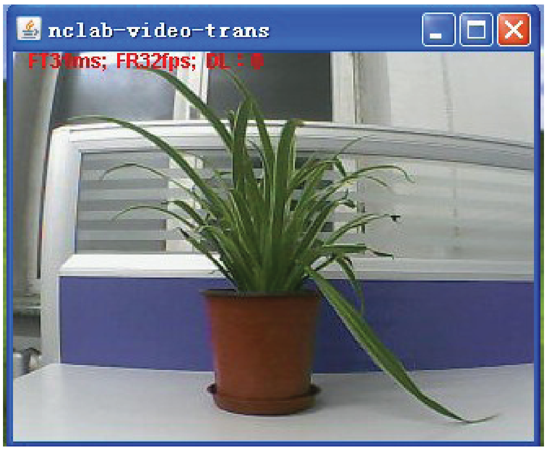
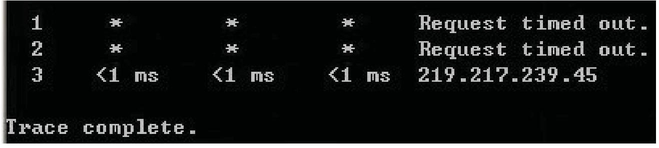


Figure 7. Captured video on IP 219.217.242.18

It is a direct connection between client and the server as shown in Figure 8.



```
 1     *        *        *     Request timed out.
 2     *        *        *     Request timed out.
 3    <1 ms    <1 ms    <1 ms  219.217.239.45

Trace complete.
```

Figure 8. Path between 219.217.242.18 and 219.217.239.45

Secondly, we run our application in another laboratory with IP 172.17.52.83, seemingly not in the same network segment to server's IP. The software received video data from server, and showed it by the player, as Figure9.

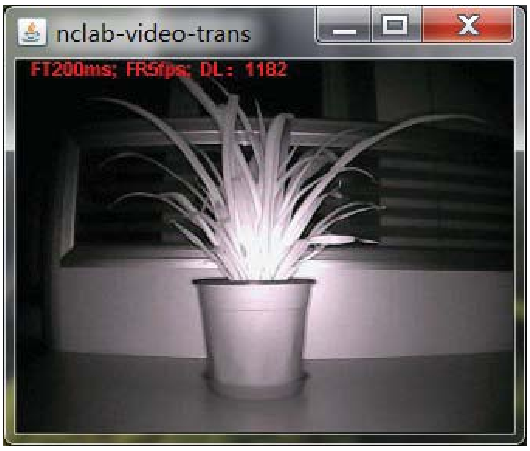
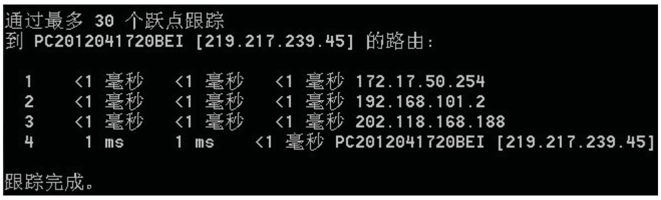


Figure 9. Captured video on IP 172.17.52.83

Figure 10 shows that there are four hops between this receiver and the server.



```
通过最多 30 个跃点跟踪
到 PC2012041720BEI [219.217.239.45] 的路由:

  1    <1 毫秒   <1 毫秒   <1 毫秒 172.17.50.254
  2    <1 毫秒   <1 毫秒   <1 毫秒 192.168.101.2
  3    <1 毫秒   <1 毫秒   <1 毫秒 202.118.168.188
  4     1 ms      1 ms    <1 毫秒 PC2012041720BEI [219.217.239.45]

跟踪完成。
```

Figure 10. Path between 172.17.52.83 and 217.219.239.45

Finally, we test it in another campus of our university, and the IP is 202.118. 233.57, seemingly not in the same network segment to the IP of server either, and the data is received, as we can see in a display window of Figure11.

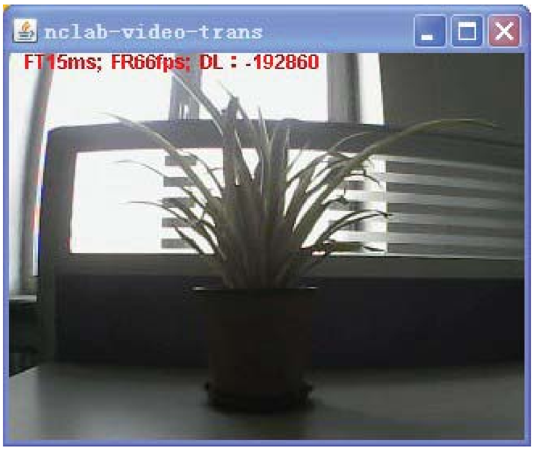
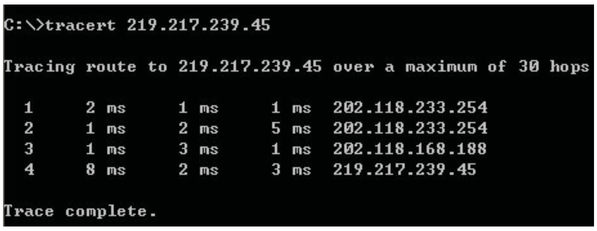


Figure 11. Captured video on IP 202.118.233.57

We find the three hops path from the local PC to server. That denotes the data can be delivered crossing different network segment, as shown in Figure 12.



```
C:\>tracert 219.217.239.45

Tracing route to 219.217.239.45 over a maximum of 30 hops

  1     2 ms     1 ms     1 ms  202.118.233.254
  2     1 ms     2 ms     5 ms  202.118.233.254
  3     1 ms     3 ms     1 ms  202.118.168.188
  4     8 ms     2 ms     3 ms  219.217.239.45

Trace complete.
```

Figure 12. Path between 202.118.233.57 to 217.219.239.45

That denotes the data can be delivered crossing different network segment.

## VII. Conclusion

The proposed video transmission system achieves sending and receiving video data via internet based on JXTA. It may be quite useful for developers who want to develop their own P2P instant message systems containing real time video image. By analyzing source code of JXT and CGI document of IP camera, we capture the video data from HTTP and divide to individual byte array for each MJpeg format picture. Then JXTA pipe or JXTA multicast socket is created to send them out. Finally, these video data can be received by other JXTA terminals on internet those run as the receiver. For original edition JXTA, for example version 2.7 and earlier releases, only text messages and XML format messages can be used by provided interfaces.

In this paper, we get a simple improvement of transmitting video via JXTA without any optimizing. So, the performance depends on network environment absolutely, this can be inferred indirectly in the result section of this article. As a large scaled application, P2P software should have an optimized structure to enhance its processing ability and increase it efficiency[6][7]. It is the target of our future research.

## References

[1] Jerome Verstrynge, Practical JXTA II, 2nd ed, Lulu Enterprises, Inc. (www.lulu.com), July, 2010, p 166.

[2] Agostino Forestiero, Emilio Leonardi, Carlo Mastroianni, “Self-Chord: A Bio-Inspired P2P Framework for Self-Organizing Distributed Systems”, IEEE/ACM Transactions Networking, vol.18, p1651-1664, October 2010.

[3] Silvia Rueda, Pedro Morillo, Juan M. Orduna, “A Peer-To-Peer Platform for Simulating Distributed Virtual Environments,” Proceedings of the 13th International Conference on Parallel and Distributed Systems – ICPADS, vol 2, 2007.

[4] Wang Dongjin, Xu Ke, “Red5 Flash Server Analysis and Video Call Service Implementation,” Proceedings - 2010 IEEE 2nd Symposium on Web Society, Beijing, China, pp 397-400, August 2010.

[5] Sébastien Vénot, Lu Yan, “On-demand mobile peer-to-peer streaming over the JXTA overlay,” Proceedings - International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, UBICOMM. Turku, Finland, vol. 16, pp. 131–136, 2007.

[6] Chang Le, Liu Yangyang, Wei Zhonghua, Pan Jianping , “Optimizing BitTorrent-like peer-to-peer systems in the presence of network address translation devices,” Peer-to-Peer Networking and Applications, vol. 4, num. 3, pp. 247–288, 11. 2011.

[7] Gaspare Giuliano E. Bruno, Daniel R. Figueiredo, “Optimal Multi-Tree Peer-to-Peer Video Streaming,” Proceedings of 19th International Conference on Computer Communications and Networks, ICCCN 2010. Zurich, Switzerland

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# A real-time optimization control strategy for power management in fuel cell/battery hybrid power sources

Chun-Yan Li
Institute of Automation
Chinese Academy of Sciences
Beijing, China
Email: lichunyan820@126.com

Geng Zheng
Institute of Automation
Chinese Academy of Sciences
Beijing, China
Email: geng.zheng@ia.ac.cn

Guo-Ping Liu
Faculty of Advanced Technology
University of Glamorgan
Pontypridd,Wales,UK
Email: gpliu@glam.ac.uk

***Abstract*—The power management strategy can greatly affect the fuel economy of hybrid power systems. This work presents an optimization based power management strategy of hybrid fuel cell power sources during real-time operation. In this approach, local optimization strategy is adopted because it doesn't need the priori knowledge of the future power demand. Every step, the current power demand is measured and the real-time optimal power distribution is determined by maximizing the efficiency of the hybrid system. Simulation and experimental results are presented to show that this real-time power management optimization strategy is feasible and can provide good fuel economy.**

***Keywords*—Fuel cell hybrid, Power management, Real-time optimization**

## I. Introduction

Fuel cell is considered as one of the most attractive power sources with applications ranging from automobiles to stand alone power generation plants due to its environmental-friendly [1], [2]. Among the various kinds of fuel cells, proton exchange membrane fuel cells is one of the promising energy sources due its higher power density, lower operation temperature, quick start up and long cycle life [3]–[5]. However, the dynamics of the fuel cell stack is limited by the air and hydrogen delivery system. A fuel cell only power system may not be sufficient to meet the load demands, especially in peak power demand or transient situation. Other energy storage devices, such as batteries and supercapacitors, are needed to supplement the fuel cell system in application [6]–[8].

In research of the fuel cell hybrid power sources, study of the power management strategy is one of the important tasks, especially in fuel cell hybrid vehicles. Many literatures about power management control strategy, based on optimization, can be found. Brahma et al. [9] used the dynamic programming technique in the optimization of instantaneous generation/storage power split in series hybrid electric vehicles. Delprat et al. [10] presented a global optimization method based on optimal control theory. All of these optimized power management strategies are based on a prior knowledge of the future power demand and not suitable for real time control.

Also, Some literatures studied the real time power management strategies, based on real time optimization. Delprat et al. [11] derived a real-time control strategy from optimal control theory. Rodatz et al. [13] also used ECMS to determine the real time optimal power distribution of a fuel cell/supercapacitor hybrid vehicle.

In general, it is need a priori knowledge of the future power demand to find a global optimal solution of power management. So, global optimization is infeasible in real time power management control. On the other hand, strategies that deal with local optimization are suitable to real implementation. In this paper, a fuel cell/battery hybrid power source is studied and a local optimal power management strategy is presented. The performance of this local optimal solution is compared with that of a optimal fuzzy power control and management strategy that presented in literature [14].

The organization of this paper is as follows. In section II, the structure, characteristics and models of hybrid power sources are introduced. Then, the proposed power management strategy based on local optimization is presented in section III. In section IV, the simulation results are shown and the performance of the proposed strategy is compared with that of other strategies. The experimental results are also reported in this section. Finally, the conclusions of this paper are included in section V.

## II. Hybrid power sources

The fuel cell/battery hybrid power sources combine the high power density of batteries with high energy density of fuel cells. The fuel cell hybrid power sources consist of the fuel cell stack, a battery bank, the DC/DC converter to stable the fuel cell output voltage.

Fig. 1 is represented the proposed topology of the fuel cell hybrid power system. The fuel cell system is connected to the DC bus with a DC/DC converter, whereas the battery bank is directly connected to DC bus passively. As to the load, a AC motor is considered. The current flow to the DC bus from fuel cell system can be controlled by the DC/DC converter, the difference between the current draw from the inverter and the current out from DC/DC converter is compensated by the battery bank. Given a certain load power $P_{load}$, this power

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

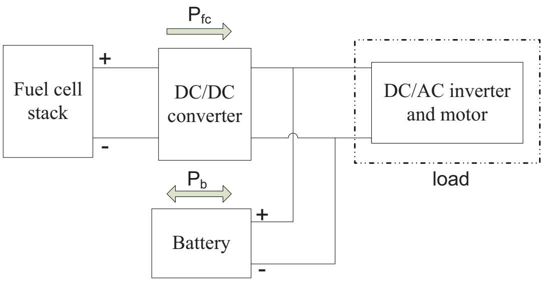


Fig. 1. configuration of fuel cell hybrid power sources

should be supplied by fuel cell system, $P_{fc}$ and the battery bank, $P_b$. At every time, the power balance should be satisfied. That is

$$P_{fc}(t_k)\eta_{dc} + P_b(t_k)\eta_b = P_{load}(t_k) \quad \forall t_k \tag{1}$$

where $\eta_{dc}$ and $\eta_b$ are the efficiency of DC/DC converter and the efficiency of battery bank respectively. Here, we assume that the DC/DC converter is well controlled and the efficiency is known.

The main objective of the power management strategy is to reduce the hydrogen consumption and improve the efficiency of hybrid power system. For a given fuel cell system, the hydrogen consumption along with the output power of fuel cell system is a matter of much concern. So, here, we developed a static fuel cell model. It is assumed that the temperature of the fuel cell system is well maintained at the operating condition (around 65°C) and the pressure difference between the cathode and the anode is ignored. A typical efficiency characteristic of a fuel cell system with a 50-kW rate power is shown in Fig. 2.

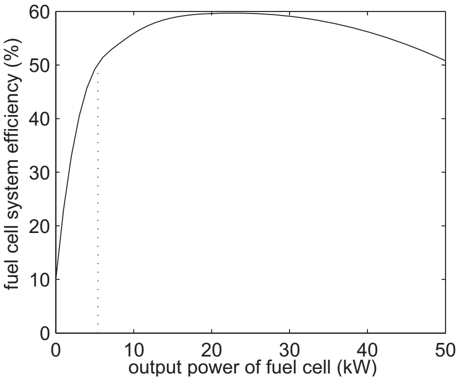


Fig. 2. ADVISOR efficiency map for a 50-kW fuel cell system as a function of output power

The battery pack consists of serially connected battery cells. The internal resistance is the major factor to limit charging and discharging capability. The internal resistance model is used in this study. This model is related to work which was originally performed by Idaho National Engineering Laboratory to model flooded lead-acid batteries [15]. A battery cell is modelled with a voltage source and an internal resistor with temperature ignored (Fig. 3). The resistance and open circuit voltage both are the nonlinear functions of battery state of charge (SOC) (Fig. 4).These relationships are implemented as look-up tables with test data. This simple battery model enables fast calculation for optimization and makes it possible to apply the real time optimization power management strategy.

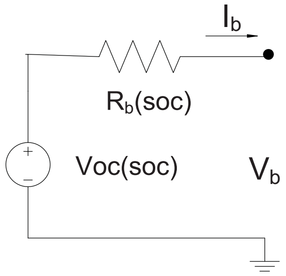


Fig. 3. Internal resistance battery model.

As shown in Fig. 3, the terminal voltage of battery pack $V_b$ can be written by

$$V_b = n_b(V_{oc} - R_b I_b) \tag{2}$$

where $n_b$ is the number of battery cells, $V_{oc}$ is the open circuit voltage of the battery cell, $R_b$ is the internal resistance and $I_b$ is the current flow out the battery. $I_b$ can be calculated by

$$I_b = \frac{V_{oc} - \sqrt{(V_{oc}^2 - \frac{4R_bP_b)}{n_b}}}{2R_b} \tag{3}$$

The SOC of battery is denoted by

$$SOC(k) = SOC_0 - \frac{1}{C_b}\int_{t_0}^{t_k} I_b dt \tag{4}$$

where $k$ is the time step and $C_b$ is the capacity of battery cell.

When the battery pack is discharging, the discharge efficiency of the battery pack can be written as

$$\eta_{dis} = \frac{V_bI_b}{V_{oc}I_b} = \frac{1}{2} + \frac{1}{2}\sqrt{1 - \frac{4R_{dis}P_b}{V_{oc}^2}} \tag{5}$$

where $R_{dis}$ is the discharge resistance of battery cell.

Similarly, for the battery charge process, the charge efficiency is given by

$$\eta_{chg} = \left(\frac{1}{2} + \frac{1}{2}\sqrt{1 - \frac{4R_{chg}P_b}{V_{oc}^2}}\right)^{-1} \tag{6}$$

where $R_{chg}$ is the charge resistance of the battery cell.

## III. Power management strategy of the hybrid power sources

In this section, a power management strategy based on real time optimization is addressed. The main objective of power management strategy is to improve the efficiency of the hybrid system while maintaining the SOC of the battery pack in a certain range.

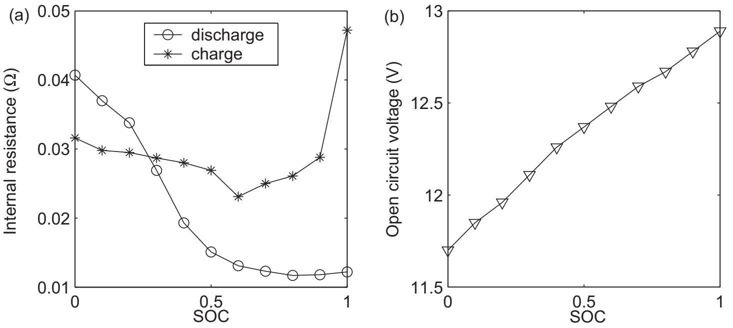


Fig. 4. The relationship between (a)internal resistance and SOC, and (b) open circuit voltage and SOC in ADVISOR.

### *A. The concept of equivalent fuel consumption*

Ideally, The overall efficiency of the hybrid system is defined by

$$\eta_{sys} = \frac{\sum_0^{t_f} P_{load}(t_k)\Delta t_k}{E_{fc} + \sum_0^{t_f} \lambda_{pb}(t_k) P_b(t_k)\Delta t_k} \tag{7}$$

where $\eta_{sys}$ is the overall efficiency of the hybrid system , $P_{load}(t_k)$ is the power supplied in to the vehicle at time step $\Delta t$, $E_{fc}$ is the energy of hydrogen fuel supplied into the fuel cell stack, $P_b(t_k)$ is the battery power of charge or discharge at time step $\Delta t$, $\lambda_{pb}(t_k)$ is the equivalence factor which can be evaluated by the concept of equivalent consumption at time step $\Delta t$.

The energy of hydrogen fuel supplied to the fuel cell during the given mission is calculated according to

$$E_{fc} = \sum_0^{t_f} \frac{P_{fc}(t_k)}{\eta_{fc}(t_k)} \Delta t \tag{8}$$

where $\eta_{fc}(t_k)$ is the efficiency of the fuel cell system at time step $\Delta t_k$ when the output power is $P_{fc}(t_k)$. It can be obtained from the efficiency map of the fuel cell system shown in Fig. 2.

To make the electrical energy consumption of the battery and fuel energy of hydrogen comparable, the electrical energy consumption of the battery is converted into equivalent fuel consumption. Paganelli et al. [16] proposed the concept of equivalent fuel consumption. The concept is that if the battery discharged some power $P_b(t_k)$ at time step $\Delta t$, to maintain the SOC, the battery will be recharged using the energy of the fuel cell in the future. The discharge efficiency can be written as

$$\eta_{dis}(t_k) = \frac{1}{2} + \frac{1}{2}\sqrt{1 - \frac{4R_{dis}(t_k)P_b(t_k)}{V_{oc}(t_k)^2}} \tag{9}$$

where $R_{dis}$ is the discharge resistance of battery.

Similarly, for the battery charge process, the charge efficiency is given by

$$\eta_{chg}(t_k) = \left(\frac{1}{2} + \frac{1}{2}\sqrt{1 - \frac{4R_{chg}(t_k)P_b(t_k)}{V_{oc}(t_k)^2}}\right)^{-1} \tag{10}$$

The battery equivalent fuel consumption is defined as

$$\begin{aligned} &C_b(t_k) = \lambda_{cb}(t_k)\ P_b(t_k) \\ &\lambda_{cb}(t_k) = \begin{cases} \dfrac{C_{fc,avg}}{P_{fc,avg}\eta_{dis}(t_k)\eta_{chg,avg}} & P_b(t_k) \geq 0 \\ \dfrac{C_{fc,avg}\eta_{chg}(t_k)\eta_{dis,avg}}{P_{fc,avg}} & P_b(t_k) < 0 \end{cases} \end{aligned} \tag{11}$$

Because the future operating points are not known, the average charge efficiency of the battery is used and also, the average fuel cell power and its fuel consumption are used.

According to (11), the equivalence factor $\lambda_{pb}(t_k)$ can be calculate by

$$\lambda_{pb}(t_k) = \frac{P_{fc,avg}}{\eta_{fc}C_{fc,avg}}\lambda_{cb}(t_k) \tag{12}$$

### *B. Optimization problem statement*

The problem is to solve this global optimization problem, the load power demand in the given mission has to be known a priori. But in many cases, we can not get this power demand until it is generated by the load, especially in automotive applications. So instead of the global optimization, we reduce the global optimization to a local one. That is, for each time $t_k$ with a time step $\Delta t_k$, we solve the local optimization problem by maximizing the objective $J(t_k)$, defined as

$$J(t_k) = \frac{P_{load}(t_k)\Delta t_k}{(\frac{P_{fc}(t_k)}{\eta_{fc}(t_k)} + \lambda_{pb}(t_k)P_b(t_k))\Delta t_k} \tag{13}$$

The global optimization is not equal to the local problem described above. But it can be easily used for real time control whereas its global counterpart is non-causal and non-realizable [12].

For all $t_k$ the constraints in the fuel cell power and the battery pack power are

$$\begin{array}{ccccc} 0 & \leq & P_{fc}(t_k) & \leq & P_{fc,max} \\ \Delta P_{fc,fallrate} & \leq & \dfrac{\Delta P_{fc}(t_k)}{\Delta t_k} & \leq & \Delta P_{fc,riserate} \\ P_{b,chg,max} & \leq & P_b(t_k) & \leq & P_{b,dischg,max} \\ SOC_{min} & \leq & SOC(t_k) & \leq & SOC_{max} \end{array} \tag{14}$$

where $P_{fc,max}$ is the maximum power that fuel cell system can deliver, $\Delta P_{fc,fallrate}$ and $\Delta P_{fc,riserate}$ are maximum fall rate of $P_{fc}$ and maximum rise rate of $P_{fc}$ respectively. With regard to the battery pack, the maximum power flows are also limited. The maximum power that the battery pack can deliver $P_{b,dischg,max}$ or store $P_{b,chg,max}$ depends on the actual voltage of the battery pack $V_{oc}$, the maximum voltage $V_{b,max}$, and the minimum voltage $V_{b,min}$ [17].

$$P_{b,chg,max} = \frac{n_b V_{oc}(V_{oc} - V_{b,max})}{R_d} \tag{15}$$

$$P_{b,dischg,max} = \frac{n_b V_{oc}(V_{oc} - V_{b,min})}{R_d} \tag{16}$$

Because $V_{oc}$ and $R_d$ both are depend on the SOC of the battery pack, whereas the SOC is various in the duration of the given mission. The values of $P_{b,chg,max}$ and $P_{b,dischg,max}$ are various at each time step in optimization.

To get the optimal power distribution, the local optimization problem that should be solved at each time $t_k$ with a time step $\Delta t_k$ is

$$\begin{array}{lccccc}
 & \underset{P_{fc}}{Maxmize}\, J(t_k) = \frac{P_{load}(t_k)\Delta t_k}{\left(\frac{P_{fc}(t_k)}{\eta_{fc}(t_k)} + \lambda_{pb}(t_k)P_b(t_k)\right)\Delta t_k} \\
s.t. & P_{fc}(t_k)\eta_{dc} + P_b(t_k)\eta_b = P_{load}(t_k) \\
 & 0 & \leq & P_{fc}(t_k) & \leq & P_{fc,max} \\
 & \Delta P_{fc,fallrate} & \leq & \frac{\Delta P_{fc}(t_k)}{\Delta t_k} & \leq & \Delta P_{fc,riserate} \\
 & P_{b,chg,max}(t_k) & \leq & P_b(t_k) & \leq & P_{b,dischg,max}(t_k) \\
 & SOC_{min} & \leq & SOC(t_k) & \leq & SOC_{max}
\end{array} \tag{17}$$

To avoid the "starvation" of reactants in the fuel cell system and take the slow dynamics into account, the output power of fuel cell system is increased no faster than a certain power rise rate $\Delta P_{fc,riserate}$. Also, the power fall rate of the fuel cell system is restricted to prevent overpressure into the stack. $SOC_{max}$ is the upper bound of SOC and $SOC_{min}$ is the lower bound of SOC. As a conservative target, 0.8 and 0.4 are used in this study.

### C. Implementation and practical considerations

The fuel cell system has serval subsystems such as the gas supply subsystem, the humidifying subsystem, the temperature control subsystem and so on. The anode pressure, cathode pressure, the temperature and the moisture should be appropriately controlled. All these control algorithms are achieved through a so-called NetController, which has been developed by CASIA. To reduce the calculation work of NetController, the real time optimization problem is solved in MATLAB, which can also take advantage of MATLAB optimization toolbox. User Datagram Protocol (UDP) communication is adopted to exchange the necessary data between the NetContorller and MATLAB. The proposed real time optimization power management strategy here is shown in Fig. 5.

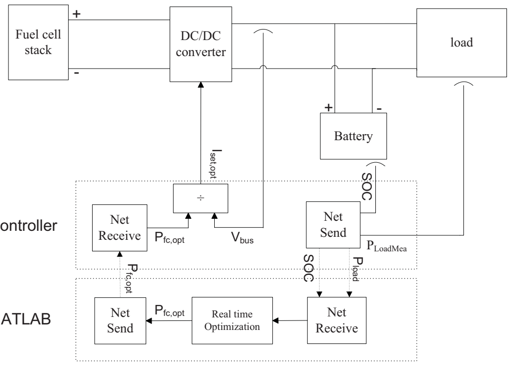


Fig. 5. Proposed power management strategy based on real time optimization

At each time $t_k$ with a time step $\Delta t_k$, the following steps are performed in real time optimization process:

- The load power and the battery SOC are measured by the sensor connected with NetController, and then, these values are sent to MATLAB by UDP communication program.
- MATLAB receives the values of load power and battery SOC, solve the optimization problem shown by 17) and get the optimal set point of fuel cell power $P_{fc,opt}(t_k)$ and the optimal value of $J(t_k)$.
- MATLAB sends the optimal value $P_{fc,opt}(t_k)$ to the NetController.
- The Netcontroller receives the value of $P_{fc,opt}(t_k)$, calculate the set point current of the DC/DC converter.

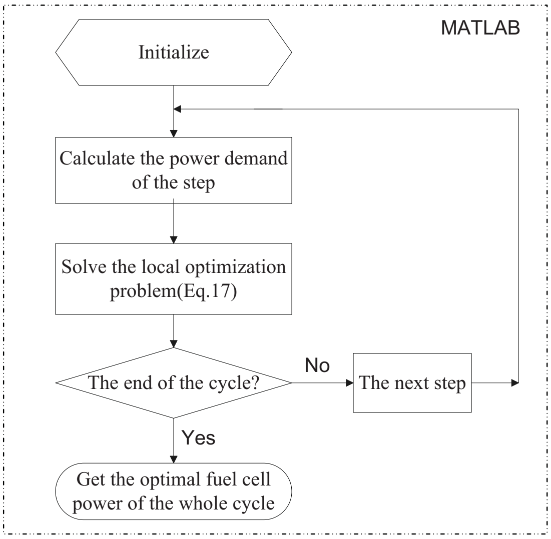

MATLAB
Initialize
Calculate the power demand of the step
Solve the local optimization problem(Eq.17)
The end of the cycle?
No
The next step
Yes
Get the optimal fuel cell power of the whole cycle

Fig. 6. The simulation flowchart of the proposed power management strategy

## IV. Optimization and experimental results

The proposed power management strategy based real time optimization are tested by simulation and experiments. Section

A is devoted to simulation results and analysis. The implementation of the proposed strategy in an experimental test setup is presented in section B.

## A. Simulation results

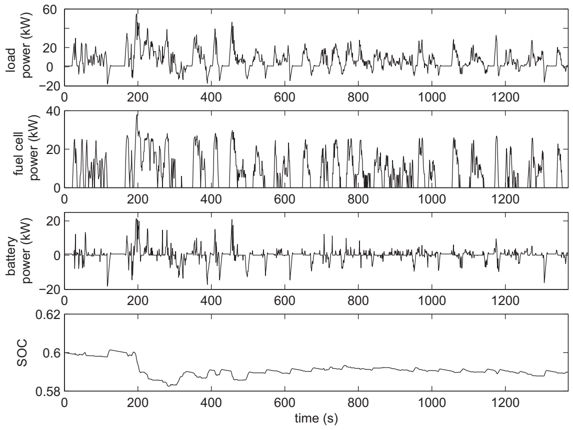


Fig. 7. The simulation results of the proposed strategy for UDDS cycle.

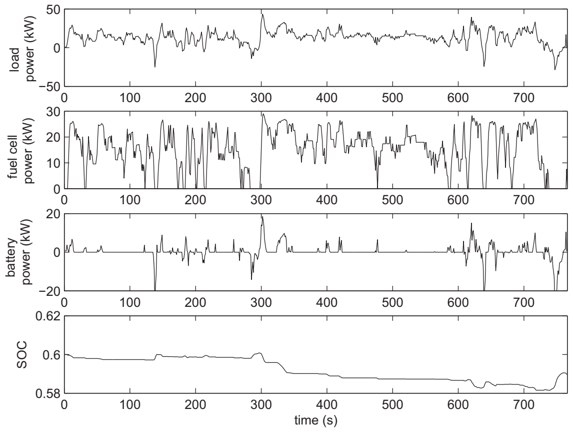


Fig. 8. The simulation results of the proposed strategy for HWFET cycle.

To compare the performance of the proposed real time optimization strategy with other power management strategies especially the global optimization control strategy, we simulated the power demand of a typical vehicle in three driving cycles: UDDS, HWFET, and NEDC. The optimization results of the proposed power management strategy and the optimal fuzzy power management strategy described in literature [14] are shown in Table I. In this table, the degree of hybridization (DOH) is defined as the ratio of electric power can be delivered by the energy storage system (here it means the battery pack) to the total power that can be delivered by ESS and fuel cell system [18]. The simulation is carried out in MATLAB, process of the simulation is shown in Fig. 6.

Fig. 7 shows the simulation results with the real time optimization power management strategy for UDDS cycle. The simulation results for HWFET cycle and NEDC cycle are shown in Fig. 8 and Fig. 9.

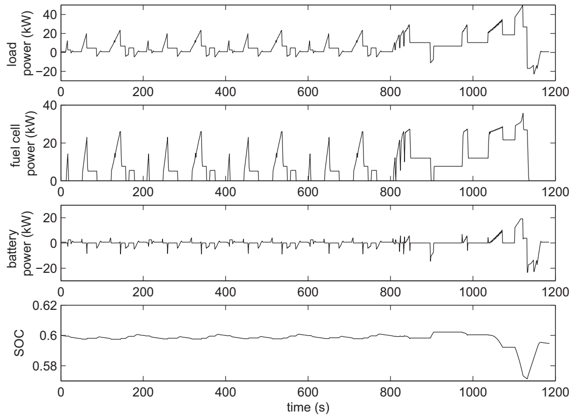


Fig. 9. The simulation results of the proposed strategy for NEDC cycle.

Optimization results in Table I show that the hydrogen economy of the real time optimization power management strategy is not as good as that of the global optimization. That is reasonable because the global optimization strategy is based on power demand of the entire driving cycle which is infeasible in practice, whereas the real time optimization is just based on the present power need.

## B. Experimental validation

In this subsection, the experimental results are presented to validate the feasibility and practicability of the proposed real time optimization power management strategy.

The experimental setup is composed of a PEM fuel cell test rig, a lead-acid battery pack, a constant-voltage restricted-current DC/DC converter and car lights to emulate the power consumption The PEM fuel cell test rig are designed and built by Institute of Automation, Chinese Academy of Sciences(CASIA). 24 cells are connected in series to make up the fuel cell stack. The voltage level can vary from 22V at no load to about 16V at full load. The rated power of this small fuel cell system is 150W and The max rise rate of the fuel cell is 30W/s. The DC/DC converter is connected after the fuel cell system to stable the output voltage and control the fuel cell output power. A 24 AH,12V lead acid is connected to dc bus. Car lights are used as the power load.

The optimization process is executed at every second and the experimental results are shown in Fig. 10. The results show that when the load power demand is the range that the fuel cell and battery can afford, the fuel cell system is apt to work at the power range that the system has the maximum efficiency.

## V. CONCLUSIONS

In this work, a new power management strategy based on real time optimization for fuel cell hybrid power sources system was addressed. Compared to the global optimization strategies that need the power demand of the entire task, which is not feasible in practice, the proposed strategy only

TABLE I

OPTIMIZATION RESULTS COMPARISON OF THE LOCAL OPTIMIZATION STRATEGY AND THE GLOBAL OPTIMIZATION STRATEGY

| Simulation outputs | Units | UDDS cycle | | HWFET cycle | | NEDC cycle | |
|---|---|---|---|---|---|---|---|
| | | local optimization strategy[a] | global optimization strategy | local optimization strategy[a] | global optimization strategy | local optimization strategy[a] | global optimization strategy |
| Total fuel consumption($H_2$)[b] | g | 145.7 | 142.6 | 163.6 | 154.5 | 125.9 | 123.1 |
| $\Delta SOC$ | | -0.0103 | 0.0249 | -0.0103 | 0.0299 | -0.0093 | 0.0058 |
| Cycle length | km | 11.99 | 11.99 | 16.51 | 16.51 | 10.94 | 10.94 |
| Specific energy consumption[c] | $MJkm^{-1}$ | 1.48 | 1.39 | 1.21 | 1.09 | 1.37 | 1.33 |

[a] Fixed DOH=0.3786.

[b] We use the lowest energy content of hydrogen, $120MJkg^{-1}$.

[c] The energy that the battery pack delivered or stored during the driving cycles is transformed to the hydrogen consumption by the concept of equivalent fuel consumption.

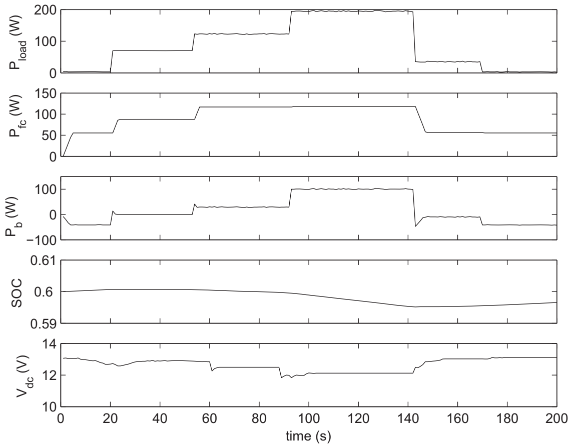


Fig. 10. Experimental results of the power management strategy based on real time optimization.

need the current power demand and can be easily applied in practice. The proposed strategy was tested both in simulation environment using three standard driving cycles and in an experiment. Compared to other power management strategies, the followings are verified: First, although the hydrogen economy of power management strategy based on real time optimization is not as good as the strategy based on global optimization, the proposed strategy still has a comparative good fuel economy. Second, if the power demand is near constant and the power demand deceleration is low, the fuel economy improved by optimization, no matter local optimization or global optimization, is insignificant. Finally, the experimental test shows that the proposed strategy is very easy to be implemented in practice.

## REFERENCES

[1] M. M. Hussain, I. Dincer, and X. Li, "A preliminary life cycle assessment of pem fuel cell powered automobiles," *Appl. Therm. Eng*, vol. 27, no. 13, pp. 2294 – 2299, 2007.

[2] L. Schlecht, "Competition and alliances in fuel cell power train development," *Int. J. Hydrogen Energy*, vol. 28, no. 7, pp. 717 – 723, 2003, market Challenges of Fuel Cell Commercialization.

[3] P. Costamagna and S. Srinivasan, "Quantum jumps in the pemfc science and technology from the 1960s to the year 2000: Part ii. engineering, technology development and application aspects," *J. Power Sources*, vol. 102, no. 1-2, pp. 253 – 269, 2001.

[4] D. Chu, R. Jiang, K. Gardner, R. Jacobs, J. Schmidt, T. Quakenbush, and J. Stephens, "Polymer electrolyte membrane fuel cells for communication applications," *J. Power Sources*, vol. 96, no. 1, pp. 174 – 178, 2001.

[5] F. Barbir and T. Gomez, "Efficiency and economics of proton exchange membrane (pem) fuel cells," *Int. J. Hydrogen Energy*, vol. 22, no. 10-11, pp. 1027 – 1037, 1997.

[6] M. Uzunoglu and M. Alam, "Dynamic modeling, design, and simulation of a combined pem fuel cell and ultracapacitor system for stand-alone residential applications," *IEEE Trans. Energy Convers.*, vol. 21, no. 3, pp. 767–775, Sept. 2006.

[7] M. Kim, Y.-J. Sohn, W.-Y. Lee, and C.-S. Kim, "Fuzzy control based engine sizing optimization for a fuel cell/battery hybrid mini-bus," *J.Power Sources*, vol. 178, no. 2, pp. 706 – 710, 2008.

[8] P. Thounthong, S. Rael, and B. Davat, "Analysis of supercapacitor as second source based on fuel cell power generation," *IEEE Trans. Energy Convers.*, vol. 24, no. 1, pp. 247–255, March 2009.

[9] A. Brahma, Y. Guezennec, and G. Rizzoni, "Optimal energy management in series hybrid electric vehicles," *in Proc. 2000 American Control Conf.*, vol. 1, no. 6, pp. 60–64, Sep 2000.

[10] S. Delprat, T. Guerra, G. Paganelli, J. Lauber, and M. Delhom, "Control strategy optimization for an hybrid parallel powertrain," *in Proc. 2001 American Control Conf.*, vol. 2, pp. 1315–1320, 2001.

[11] S. Delprat, T. Guerra, and J. Rimaux, "Optimal control of a parallel powertrain: from global optimization to real time control strategy," *IEEE 55th Vehicular Technology Conf., 2002.*, vol. 4, pp. 2082–2088, 2002.

[12] A. Sciarretta, M. Back, and L. Guzzella, "Optimal control of parallel hybrid electric vehicles," *IEEE Trans. Control Syst. Technol.*, vol. 12, no. 3, pp. 352–363, May 2004.

[13] A. L. P.Rodatz, G.Paganelli, "Optimal power management of an experimental fuel cell/supercapacitor-powered hybrid vehicle," *Control Eng. Pract.*, vol. 13, no. 1, pp. 41 – 53, 2005.

[14] C.-Y. Li and G.-P. Liu, "Optimal fuzzy power control and management of fuel cell/battery hybrid vehicles," *J. Power Sources*, vol. 192, no. 2, pp. 525 – 533, 2009.

[15] V. H. Johnson, "Battery performance models in advisor," *J. Power Sources*, vol. 110, no. 2, pp. 321 – 329, 2002.

[16] G. Paganelli, S. Delprat, T. Guerra, J. Rimaux, and J. Santin, "Equivalent consumption minimization strategy for parallel hybrid powertrains," *IEEE 55th Vehicular Technology Conf., 2002.*, vol. 4, pp. 2076–2081, 2002.

[17] D. Feroldi, M. Serra, and J. Riera, "Energy management strategies based on efficiency map for fuel cell hybrid vehicles," *J. Power Sources*, vol. 190, no. 2, pp. 387 – 401, 2009.

[18] R. K. Ahluwalia, X. Wang, and A. Rousseau, "Fuel economy of hybrid fuel-cell vehicles," *J. Power Sources*, vol. 152, pp. 233 – 244, 2005.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# A Distributed Newton Iteration Based Localization Scheme in Underground Tunnels

Yuanqing Qin, Chunjie Zhou
Dept. of Control Science & Engineering
Huazhong University of Science & Technology
Wuhan, China
e-mail: qinyuanqing@mail.hust.edu.cn, cjiezhou@126.com

Shuang-Hua Yang, Fang Wang
Computer Science Department
Loughboroug University
Loughborough, UK
e-mail: s.h.yang@lboro.ac.uk, scarlew@sina.com

*Abstract*— **One of the main concerns in underground working tunnels is ensuring the safety of the workers and their equipment. Being aware of the real-time position of personnel in such harsh environment is challenging and requires a sophisticated localization system. With traditional Received Signal Strength (RSS) failing to accurately estimate the distance between nodes due to multipath effect in such long and narrow space, Radio Frequency Time-of-Flight (RF-TOF) is proved to be an alternative method for more accurate distance estimation. To reduce the communication cost, a distributed localization scheme is proposed, where a simple Newton Iteration location estimation algorithm is embedded in the blind node. Linear least square estimation is used as the initial value to accelerate the convergence of the iteration. Experimental results show the effectiveness of the proposed scheme.**

*Keywords- **wireless sensor networks; localization; time of flight; Newton Iteration***

## I. INTRODUCTION

With the emergence of various location-based services and other potential application in wireless communication networks, localization in wireless networks has received a great deal of attention in the past decades. Commercial examples range from low-accuracy methods based on cell identification to high-accuracy methods combining wireless network information and satellite positioning [1]. These methods are typically network centric, where the position is determined in the network and presented to the user via a specific service. Applications of such kind of systems are limited where infrastructures or signal coverage are not perfect. Wireless Sensor Networks (WSNs) provide another option to localize targets in their covered area, which is an important complement to the infrastructure based wireless localization systems.

In this paper, a special environment, an underground working tunnel, is focused. Underground working tunnels referred here are railway or road tunnels which are under construction, or coal mines. The common characteristics of such environment are as follows: the space is usually long and narrow, with length of several kilometres and width of several meters, and its structure is changing with construction or production; the power supply and the communication infrastructure are not always available, and there are usually no reliable wired or wireless communication link; the tunnels are underground or in mountain bodies, such that the humidity is high, the air is dirty due to the dust and there are even dangerous gases, such as methane, carbon dioxide, carbon monoxide etc.; the environment is noisy and is full of equipment and workers. These characteristics make the tunnel under construction a dangerous working environment. Accidents often happen causing severe casualty and capital lost. It is urgent to establish an advanced monitoring system, which can obtain the real-time information about the worker and the environment, evaluate the risk level to safeguard the workers.

A distributed localization scheme is proposed in this paper. Using JN5148 wireless module which is embedded with a RF-TOF ranging engine [2], a Newton Iteration based localization algorithm is designed and implemented on the blind node. With no overhead hardware requirement and the distributed characteristics, a cheap but efficient localization system adaptable to constructing tunnel environment can be achieved.

The rest of the paper is organized as follows. In section II, popular localization methods of WSN and the state of the art of tunnel localization are briefly reviewed. Section III presents a Newton Iteration localization algorithm with Linear Least Square Estimation (LLSE) as initial value and a distributed localization scheme is proposed in section IV. Section V demonstrates the experimental results and concluding remarks are made in section VI.

## II. RELATED WORKS

### A. Localization Techniques in WSN

The subject of localization in wireless sensor networks has been drawing considerable attention due to its potential applications, such as inventory tracking, intruder detection, tracking of fire-fighters and miners, home automation and patient monitoring etc. [3, 4]. These potential applications of wireless positioning were also recognized by IEEE, which approved a new amendment, IEEE 802.15.4a, that provides a new physical layer for low data rate communications combined with positioning capabilities [5, 6].

Depending on the mechanisms used, localization schemes in wireless networks can be classified into two categories:

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

range-based and range-free. Range-based approach involves estimation of location in two steps. In the first step, location related parameters, such as Time of Flight (TOF) [7, 8] of signals traveling between the target node (or blind node), i.e. the node to be located, and a number of reference nodes (or anchor nodes) are estimated. Then, in the second step, the location is estimated based on the signal parameters obtained in the first step. The location related parameters estimated in the first step include Received Signal Strength (RSS) [9], Time of Arrival (TOA), Time Difference of Arrival (TDOA) [10, 11], Near Field Electromagnetic Ranging (NFER) [12], which provide an estimation of distance, and Angle Of Arrival (AOA) [13], which estimates the angle between the nodes. For distance based localization algorithms, the maximum likelihood (ML) solution can be obtained by a Nonlinear Least Squares (NLS) approach, under certain conditions [1].

Range-free localization schemes do not need distance or angle information, but performs the localization algorithm based on the connection characteristics and anchor nodes' location information instead. There are some typical algorithms such as centric algorithm [14], DV-HOP algorithm [15], Area-based Point-In-Triangulation Test algorithm (APIT) [16] etc. Range-free localization schemes do not need overhead hardware, so that they are cost-effective and power-effective. But they are usually central schemes and are suitable for simple topology and high densities networks only.

### *B. State of Art of Tunnel Localization*

Range-free localization schemes are not suitable for tunnel environment because of the low density and complex topology of WSN. RSS range-based localization methods have been studied in coal mine galleries [17]. Qiao proposed a dynamic RSS localization algorithm for chain-type WSN in tunnels [18], in which the distance and the corresponding RSS between the adjacent beacon nodes were taken into account to get a better path loss parameter. Impulse Radio Ultra Wideband (IR-UWB) is a promising technology for indoor localization applications due to its high-temporal resolution, multipath immunity, and simultaneous ranging and communication capability. But the receivers need to be connected by cable for high accurate synchronization requirement. Zhou proposed an asynchronous position measurement scheme for indoor localization by adding an additional UWB transmitter besides the anchor nodes and the target nodes [19]. The challenge is that the high accuracy can only be achieved in a smaller coverage. Chehri studied the feasibility of using UWB-based WSN as future solution for localization in underground mine via simulation and measurement [20].

Fingerprinting technique was used in mine localization to avoid the difficulty of measuring distance or angle in such harsh environment [21]. The main disadvantage of such methods is the requirement that the training database should be large enough and representative of the current environment for accurate localization. In underground working tunnels, such data collection task can be laborious or even impossible because of the dynamic change of the structure.

Localization schemes can be categorized into centralized and distributed based on the communications between nodes. Centralized schemes involve transmitting all measured data to a central node to compute the location of the target node and the central node has enough computation resources to carry out complicated localization algorithms. Distributed localization schemes do not require centralized computation, and rely on each node to calculate its location with only limited communications with nearby limited nodes. Distributed scheme is more suitable for tunnel environment where the communication cost to the central node is much higher and the time delay is much bigger because of multi-hop transmission.

### *C. Radio Frequency Time of Flight Ranging*

RF-TOF refers to the time needed for a message to be sent from one node to another. Since the spread velocity of radio is invariable, which is $3\times10^8$ m/s, with RF-TOF obtained, the distance between two nodes can be calculated easily. With the same transceiver used for data communication, an RF-TOF ranging engine requires little hardware overhead and can achieve meter level accuracy in complex environments. RF-TOF ranging occurs in short bursts and in a frequency hopped fashion thereby reducing the chance of interference. RF-TOF ranging is such an attractive option for WSNs that some prototypes have been demonstrated, but work has been largely limited to wide bandwidths and high power devices [22].Optimization of RF-TOF for WSNs has recently received attention with some interesting results in the wideband signal domain [6]. In bandwidth limited systems, measuring the TOF requires accurately resolving the phase offset of a signal. Pseudorandom Noise (PN) codes are good candidate signals for measuring small phase offsets because the autocorrelation function of a PN code exhibits a single large peak that moves with phase offset. Reference [7] proposed a pair-wise ranging called Code Modulus Synchronization (CMS) that does not require either node to determine the absolute phase offset of system clocks, the correlation function or the TOF in real time. This reduces the hardware overhead and measurement time by not requiring a real time co-relator.

Two classes of RF-TOF measurement systems exist. The first is a scheme where a number of significant devices have highly accurate, synchronized clocks. In the simplest case, a signal is sent from a device with a known location and an accurate clock to another device with an accurate clock, and the departure time of the signal is compared to the actual time of arrival. This scheme is not practical in WSNs due to the high accuracy requirement to the hardware. The second type of RF-TOF system is a pair-wise round-trip measurement, which does not require absolute clock. By sending a ranging signal and waiting for a reply, the individual clock biases are subtracted away. Reference [23] proposed a two-way TOF ranging scheme using narrow-band RF chip CC2430.

In this paper, JN5148 microcontroller is utilized as the CPU of wireless sensor nodes, as it embeds an RF-TOF ranging engine [2], which is an alternative of RSS to estimate distance between two nodes without overhead hardware. A chain type of Zigbee network is deployed in underground tunnels, and a distributed Newton Iteration based localization algorithm is designed on the blind node.

## III. Newton Iteration Based Localization Algorithm

### A. Traditional Trilateration Algorithm

In range-based localization scheme, the position of a blind node can be determined with the knowledge of the distance to its neighbouring anchors (i.e. reference nodes) and the coordinates of those anchors:

$$(x-x_i)^2+(y-y_i)^2=d_i^{\,2}, \quad i=1,2,\cdots,N \tag{1}$$

Where $(x, y)$ is the coordinates of the blind node;

$(x_i, y_i)$ is the coordinates of the ith reference node;

$d_i$ is the distance between the blind node and the ith reference node and N is the number of the reference nodes.

In the absence of noise in a system, each distance measurement specifies a circle for the possible positions of the blind node, and the intersection of those circles determines the target position. This geometric technique, called trilateration, yields ambiguous solutions in the presence of noise in the system, since the circles defined by (1) may intersect at multiple points due to erroneous distance estimation. A popular statistical localization algorithm is the Nonlinear Least Squares (NLS) techniques, by which the location of the blind node is calculated as follows:

$$\begin{aligned}[\hat{x},\hat{y}] &= \arg\min_{(x,y)} s(x,y)\\ &= \arg\min_{(x,y)} \sum_{i=1}^{N} \beta_i\left(\sqrt{(x-x_i)^2+(y-y_i)^2}-d_i\right)^2\end{aligned} \tag{2}$$

Where $s(x, y)$ is the cost function, $N \geq 3$ is the number of the reference nodes, and $\beta_i$ represents a weighted coefficient for the ith measurement, which commonly reflects the reliability of the measurement. The solution of (2) usually requires numerical search methods such as the steepest descent or the Gauss-Newton techniques, which can have high computational complexity and typically requires good initial value in order to avoid converging to the local minima of the cost function.

Alternatively, Linear Least Square Estimation (LLSE) can provide suboptimal location estimation with low computational complexity. Let the rth equation represented in (1) subtract all the other equations, (i.e. equation 1, 2, $\cdots$, r-1, r+1, $\cdots$, n), the following linear relation can be obtained:

$$AX=b$$

Where $X=[x, y]^T$ is the location of the blind node,

$$A=2\begin{bmatrix} x_1-x_r & y_1-y_r\\ \vdots & \vdots\\ x_{r-1}-x_r & y_{r-1}-y_r\\ x_{r+1}-x_r & y_{r+1}-y_r\\ \vdots & \vdots\\ x_N-x_r & y_N-y_r \end{bmatrix} \tag{3}$$

$$b=\begin{bmatrix} d_r^2-(x_r^2+y_r^2)-d_1^2+(x_1^2+y_1^2)\\ \vdots\\ d_r^2-(x_r^2+y_r^2)-d_{r-1}^2+(x_{r-1}^2+y_{r-1}^2)\\ d_r^2-(x_r^2+y_r^2)-d_{r+1}^2+(x_{r+1}^2+y_{r+1}^2)\\ \vdots\\ d_r^2-(x_r^2+y_r^2)-d_N^2+(x_N^2+y_N^2) \end{bmatrix} \tag{4}$$

Note that A is an $(N-1)\times 2$ matrix, and b is a $(N-1)$ vector, since the rth measurement is used as a reference for all the other measurement. Then the LLSE can be obtained as

$$\hat{X}=(A^TA)^{-1}A^Tb \tag{5}$$

Reference [24] analysed the performance of several LLSE algorithms, where different information is used as reference. Reference [11] proposed a linear suboptimal location estimation algorithm by constructing a triangle and selecting the best estimation from Seven Potential Estimation (SPE) according to the cost function. With relatively low computational complexity, such algorithms can be implemented in a distributed way.

### B. Newton Iteration with LLSE as Initial Value

To obtain more precise estimation, high-accuracy techniques, such as NLS approach and linearization based on Taylor series can be considered. A good initial value can make the sequence converge quickly and significantly reduce the calculation complexity, thus making it possible to be implemented in a distributed way. In this paper, LLSE solution provided by (5) is utilized as the initial value to accelerate the convergence of Newton iteration, LLSNI for brevity.

Calculate the partial derivatives of $s(x, y)$ in (2) with respect to x and y, denoted as $f(x, y)$ and $g(x, y)$; and let them equal to zero:

$$\begin{cases} f(x,y)=\dfrac{\partial s(x,y)}{\partial x}=\displaystyle\sum_{i=1}^{N}\beta_i\left[2(x-x_i)-\dfrac{2r_i(x-x_i)}{\sqrt{(x-x_i)^2+(y-y_i)^2}}\right]=0\\ g(x,y)=\dfrac{\partial s(x,y)}{\partial y}=\displaystyle\sum_{i=1}^{N}\beta_i\left[2(y-y_i)-\dfrac{2r_i(y-y_i)}{\sqrt{(x-x_i)^2+(y-y_i)^2}}\right]=0 \end{cases} \tag{6}$$

Then the Newton-iteration equation is:

$$\begin{cases} x_{k+1}=x_k+\dfrac{f(x_k,y_k)g_y(x_k,y_k)-g(x_k,y_k)f_y(x_k,y_k)}{g_x(x_k,y_k)f_y(x_k,y_k)-f_x(x_k,y_k)g_y(x_k,y_k)}\\ y_{k+1}=y_k+\dfrac{g(x_k,y_k)f_x(x_k,y_k)-f(x_k,y_k)g_x(x_k,y_k)}{g_x(x_k,y_k)f_y(x_k,y_k)-f_x(x_k,y_k)g_y(x_k,y_k)} \end{cases} \tag{7}$$

Where $f_x(x,y)$, $f_y(x,y)$, $g_x(x,y)$ and $g_y(x,y)$ represent partial derivatives of $f(x,y)$ and $g(x,y)$ with respect to x and y. The detail is omitted for the sake of simplicity. One numerical solution of (6) is

$$\begin{cases} x=\lim\limits_{k\to\infty} x_k\\ y=\lim\limits_{k\to\infty} y_k \end{cases}$$

when the Newton sequence is convergent.

RSS information is included in a RF-TOF ranging package [25]. According to experiment, large RSS value means the distance between two nodes is short and thus the RF-TOF ranging is relatively reliable. Therefore, we choose the measurement with the largest RSS value as the reference in LLSE algorithm, i.e. r in (3) and (4) is defined as:

$$r = \arg\max_{i}(RSS_i)$$

and RSS information is also utilized to calculate the weighting coefficient in the cost function:

$$\beta_i = RSS_i \Big/ \sum_{j=1}^{N} RSS_j \qquad (8)$$

After obtaining all the information needed for localization, including the coordinates of the reference nodes, distance between the blind nodes and each reference node measured by RF-TOF ranging engine, and RSS value when performing RF-TOF, the blind node calculates LLSE according to (5) and uses it as the initial value of Newton Iteration, which is used to calculate the minimum point of the cost function $s(x, y)$ according to (7). Finally, the blind node reports the localization result to the surveillance centre.

## IV. A Distributed Localization System in Underground Tunnels

### A. Architecture of The Localization System

The localization system proposed in this paper consists of a surveillance PC, a coordinator, anchor nodes and one or more blind nodes. The structure of the system is shown in Fig. 1.

The WSN in tunnels is a ZigBee network and the coordinator is responsible for establishing the network. The coordinator also acts as a gateway to the surveillance PC through a serial port. The surveillance PC is responsible for the configuration of the anchor nodes and localization data management.

The anchor nodes collect data of tunnel environment and participate in localization. Anchor nodes are routers of the ZigBee network. The blind node performs a distributed localization algorithm. There can be one or more blind nodes in WSN simultaneously and they must be routers of the ZigBee network, because they need to communicate with multiple anchor nodes directly within their communication range.

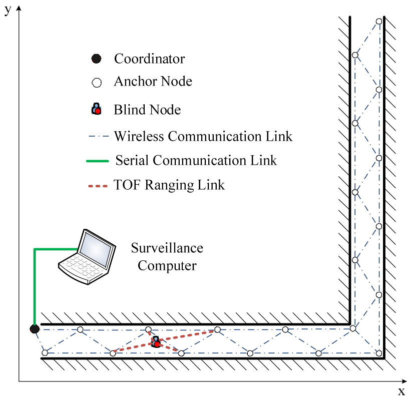


Figure 1. Architecture of the proposed localization system

There is a configurable timer on the blind nodes. When the timer expires, the blind nodes execute the localization task and report the result to the coordinator, and then the timer is restarted again.

### B. Deployment and Configuration of The System

To ensure the network communication having certain redundancy and the blind node finding at least 4 reference nodes, the anchor nodes should be deployed along both sides of the tunnel. The distance between any two adjacent nodes on the same side remains the same, and it should be shorter than their valid communication range. The anchor nodes on different sides should be placed alternately, in other words, one anchor node on one side is to be placed in the middle point of two nodes on the opposite side, as shown in Fig. 3.

There are two parameters that should be configured before the localization system works: the ID number and the coordinates of each anchor node. These parameters should be non-volatile. A unique ID number for each anchor node is defined and configured after the node joins the network. The serial numbers of the anchor nodes on one side are all odd numbers, and are all even numbers on the other side. This rule helps the blind node to choose proper reference nodes on both sides, because if all the reference nodes are on the same side, which means they may be in a line, it will lead to a failure of our localization algorithm.

### C. Distributed Localization Scheme

A distributed localization algorithm is designed and implemented on the blind node, consisting of the following steps:

Step 1) When the blind node enters into the area covered by the ZigBee network, it requests to join the network as a router. After joining the network successfully, the blind node broadcasts a localization request in one hop range, starts a timeout timer and waits for the anchor nodes' response. If the following conditions are satisfied, turn to step2:

Condition 1: There are at least four anchor nodes responding the request with their ID numbers and coordinates, which are the reference nodes for this localization;

Condition 2: Two of the reference nodes have odd ID numbers and the other two have even numbers. It ensures that the reference nodes are not in a line;

If the timeout timer expires before the above conditions are met, the blind node reports the information of not finding enough reference nodes to the coordinator and repeat Step 1.

Step 2) The blind node uses the RF-TOF engine to measure the distance between each reference node and itself. To minimize the measurement errors caused by clock shifts between different nodes, a bidirectional round trip measurement strategy is adopted: the blind node performs M times forward measurement and M times reverse measurement [25]. The average of these 2M results is regarded as the final value. The parameter M can be configured through the surveillance software, and usually within the range of 5~10.

Step 3) The blind node estimates its own coordinates according to the distances between itself and each reference node and the coordinates of the reference nodes, using LLSNI algorithm presented in section III.

Step 4) The blind node reports its coordinates to the coordinator.

When the coordinator receives the localization message, it will hand the message over to the surveillance PC immediately.

## V. Experimental Results

The proposed localization system is tested in an abandoned air-raid shelter. It has an "L" shape, as shown in Fig. 4, with similar environment characteristics to underground tunnels.

### A. *Rang Accuracy Comparison between RF-TOF and RSS*

To show the advantage of RF-TOF ranging method in tunnel environment, a contrast measuring experiment was carried out in a point to point way, using RF-TOF and RSS ranging methods respectively. Both two nodes use the same wireless module, JN5148-001-M03 with a standard power. One is placed at the entry of the air-raid shelter and the other is moving along the air-raid shelter. At each distance, 20 times of TOF ranging and 20 times of RSS ranging were performed. The distance calculation formula is as the following:

$$d = TOF \times 0.0003 \qquad \text{(TOF ranging)}$$

$$d = 0.02 \times 10^{(\frac{108-RSS}{20})} \qquad \text{(RSS ranging)[25]}$$

The average and the standard deviation of the measuring results are shown in Table I. As can be seen from the result, the distance estimation errors according to RSSI increase significantly and the standard deviation of multiple measurements increases with the increase of distance. The energy of the radio signal distorts seriously because of the multi-path fading effect. Therefore, RSSI is not suitable for distance estimation in tunnel environment. On the contrary, TOF ranging shows excellent performance with little standard deviation and the ranging error does not increase significantly with the increase of distance. Most of the ranging errors are less than 3 metres. It proves the choice of our RF-TOF for the study.

### B. *Localization Experiment in Air-raid Shelter*

The structure of an air-raid shelter we used is shown as Fig. 1. The length in X axis and Y axis is 150 meters respectively and the width is 5 meters. The distance between the adjacent anchor nodes on the same side is 30 meters, and the deployment of the two sides is alternate. The ZigBee localization network consists of one coordinator and 21 routers, which are all based on JN5148-M03 wireless modules. One of the routers acts as the blind node, which is needed to be localized in real time. Locations were estimated at 14 test points, with each point being localized 20 times.

TABLE I. The Accuracy Comparison of TOF and RSS Ranging

| **Real distance(m)** | | 10 | 20 | 30 | 40 | 50 | 60 |
|---|---|---|---|---|---|---|---|
| **TOF ranging** | ***Average(m)*** | 11.5 | 19.4 | 29.3 | 41.8 | 48.9 | 58.2 |
| | ***Standard deviation(m)*** | 0.7 | 2.1 | 2.2 | 2.5 | 2.6 | 2.8 |
| **RSS ranging** | ***Average(m)*** | 7.1 | 48.5 | 91.4 | 113.5 | 120.8 | 104.7 |
| | ***Standard deviation(m)*** | 0.4 | 14.1 | 28.9 | 19.6 | 23.9 | 64.8 |

The termination condition of Newton Iteration is

$$\sqrt{(x_{k+1}-x_k)^2+(y_{k+1}-y_k)^2} < 0.001$$

or the iteration has been performed 100 times. Two kinds of initial values were tested: random initial value and LLSE as the initial value. With random initial value, 81% of Newton Iterations converged after 18.5 iterations in average, and 19% reached the maximum iteration boundary. With LLSE as the initial value, 100% of Newton iterations converged after averaged 5.8 iterations in average. The experimental results proved that LLSE initial value accelerated the convergence of Newton Iteration.

Other two existing algorithms, LLSE and SPE, were implemented as well in the same experiment environment.

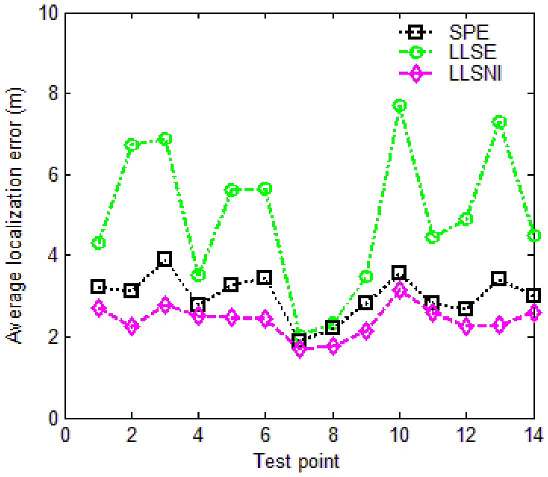


Figure 2. Average localization errors of LLSNI, LLSE and SPE

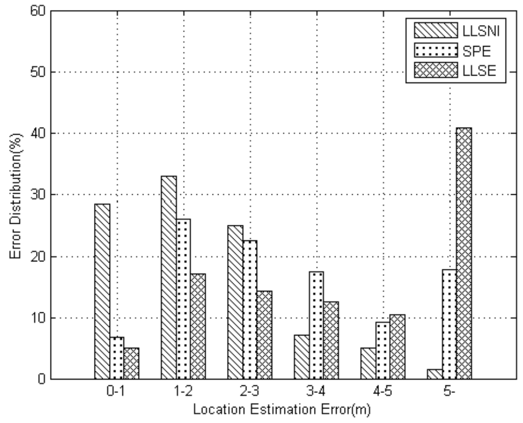


Figure 3. Localization error distribution of LLSE, SPE and LLSNI

These three algorithms were tested under the same condition. Fig. 2 showed the average localization errors of these three algorithms at each test point. Three curves had the similar trend, which means that large range error degraded the performance of all those localization algorithms, but the influence to LLSNI was much smaller than the influence to the other two algorithms. Fig. 3 showed the error distribution of those three algorithms. As can be seen, LLSNI algorithm outperforms theother two algorithms with acceptable computation increase, and 86.4% of the localization errors are less than 3 meters.

## VI. Conclusions

According to the special characteristics of the underground working tunnels, a distributed range-based localization scheme is proposed. RF-TOF range engine embedded in JN5148 microcontroller is utilized to estimate the distances between nodes. A Newton Iteration location estimation algorithm is proposed, with LLSE as the initial value to accelerate the convergence. With low calculation complexity, the localization algorithm can be embedded in the blind node and only the localization result need to be transmitted to the coordinator. This distributed scheme is especially meaningful in multi-hop WSN in underground tunnels, because it can greatly reduce the communication cost and improve the real-time performance.

Experimental results show that the proposed system can provide precise distributed localization without any overhead hardware, which enables the establishment of a cheap but effective constructing tunnel surveillance system to safeguard the workers there.

Several research challenges remain to be addressed. None line of sight propagation is not taken into account in our scheme, which can be a main cause of range error. Some localization results are clearly outlier due to the range error. How to assess and improve the localization performance with geographic information of the tunnels can be considered. Time delay of communication is another issue that should be addressed when the scale of the system becomes bigger.

## Acknowledgment

This work was supported by the NSFC grants 61074145. Yuanqing Qin thanks Chinese Scholar Council for their support of academic visiting to Loughborough University.

## Reference

[1] F. Gustafsson and F. Gunnarsson. “Mobile positioning using wireless networks: possibilities and fundamental limitations based on available wireless network measurements,” IEEE Signal Processing Magazine, vol. 22, Jun. 2005, pp.41-53

[2] JN5148-001 data sheet. [Online]. Available: http://www.jennic.com/files/product_briefs/JN-DS-JN5148-1v6.pdf

[3] IEEE 15-03-0489-03-004a application requirement analysis-031127 v0.4. [Online]. Available: http://www.802.org/15/pub/TG4.html

[4] S. Gezici, Z. Tian, G. B. Giannakis, H. Kobayashi, A. F. Molisch, H. V. Poor and Z. Sahinoglu. “Localization via ultra-wideband radios: a look at positioning aspects for future sensor networks,” IEEE Signal Processing Magazine, vol. 22, Jun. 2005, pp.70-84

[5] IEEE P802.15.4a/D4 (Amendment of IEEE Std 802.15.4), Part 15.4: wireless medium access control (MAC) and physical layer (PHY) specifications for low-rate wireless personal area networks (LRWPANs). 2006

[6] Y. S. Kwok, F. Chin and X. Pen. “Ranging mechanism, preamble generation, and performance with IEEE 802.15.4a low-rate low-power UWB system,” The IEEE International Conference on Ultra-Wideband,Agu.2006, pp.525-530

[7] T. C. Karalar and J. Rabaey. “An RF TOF based ranging implementation for sensor networks,” In IEEE Int. Conf. on Communications, Jun. 2006, Vol. 7, pp. 3347-3352

[8] S. Lanzisera, D. T. Lin and K. S. J. Pister. “RF time-of-flight ranging for wireless sensor network localization,” In Int. workshop on Intelligent Solutions in Embedded systems, Jun. 2006, pp. 1-12

[9] X. Li. “Performance study of RSS-based location estimation techniques for wireless sensor networks,” In Military Communications Conference, Oct. 2005, pp. 1064-1068

[10] R. J. Fontana, E. Richley and J. Barney. “Commercialization of an ultra-wideband precision asset location system,” In IEEE Conf. on Ultra Wideband Systems and Technologies, Nov. 2003, pp. 369-373

[11] Z. M. Merhi, M. A. Elgamel and M. A. Bayoumi. “A lightweight collaborative fault tolerant target localization system for wireless sensor networks,” IEEE Transactions on Mobile Computing, vol. 8, Dec. 2009, pp.1690-1703

[12] H. G. Schantz. “A real-time location system using near-field electromagnetic ranging,” IEEE Antennas and Propagation Society International Symposium, Jun. 2007, pp. 3792-3795

[13] P. Rong and M. L. Sichitiu. “Angle of arrival localization for wireless sensor networks,” In 2006 3rd Annual IEEE Communications Society on Sensor and Ad Hoc Communications and Networks, Setp.2006, pp. 374-382

[14] N. Bulusu, J. Heidemann and D. Estin. “GPS-less low cost outdoor localization for very small devices,” IEEE Personal Communication Magazine, vol. 7, Oct. 2000, pp.28-34

[15] D. Niculescu and B. Nath. “DV based positioning in Ad hoc networks,” Journal of Telecommunication systems, vol. 22, Apr. 2003, pp.267-280

[16] T. He, C. D. Huang and B. M. Blum. “Range-free localization schemes for large scale sensor networks,” In Proc. of the 9th annual Int. Conf. on Mobile Computing and Networking, Sept. 2003, pp.81-95

[17] X. Liu, Z. Wang, S. Wang and X. Zhou. “Study of WSN localization based on RSSI in coal mine,” Coal Mine Machinery, vol. 30, Mar. 2009, pp. 59-60(in Chinese)

[18] Qiao G Z, Zeng J C. “Localization algorithm of beacon nodes chain deployment based on coal mine underground wireless sensor networks,” Journal of China Coal Society, vol. 35, Jul. 2010, pp.1229-1233(in Chinese)

[19] Y. Zhou, C. L. Law, Y. L. Guan and F. Chin. “Indoor Elliptical Localization Based on Asynchronous UWB Range Measurement,” IEEE Transactions on Instrumentation and Measurement, vol. 60, Jan. 2011, pp. 248-257

[20] A. Chehri, P. Fortier and P. M. Tardif. “UWB-based Sensor Networks for Localization in Mining Environment,” Ad Hoc Networks, vol. 7, Jul. 2009, pp.987-1000

[21] C. Nerguizian, C. Despins and S. Affes. “Geolocation in Mines With an Impulse Response Fingerprinting Technique and Neural Networks,” IEEE Transactions on Wireless Communications, vol. 5, Mar. 2006, pp. 603-611

[22] D. McCrady, L. Doyle, H. Forstrom, T. Dempsey and M. Martorana. “Mobile ranging using low-accuracy clocks,” IEEE Transactions on Microwave Theory and Techniques, vol. 48, Jun. 2000, pp.951-958

[23] B. Thorbjornsen, N. M. White, A. D. Brown and J. S. Reeve. “Radio frequency (RF) time-of-flight ranging for wireless sensor networks,” Measurement Science and Technology, vol. 21, Mar. 2010, pp.1-12

[24] S. Gezici, I. Guvenc and Z. Sahinoglu. “On the performance of linear least-squares estimation in wireless positioning systems,” In IEEE Int. Conf. on Communications, May 2008, pp.4203-4208

[25] Jennic Time-of-Flight API User Guide, provided by Jennic support, 2009

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Modeling and Simulation of a Coupled Double-Loop-Cooling System for PEM-Fuel Cell Stack Cooling

Martin Schultze, Michael Kirsten, Sven Helmker and Joachim Horn
Institute for Control Engineering
Helmut-Schmidt-University Hamburg
D-22043 Hamburg, Germany

*Abstract*—**PEM (Polymer electrolyte membrane) fuel cell systems are very efficient energy converters. They generate electrical power, low oxygen concentration cathode exhaust gas, water as well as heat. The fuel cell technology has become very attractive for the use on aircraft where it may serve as a replacement for the auxiliary power unit currently being used for generating electrical power. For the use on aircraft coupled double-loop-cooling systems are investigated as different coolants can be used for inner and outer cooling system. As fuel cell electrical power is nonlinearly dependent on stack temperature and current, a cooling temperature control is required.**

**In this study a nonlinear simulation model of a coupled double-loop-cooling system is presented. The effectiveness-NTU method is used to model the intercooler that couples inner and outer cooling loop. This method has the advantage that outlet temperatures are obtained explicitly based on inlet coolant temperatures and cooling mass flows. The model is valid over the entire operating range of the hydrogen fed fuel cell system. Subsequently, a linear controller with feedforward control is proposed to control for the stack inlet cooling temperature.**

***PEM fuel cell system model; nonlinear heat exchanger model; model identification; fuel cell temperature control***

## I. Introduction

For aircraft applications the fuel cell technology is investigated regarding to its multifunctional use. One of its main aspects is the use of oxygen depleted cathode exhaust air (ODA) for tank inerting [1]. For this purpose the oxygen content needs to be close or less than 10% (vol.) [2]. Another and certainly one of the most important aspects of the multifunctional use is electrical power generation. So far, auxiliary power units (APU) generate electrical power for the use on aircraft during ground operations. An APU, though, is a significant source of green house gases such as $CO_2$ and $NO_x$ and noise. Therefore, hydrogen-gas fed PEM fuel cell systems are studied as a replacement for APUs to significantly reduce these pollutants. Among the different types of fuel cells, PEM fuel cells are the most suitable ones for dynamic applications. Nonetheless, proper fuel cell system operation such as a well humidified membrane or a proper supply of reactant gases and well kept gas pressure is central for a maximum fuel cell lifetime [3]. In this study a fuel cell system with anode hydrogen recirculation for stack humidification is used. Proper fuel and air supply as well as the temperature difference across the stack is managed by an internal fuel cell system controller. The fuel cell system is connected to a controllable ohmic load that draws a current as requested. A schematic of the fuel cell system and its coupled two-loop-cooling system is shown in figure 1.

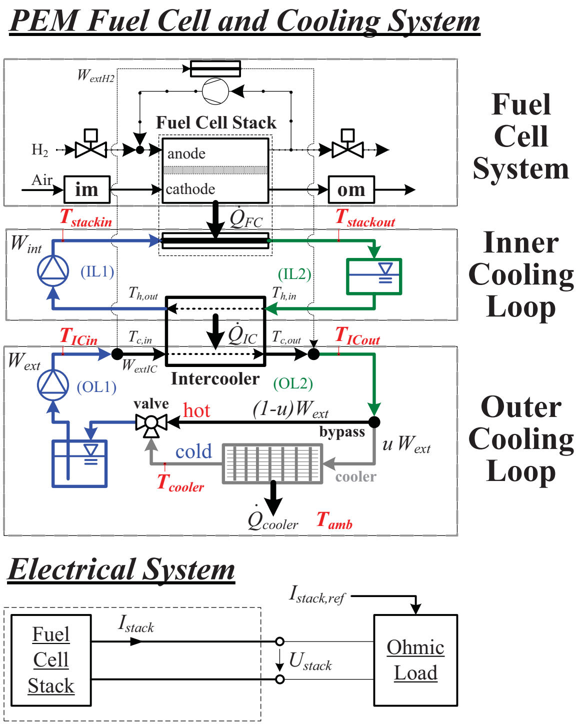


Figure 1. Schematic of the fuel cell system and the connected double-loop cooling system; the fuel cell system comprises the fuel cell stack, inlet and outlet manifolds and anode hydrogen recirculation; the cooling system comprises the inner and outer loop that are interconnected by a counter-flow heat exchanger, connecting pipes, a cooling valve and a cooler; heat flows from stack into the inner loop, inside the intercooler and in the cooler are shown (top); the fuel cell system is connected to an ohmic load (bottom)

The fuel cell system simulation model is based on models reported of in [4], [5] and [6]. Furthermore, product water leaves the fuel cell cathode as vapor and liquid. This has an influence on stack temperature and is taken into account as presented in [7] and [8]. The fuel cell system model is briefly outlined in section Simulation Model. The cooling system is modeled based on first principles. Heat exchanger models as

This study as part of the project "cabin technology and multifunctional fuel cell systems" has been supported by Airbus and the German Federal Ministry of Education and Research (support code: 03CL03A).

proposed in [9] and [6] are restricted on operating range. The effectiveness-NTU method [10] provides nonlinear models for several designs of heat exchangers. Ref. [11] uses this approach for a single-loop cooling system of an automotive fuel cell system assuming constant heat exchanger effectiveness. Constant effectiveness, however, might not hold true for the entire operating range. Here, the effectiveness-NTU method with mass flow dependent effectiveness is used to develop an intercooler model for the coupled double-loop-cooling system. The fuel cell system and intercooler model have been fit to experimental data, which is presented in section Model Identification. Actuating the cooling valve in the outer loop influences the stack inlet temperature in the inner cooling loop. To operate the fuel cell system at a certain stack inlet temperature a temperature controller with feedforward control is proposed in section Controller Design. A simulation of the nonlinear fuel cell and cooling system model with temperature controller are presented in section Simulation results.

## II. SIMULATION MODEL

The overall system simulation model consists of a model of the fuel cell system, a model of its internal fuel cell system controller and a model of the controllable ohmic electrical load to which the fuel cell system is connected. A list of the model parameters is given in table 1 at the end of this paper. Inlet air is taken from an air pressure tank that is filled by a compressor. After compression air is cooled, dried and oil-filtered.

### A. Fuel Cell System Model

The fuel cell system model comprises the fuel cell stack, a mass flow controller (MFC) for cathode air supply and the anode hydrogen recirculation for better hydrogen utilization and stack humidification. The internal fuel cell system controller operates the mass flow controller such that it delivers an air mass flow as specified by stoichiometry and stack current drawn. Furthermore, the internal controller operates the cooling pump in the inner cooling loop to set the reference temperature difference across the fuel cell stack. Inlet air is modeled as a dry and ideal gas consisting of 21% (vol.) oxygen and 79% (vol.) nitrogen. It has constant temperature.

#### 1) Inlet Manifold Model

Inlet manifold pressure $p_{im}$ (1) is gained by a pressure differential equation ($\gamma=1.4$ [5]). In- and outlet temperature are assumed to be equal and constant. Flow $W_{im}$ of dry air exiting the manifold is modeled by a linear nozzle equation (2) with nozzle constant $k_{im}$ and the pressure difference of manifold and cathode pressure $p_{ca}$. Mass flow of dry air $W_{MFC}$ (3) entering the manifold is supplied by a mass flow controller, which is modeled as a first order transfer function with time constant $T_{MFC}$. The mass flow reference is calculated using the reference stoichiometry $\lambda_{ref}$ and the stack current $I_{stack}$ being drawn.

$$\frac{dp_{im}}{dt} = \frac{\gamma R_{air}}{V_{im}} T_{im}\left(W_{MFC} - W_{im}\right) \tag{1}$$

$$W_{im} = k_{im}\left(p_{im} - p_{ca}\right) \tag{2}$$

$$W_{MFC} = \frac{1}{T_{MFC}s+1}\left[I_{stack}\lambda_{ref}\left(M_{O2} + \frac{0.79}{0.21}M_{N2}\right)\frac{n_{cells}}{4F}\right] \tag{3}$$

#### 2) Outlet Manifold Model

Vapor mass flow is obtained by water loading (4) of the dry gas mass flow. Water loading depends on total gas pressure $p_i$ and vapor partial pressure $p_{i,v}$ (5), which solely depends on temperature [12]. Mass flows of water vapor $W_{ca,v}$, oxygen depleted air $W_{ca,oda}$ and liquid water $W_{ca,l}$ are supplied to the outlet manifold at stack temperature $T_{stack}$. The manifold is considered perfectly insulated. Mass flows $W_{om,v}$, $W_{om,oda}$ and $W_{om,l}$ exit directly to the ambient environment and are governed by a linear nozzle equation with constant $k_{om}$ and manifold and ambient pressure difference (6). The flow of liquid water leaving the manifold (7) is modeled to be dependent on the gas mass flow, the liquid water mass $m_{om,l}$ and a constant $\delta_l$ [7].

$$X_i = \frac{p_{v,i}}{p_i - p_{v,i}}\frac{R_{air}}{R_v} \tag{4}$$

$$p_v^{sat} = \exp\left(17.2799 - \frac{4102.99}{\vartheta + 237.431}\right)0.611657\times10^3 \tag{5}$$

$$W_{om} = k_{om}\left(p_{om} - p_{amb}\right) \tag{6}$$

$$W_{om,l} = \delta_l W_{om} m_{om,l} \tag{7}$$

Partial pressures of dry gas $p_{om,oda}$ and vapor $p_{om,v}$ are gained by a mass balance and the ideal gas law (8) with manifold volume $V_{om}$ and the vapor gas constant $R_v$. ODA is approximated as air with gas constant $R_{oda}$=$R_{air}$. Condensation is considered happening instantaneously leading to the condensate mass $m_{om,l}$. The outlet manifold pressure is the sum of oda and vapor partial pressures $p_{om} = p_{om,oda} + p_{om,v}$.

$$\frac{dm_{om,oda}}{dt} = W_{ca,oda} - W_{om,oda}\,, \qquad W_{om,oda} = \frac{1}{1+X_{om}}W_{om}$$

$$\frac{dm_{om,H2O}}{dt} = W_{ca,v} + W_{ca,l} - W_{om,v} - W_{om,l}\,, \quad W_{om,v} = \frac{X_{om}}{1+X_{om}}W_{om}$$

$$p_{om,oda} = \frac{T_{om}R_{oda}}{V_{om}}m_{om,oda} \tag{8}$$

$$p_{om,v} = \min\left(p_v^{sat}(T_{om}), m_{om,H2O}\frac{R_vT_{om}}{V_{om}}\right) \tag{9}$$

$$m_{om,l} = \max\left(0, m_{om,H2O} - p_v^{sat}(T_{om})\frac{V_{om}}{R_vT_{om}}\right) \tag{10}$$

#### 3) Fuel Cell Stack Electrical Model

The stack voltage $U_{stack} = n_{cells}\,U_{cell}$ is the sum of all $n_{cells}$ cell voltages. The cell voltage is modeled as $U_{cell} = U_{rev} - \eta_{act} - \eta_{\Omega}$ (11) with the reversible cell voltage $U_{rev}$, activation loss $\eta_{act}$ [13] and ohmic loss $\eta_{\Omega}$ [14]. The membrane thickness is given by $d_m$ and the active surface area by $A_{sfc}$. Parameters $\zeta_1,\ldots\zeta_4$ and $b_1,\ldots b_3$ have been identified using experimental data.

$$\begin{aligned} U_{rev} &= 1.229 - 0.85\cdot10^{-3}\left(T_{stack} - 298.15\right) + 4.3\cdot10^{-5}T_{stack}\left(\ln\tfrac{p_{H2}}{p_0} + \tfrac{1}{2}\ln\tfrac{p_{O2}}{p_0}\right) \\ \eta_{act} &= \zeta_1 + \zeta_2T_{stack} + \zeta_3T_{stack}\ln\left(p_{O2}e^{\left(498/T_{stack}\right)}/5.08\cdot10^{-6}\right) + \zeta_4T_{stack}\ln\left(I_{stack}\right) \\ \eta_{\Omega} &= \frac{d_m}{\left(b_1\lambda_m - b_2\right)}e^{-b_3\left(\frac{1}{303}-\frac{1}{T_{stack}}\right)}\frac{I_{stack}}{A_{sfc}} \end{aligned} \tag{11}$$

#### 4) Fuel Cell Stack Thermal Model

An energy balance (12) leads to the fuel cell stack temperature $T_{stack}$. The stack has a heat capacity of $C_{st}$. The heat generated through the chemical reaction is given by the chemical energy of hydrogen (HHV) and stack electrical energy. The stack is cooled by convective cooling with coolant mass flow $W_{cool} = W_{int}$ and specific heat capacity $c_{cool}$. The coolant stack inflow temperature is $T_{stackin}$. The coolant leaves the stack at stack temperature. Air, cathode water, vapor and oda gas are assumed to leave the stack at stack temperature as well. Air enters the stack at temperature $T_{im}$. Oda gas is approximated as air with specific heat capacity $c_{oda}= c_{air}$. The stack is very well insulated. So, heat transfer to surroundings is neglected. Enthalpy of evaporation is $h_0$.

$$\begin{aligned} C_{st}\frac{dT_{st}}{dt} &= \left(1.48\,n_{cells} - U_{stack}\right)I_{stack} \\ &+ W_{cool}c_{cool}\left(T_{stackin} - T_{stack}\right) \\ &+ W_{im}c_{air}\left(T_{im} - T_0\right) - W_{oda}c_{oda}\left(T_{stack} - T_0\right) \\ &- W_{ca,l}c_l\left(T_{stack} - T_0\right) - W_{ca,v}\left(h_0 + c_v\left(T_{stack} - T_0\right)\right) \end{aligned} \tag{12}$$

### 5) Fuel Cell Stack Anode and Cathode Model

Anode pressure $p_{an} = p_{H2} + p_{an,v}$ is the sum of hydrogen and vapor partial pressures and is governed by mass conservation (13) and the ideal gas law with $R_{H2}$ being the hydrogen gas constant. In operation a mass flow $W_{H2rct} = (M_{H2}\, I_{stack}\, n_{cells})/(2F)$ of hydrogen is consumed. The stack is operated dead-ended and the system behavior is mimicked by a proportional controller for anode pressure $p_{an}$ with hydrogen inlet mass flow $W_{H2in}$, anode reference pressure $p_{an\_ref}$ and controller gain $k_{an}$ (14). Condensation is modeled happening instantaneously [5] if the vapor pressure exceeds saturation vapor pressure. The mass of condensate and the vapor partial pressure are modeled according to (9) and (10) with water mass (13). The anode water activity $a_{an}$ is modeled as $a_{an}= p_{an,v}/p_v^{sat}$.

$$\frac{dm_{an,H2O}}{dt} = -W_{mem} \text{ and } \frac{dp_{H2}}{dt} = \frac{T_{stack}R_{H2}}{V_{an}}\left(W_{H2in} - W_{H2rct}\right) \tag{13}$$

$$W_{H2in} = k_{an}\left(p_{an\_ref} - p_{an}\right) \tag{14}$$

The cathode inlet mass flows of oxygen and nitrogen are governed by the mass fractions $x_{O2}$ and $x_{N2}$ assuming dry inlet air consisting of 21% of oxygen and 79% of nitrogen. The cathode exit mass flow $W_{ca}$ is modeled as a 3-phase-flow (15) and is governed by a linear nozzle equation similar to (2) with constant $k_{ca}$ and pressure difference ($p_{ca}$ - $p_{om}$). Liquid water mass flow $W_{ca,l}$ is modeled according to (7). Cathode pressure $p_{ca} = p_{O2} + p_{N2} + p_{ca,v}$ is the sum of oxygen, nitrogen and vapor partial pressures and is governed by mass conservation (16) and the ideal gas law [5]. In operation a mass flow $W_{O2rct} = (M_{O2}\, I_{stack}\, n_{cells})/(4F)$ of oxygen is consumed and $W_{H2Orct} = (M_{H2O}\, I_{stack}\, n_{cells})/(2F)$ of water is generated. Condensation happens instantaneously. The vapor partial pressure and liquid mass in the cathode is modeled according to (9) and (10) using the water mass $m_{ca,H2O}$ in (16).

$$\begin{bmatrix} W_{im,O2} \\ W_{im,N2} \end{bmatrix} = \begin{bmatrix} x_{O2} \\ x_{N2} \end{bmatrix} W_{im} \text{ and } \begin{bmatrix} W_{ca,O2} \\ W_{ca,N2} \\ W_{ca,v} \end{bmatrix} = \frac{1}{1+X_{CA}} \begin{bmatrix} \frac{m_{O2}}{m_{O2}+m_{N2}} \\ \frac{m_{N2}}{m_{O2}+m_{N2}} \\ X_{CA} \end{bmatrix} W_{ca} \tag{15}$$

$$\begin{aligned} \frac{dm_{ca,H2O}}{dt} &= W_{mem} + W_{H2Orct} - W_{ca,v} - W_{ca,l} \\ \frac{dm_{O2}}{dt} &= W_{im,O2} - W_{O2rct} - W_{ca,O2} \\ \frac{dm_{N2}}{dt} &= W_{im,N2} - W_{ca,N2} \end{aligned} \quad , \quad \begin{aligned} p_{O2} &= m_{O2}\frac{T_{stack}R_{O2}}{V_{ca}} \\ p_{N2} &= m_{N2}\frac{T_{stack}R_{N2}}{V_{ca}} \end{aligned} \tag{16}$$

### 6) Fuel Cell Stack Membrane Model

Membrane material is Nafion®. The membrane water mass flow $W_{mem}$ (17) is caused by gradient driven diffusion and electro-osmotic drag of water from anode to cathode [14]. Diffusion constant $D_{diff}$ depends on the membrane hydration $\lambda_{mem}$ and stack temperature (18). The membrane water activity $a_{mem}$ is modeled as the average between cathode and anode water activity [14] (with j=*mem, ca, an*).

$$W_{mem} = A_{sfc}M_w n_{cells}\left(\lambda_{mem}\frac{2.5}{22}\frac{I_{stack}}{A_{sfc}F} - \frac{\rho_{dry}}{M_{dry}}D_{diff}\frac{\lambda_{ca}-\lambda_{an}}{d_m}\right) \tag{17}$$

$$D_{diff} = e^{2416\left(1/303-1/T_{st}\right)}\left(2.563-0.33\lambda_{mem}+0.0264\lambda_{mem}^2-0.000671\lambda_{mem}^3\right)\cdot 10^{-6}$$

$$\lambda_j = \begin{cases} 0.043 + 17.81a_j - 39.85a_j^2 + 36.0a_j^3 & 0 < a \le 1 \\ 14 + 1.4\left(a_j - 1\right) & 1 < a \le 3 \end{cases} \tag{18}$$

### 7) Electrical Load Model

The stack is connected to a controllable ohmic load drawing a current as requested by a reference. An internal controller quickly adjusts the load resistance. To prevent numerical problems, the load is modeled as a first order transfer function with time constant $T_L$.

## B. Cooling System Model

As shown in figure 1 the cooling system consists of an inner and an outer cooling loop that are interconnected by an counter-flow heat exchanger also called intercooler. In the inner loop pipes connect the intercooler and the stack.

### 1) Inner Cooling Loop Model

The cooling pump in the internal cooling circuit is controlled such that the cooling temperature difference across the stack equals the reference signal given by the stack manufacturer. The pump generates a coolant mass flow $W_{int}$, which is limited to a minimum of $W_{int,min}$ and a maximum of $W_{int,max}$. The cooling pump controller is modeled as a proportional and integral controller with anti-windup to prevent integrator windup. The pipes are modeled according to [13] and are considered perfectly insulated (20) with specific heat capacity $c_{int}$ and $m_{pipe,i}$ being the mass of coolant in pipe *i*, the coolant inlet temperature $T_{pipein,i}$ and the pipe outlet temperature $T_{pipe,i}$. Using notation of figure 1 the temperature of pipe IL1 is $T_{stackin}$ and of pipe IL2 is $T_{stackout}$. The indices for the model equations (20) are *i*=IL1, IL2.

$$m_{pipe,i}c_{int}\frac{dT_{pipe,i}}{dt} = W_{int}c_{int}\left(T_{pipein,i} - T_{pipe,i}\right) \tag{20}$$

### 2) Outer Cooling Loop Model

The outer loop consists of an air-cooled heat exchanger (cooler) that cools the coolant to a temperature of $T_{cooler}$ and pipes that connect the cooler with the intercooler. In a cooling valve warm coolant from the intercooler and cold cooling fluid from the cooler can be mixed. Warm coolant bypasses the cooler such that is does not cool down. As the bypass pipe is very short as compared to the other pipes its dynamics are neglected. The cooler is modeled as a pipe as well (21). The heat transferred to the surroundings is modeled by the coefficient $k_{cooler}$ and the difference of cooler to ambient temperature $T_{amb}$. The mass flow in the external cooling loop is $W_{ext}$ and the specific heat capacity of the coolant is $c_{ext}$. Using notation of figure 1 the temperature of pipe OL1 is $T_{ICin}$ and of pipe OL2 is $T_{ICout}$. The indices for the model equations (21) are $i$=OL1, OL2.

$$m_{pipe,i}c_{ext}\frac{dT_{pipe,i}}{dt}=W_{ext}c_{ext}\left(T_{pipein,i}-T_{pipe,i}\right)$$
$$m_{cooler}c_{ext}\frac{dT_{cooler}}{dt}=u_{valve}W_{ext}c_{ext}\left(T_{ICout}-T_{cooler}\right)-k_{cooler}\left(T_{cooler}-T_{amb}\right) \qquad (21)$$

In pipe OL2 the coolant for the hydrogen recirculation pump mixes with the coolant of the intercooler outflow. This temperature mixture is modeled by a stationary energy balance (22) with thermal input of the hydrogen pump and assuming that the mass flow through the recirculation pump $W_{extH2}$ is 6% of $W_{ext}$. $T_{c,out}$ is the intercooler outlet temperature. The mass flow through the intercooler $W_{extIC}$ is taken as 94% of the mass flow in the external cooling loop.

$$W_{ext}T_{ICout}=W_{extH2}\left(T_{ICin}+\frac{\dot{Q}_{H2pump}}{W_{extH2}c_{ext}}\right)+W_{extIC}T_{c,out} \qquad (22)$$

### 3) Cooling Valve Model

The mixing temperature of the cooling valve is modeled by a stationary energy equation (23) as the mixing process is considered very fast. The valve splits the mass flow through cooler and bypass given by its actual position $u_{valve}$.

$$W_{ext}T_{pipein,OL1}=u_{valve}W_{ext}T_{cooler}+\left(1-u_{valve}\right)W_{ext}T_{ICout} \qquad (23)$$

The cooling valve has limited dynamics. This results in a limited speed of opening and closing the valve. For small changes in position it behaves linearly. It is modeled having a constant opening and closing speed $du_{max}$ for large changes and behaves like a first order transfer function for small changes in valve position (24). Reference position is $u_{valve,ref}$.

$$\frac{du_{valve}}{dt}=\int k_{valve}\left(u_{valve,ref}-u_{valve}\right)\begin{array}{cc} du_{max} & k_{valve}\left(u_{valve,ref}-u_{valve}\right)>du_{max} \\ & otherwise \\ -du_{max} & k_{valve}\left(u_{valve,ref}-u_{valve}\right)<-du_{max}\end{array} \qquad (24)$$

### 4) Intercooler Model

The counter-flow heat exchanger that connects the inner and outer cooling loop is considered static. This assumption is motivated by the high mass flows in the cooling system and the small coolant mass inside the intercooler as compared to the entire cooling system. For modeling counter-flow heat exchangers the effectiveness NTU method [10] has shown very good results. NTU is the number of transfer units, which is an important parameter in heat exchanger design and modeling. The effectiveness NTU is advantageous as outlet temperatures are calculated explicitly on inlet temperatures and cooling mass flows. Heat flow is gained by the inlet temperature differences, the minimum heat capacity flow and the effectiveness $\varepsilon$. The outlet temperatures (25) are gained by the heat capacity flow of the hot side $C_h$=$W_{int}c_{int}$ and the cold side $C_c$=$W_{extIC}c_{ext}$. The outlet temperatures are $T_{c,out}$ and $T_{h,out}$ for cold and hot side. The inlet temperatures are $T_{c,in}$=$T_{ICin}$ and $T_{h,in}$=$T_{stackout}$.

$$T_{c,out}=T_{c,in}+\frac{\varepsilon C_{min}}{C_c}\left(T_{h,in}-T_{c,in}\right)$$
$$T_{h,out}=T_{h,in}-\frac{\varepsilon C_{min}}{C_h}\left(T_{h,in}-T_{c,in}\right) \qquad (25)$$

$$C_{min}=\min\left(C_h,C_c\right),\ C_{max}=\max\left(C_h,C_c\right),\ C^*=\frac{C_{min}}{C_{max}},\ NTU=\frac{UA}{C_{min}}$$

The heat exchanger effectiveness $\varepsilon$ is obtained by (26). $UA$ is the parameter describing the heat transfer.

$$\varepsilon=\begin{cases}\dfrac{NTU}{1+NTU} & C^*=1 \\ \dfrac{1-\exp\left(-NTU\left(1-C^*\right)\right)}{1-C^*\exp\left(-NTU\left(1-C^*\right)\right)} & otherwise\end{cases} \qquad (26)$$

## III. Model Identification

As shown in the previous paragraphs, the fuel cell system model as well as the cooling system model has parameters describing the plant geometry such as manifold volumes and coolant masses and parameters for mass flows, polarization curve and the intercooler. Parameters $\zeta_1,\ldots\zeta_4$ and $b_1,\ldots b_3$ for the polarization curve, $k_{im}$, $k_{ca}$, $k_{om}$ and $k_{an}$ for the mass flows have been identified by a least square error minimization such that simulated inlet and outlet manifold as well as cathode and anode pressure and simulated stack voltage fit to experimental data. Cathode pressure cannot be measured, therefore, it is assumed as the arithmetic mean of the inlet and outlet manifold pressures. The objective function (27) has scaling factors for pressure $\mu_p$ and voltage $\mu_U$ such that quadratic errors of the same order are gained.

$$\begin{aligned}J_{FCS,i}=&\mu_p\left(p_{im,\exp,i}-p_{im,sim,i}\right)^2+\mu_p\left(p_{ca,\exp,i}-p_{ca,sim,i}\right)^2\\&+\mu_p\left(p_{om,\exp,i}-p_{om,sim,i}\right)^2+\mu_p\left(p_{an,\exp,i}-p_{an,sim,i}\right)^2\\&+\mu_U\left(U_{stack,\exp,i}-U_{stack,sim,i}\right)^2\\J_{FCS}=&\sum_i J_{FCS,i}\end{aligned} \qquad (27)$$

Simulation results of the stationary fuel cell system model after identification of the polarization curve for different sets of stoichiometry and stack inlet cooling temperature are shown in figure 2. The results for inlet as well as outlet manifold and anode pressure are shown in figure 3. As the comparison between experimental data and simulation model shows, the model fits the experimental data very well.

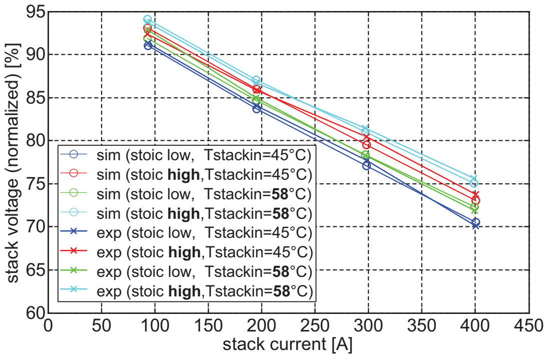


Figure 2. Comparison of experimental and simulation model fuel cell stack polarization curves for different sets of stack inlet temperature (45 and 58°C) and cathode stoichiometry (high and low stoic)

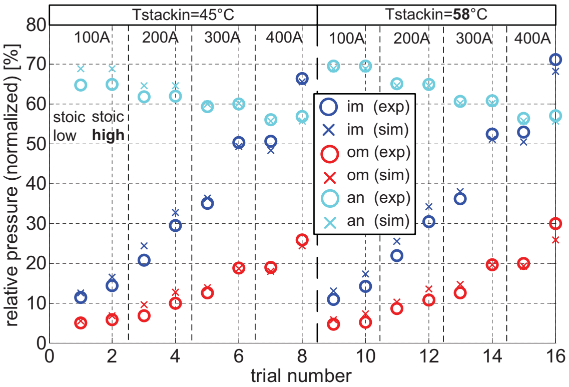


Figure 3. Comparison of experimental and simulation inlet manifold (im), outlet manifold (om) and anode (an) pressures; 16 sets of combinations of stack current, stack inlet temperature and cathode stoichiometry (stoic)

To capture the slow temperature dynamics in the simulation model as well, the stack heat capacity was determined by comparing stack temperature of simulation and experiment until results agreed well. For determining the stack heat capacity cooling mass flow and inlet temperature as well as stack current were prescribed. Simulation results with a PI-controlled cooling pump and stack current, stoichiometry and the coolant inlet temperature prescribed are shown in figure 4 and 5. The dynamic simulation results show that deviations between model and experiment in terms of pressure as well as stack voltage are small. The dynamic simulation model fits the experimental data very well.

### 1) *Identification of the Intercooler Model*

Parameter *UA* of the intercooler model has been identified by a least square error minimization as well. For model identification it was assumed that temperature measurement errors compensate as the temperature difference across the heat exchanger is taken. The cooling mass flow $W_{int,measure}$ of the internal cooling loop can be considered an exact measurement. The cooling mass flow in the external cooling loop $W_{ext,measure}$, however, is corrected by a constant factor $\alpha$ to minimize the difference between the stationary energy balances of hot and cold side (28). The stationary balance is determined by cooling mass flow, specific heat capacity of the coolant and the temperature difference.

$$J_Q = \sum_i \dot{Q}_{h,i} - \dot{Q}_{c,i} \\ = \sum_i W_{\text{int},measure,i} c_{\text{int}} \Delta T_{\text{int},measure,i} - \alpha W_{ext,measure,i} c_{ext} \Delta T_{ext,measure,i} \tag{28}$$

Heat flows resulting from an energy balance about inner and outer cooling loop are shown in figure 6. Using $W_{int,measure}$, the corrected mass flow $\alpha W_{ext,measure}$ and temperatures of hot and cold fluid at in- and outlet the intercooler parameter *UA* has been identified by a least square error minimization with objective function (29). Coolant mass flows and temperatures at the heat exchanger inflow at inner and outer loop were prescribed. *UA* has been fit to minimize the difference between simulated and measured coolant temperatures at the intercooler outflows.

$$J_T = \sum_i \left(T_{hout,\exp,i} - T_{hout,sim,i}\right)^2 + \left(T_{cout,\exp,i} - T_{cout,sim,i}\right)^2 \tag{29}$$

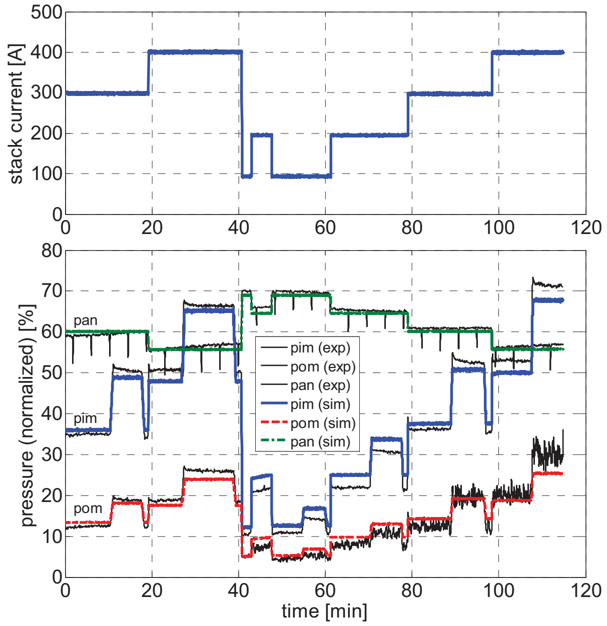


Figure 4. Simulation results dynamic fuel cell system model: experimental stack current profile also used as current profile for the simulation model (top); in- and outlet manifold and anode pressure of simulation and experiment (bottom)

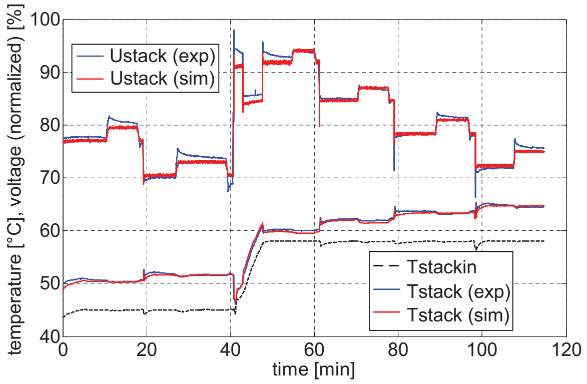


Figure 5. Simulation results fuel cell system model: stack cooling and stack temperature and stack voltage of simulation and experiment

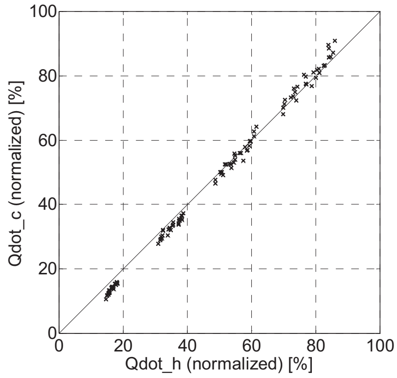


Figure 6. Stationary heat flows leaving the inner cooling loop (Qdot_h) and being brought into the outer cooling loop (Qdot_c) at the intercooler with corrected cooling mass flow of the outer cooling loop

A comparison of simulation results and experimental data for different sets of stack current, outer loop cooling temperature and cooling mass flow are shown in figure 5. An increase of stack current leads to an increase of coolant temperature in the inner loop as the thermal load is increased which can be seen by the polarization curve in figure 2.

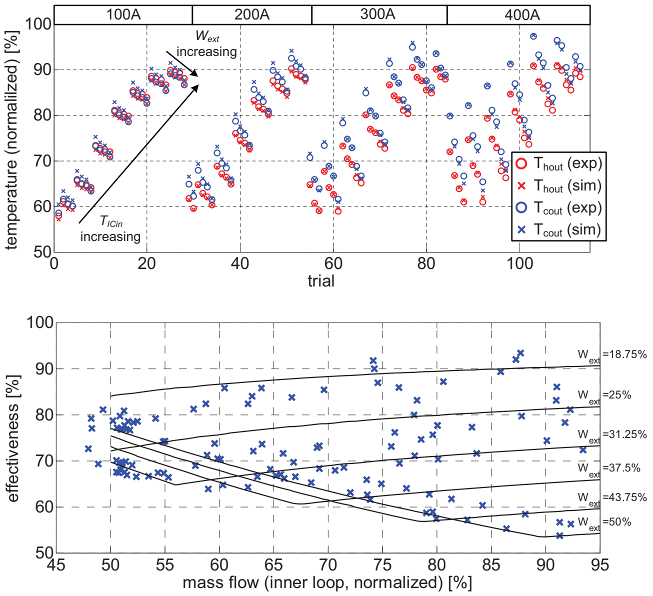


Figure 7. Comparison of experimental and simulation model outlet temperature of inner ($T_{hout}$) and outer ($T_{cout}$) cooling loop; for different sets of stack current, outer loop cooling temperature $T_{ICin}$ and outer loop cooling mass flow $W_{ext}$; cathode stoichiometry was kept constant (top); effectiveness of the intercooler model for all experimental trials and shown for various cooling mass flows in the inner and outer loop (bottom)

As shown in figure 7 the intercooler simulation results of the stationary model fit the experimental data very well. Furthermore, the heat exchanger effectiveness changes significantly within an interval of 50-95%. Figure 7 also shows the evolution of the effectiveness for constant mass flows $W_{ext}$ in the outer loop while the flow in the inner loop $W_{int}$ is varied. For constant $W_{ext}$ the effectiveness could be approximated by a constant, however still can change significantly depending on the choice of $W_{ext}$. As future controller designs possibly work with various pump speeds, a simulation model that is correct over the whole operating range of the fuel cell system is necessary. Therefore, the effectiveness is modeled as in [10] being dependent on cooling mass flows. For the dynamic simulation model parameters such as manifold, cathode and anode volumes and pipe coolant masses were determined by the system geometry and data provided by the manufacturer.

## IV. Cooling Controller Architecture

For the design of a stack inlet cooling temperature control the fuel cell system is considered a stack current dependent heat source generating a heat flow into the inner cooling loop as shown in figure 1. Evaluating the cooling system model at steady state and neglecting the hydrogen compressor's influence, the steady state valve position $u_{ss}$ to reach a certain stack inlet temperature $T_{stackin,ref}$ can be obtained by (30) with the cooler temperate $T_{cooler}$.

$$u_{ss} = \frac{\dot{Q}_{FC}}{C_c\left(\frac{\dot{Q}_{FC}}{C_h} - \frac{\dot{Q}_{FC}}{\varepsilon C_{\min}} + \frac{\dot{Q}_{FC}}{C_c} + T_{stackin,ref} - T_{cooler}\right)} \tag{30}$$

An estimation of the thermal load is obtained by (31) with the actual electrical power $P_{stack} = U_{stack} I_{stack}$ and a cell voltage corresponding to the lower heating value (LHV) of hydrogen as the stack is partly cooled by evaporation of product water in the cathode. The cooler temperature $T_{cooler}$ is being measured. The heat exchanger effectiveness $\varepsilon$ and heat capacity flows $C_h$, $C_c$ and $C_{min}$ are obtained by (25)-(26) using the measured cooling mass flows $W_{int}$ and $W_{extIC}$.

$$\dot{Q}_{FC} = U_{LHV} n_{cells} I_{stack} - P_{stack} \tag{31}$$

The control law proposed is a combination of the nonlinear feedforward control and a linear proportional and integral controller with an anti-windup to prevent integrator windup (32). The anti-windup circuit is active for controller outputs greater than 100% ($a_{windup}$=100%-$u$) and outputs less than 0% ($a_{windup}$ = 0%-$u$) and is inactive otherwise ($a_{windup}$=0). The controller gains $k_p>0$ and $k_i>0$ add with a negative sign as shown in (32) as the cooling valve needs to open for negative control errors ($T_{stackin} > T_{stackin,ref}$) and to close for positive ones.

$$u = u_{ss} - k_p\left(T_{stackin}^{ref} - T_{stackin}\right) - k_i \int T_{stackin,ref} - T_{stackin} - a_{windup} dt \tag{32}$$

A constant cooling mass flow $W_{ext}$ in the outer loop is set for the controller. Increasing the cooling mass flow would improve the cooling system dynamics as states in (21). Higher mass flows reduce the time constants for temperature in the pipe volumes. Higher pump speeds, however, result in higher power consumption. Nevertheless, the stack temperature can be set by either actuating the cooling valve or actuating the pump within certain limits set by the intercooler. This circumstance can be exploited for future control designs such that an optimal balance between pump and valve is found to optimize for power consumption or time. Here, the cooling mass flow $W_{ext}$ is set constant.

## V. Simulation Results

The simulation model is developed and run in Matlab/Simulink®. The controller for stack inlet cooling temperature is connected with the nonlinear fuel cell system model. The simulation was run for a stack cooling reference temperature of 55°C and high cathode stoichiometry. Figure 8 shows the simulation results of three controller variants:

(A) PI-control activated, feedforward control deactivated

(B) PI-control deactivated and feedforward control activated

(C) PI- and feedforward control activated.

For a constant stack cooling reference temperature of 55°C a stack current profile as shown in figure 8 was taken to test the controller for different thermal loads. With a pure feedforward control (type B) stack temperature in steady state is close to the reference temperature. The control behavior is further improved by a linear PI-controller (type C). The temperature matches the reference in steady state and under- as well as overshoots are less than in the case of a linear controller without feedforward control (type A).

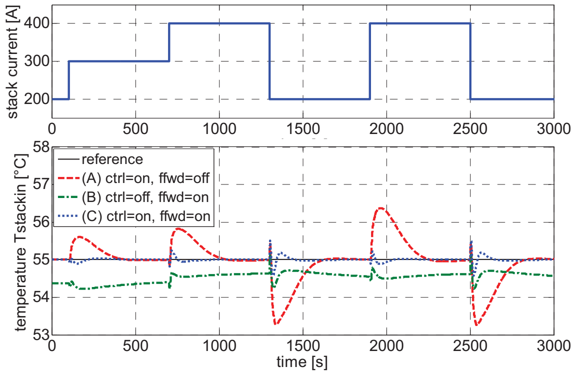


Figure 8. Simulation results of stack cooling controller for 3 controller configurations; stack current profile (top), stack cooling temperature (bottom)

## VI. Conclusion

A polymer electrolyte membrane fuel cell system electrically connected to an ohmic electronic load is connected to a coupled double-loop-cooling system with two different coolants of different specific heat capacity. The cooling loops are interconnected by a counter-flow heat exchanger (intercooler). The entire system model is derived based on physical principals and has been modeled in Matlab/Simulink®. The fuel cell system model has been fit to experimental data with very good agreement. The intercooler is modeled by the effectiveness NTU method by which a static nonlinear model is obtained. This model is valid over the entire operating range of the fuel cell system. Outlet temperatures are calculated explicitly based on input variables such as temperatures and mass flows. The intercooler model has been fit to experimental data and shows very good agreement over the entire operating range. A linear PI-controller with nonlinear feedforward control is proposed to control for stack inlet cooling temperature. The temperature controller is implemented in the nonlinear simulation model. The fuel cell and cooling system model will by used for future controller designs such as for an optimal change of stack temperature operating points or for an optimal system heat up.

## References

[1] E. Vredenborg, H. Lüdders and F. Thielecke, “Methodology for Sizing and Simulation of complex Fuel Cell Systems” orig.(german) ”Methodik zur Auslegung und Simulation komplexer Brennstoffzellensysteme”, Deutscher Luft- und Raumfahrtkongress 2010, DocumentID: 161248.

[2] J. Bleil, “Fuel Cells for onboard Power Supply of Aircraft”, orig. (german) “Brennstoffzellen zur Bordstromversorgung von Flugzeugen”, HZwei-Das Magazin für Wasserstoff und Brennstoffzellen, 04/2007

[3] R. Borup et al., “Scientific Aspects of Polymer Electrolyte Fuel Cell Durability and Degradation”, Chem. Rev, vol. 107, pp. 3904-3951, 2007

[4] A. J. del Real, A. Arce, and C. Bordons, “Development and experimental validation of a PEM fuel cell dynamic model”, Journal of Power Sources, vol. 173, pp. 310-324, 2007

[5] J. T. Pukrushpan, A. G. Stefanopoulou and H. Peng, Control of Fuel Cell Power Systems. London: Springer-Verlag, 2004.

[6] A. Y. Karnik, J. Sun, A. G. Stefanopoulou and J. H. Buckland, “Humidity and Pressure Regulation in a PEM Fuel Cell Using a Gain-Scheduled Static Feedback Controller“, IEEE Transactions on Control Systems Technology, vol. 17, No. 2, pp. 283-297, 2009

[7] M. Schultze and J. Horn, “A Control Oriented Simulation Model of an Evaporation Cooled Polymer Electrolyte Membrane Fuel Cell System”, 18th IFAC World Congress, Milan, Italy, 2011

[8] M. Schultze and J. Horn, “Current Control of a PEMFC System connected to an Electrical Load through a DC/DC Converter”, 19th Mediterranean Conference on Control & Automation, pp. 55-60, 2011

[9] J. Niemeyer, Model Predicitve Control of a PEM-Fuel Cell System, orig. (german) Modellprädiktive Regelung eines PEM-Brennstoffzellensystems, Schriften des Instituts für Regelungs- und Steuerungssysteme, Universität Karlsruhe, Band 05, 2009

[10] R. K. Shah and D. P. Sekulic, Fundamentals of Heat Exchanger Design, Hoboken, NJ: John Wiley & Sons, 2003

[11] J. Nolan and J. Kolodziej, “Modeling of an automotive fuel cell thermal system”, Journal of Power Sources, vol. 195, pp. 4743-4752, 2010

[12] H. D. Baehr and S. Kabelac, Thermodynamics, orig. (german) Thermodynamik, Berlin: Springer-Verlag, 2006

[13] J. C. Amphlett, R. M. Baumert, R. F. Mann, B. A. Peppley and P. R. Roberge, “Performance Modeling of the Ballard Mark IV Solid Polymer Electrolyte Fuel Cell”, J. Electrochem. Soc., vol. 142, pp. 1-8, 1995

[14] R. O’Hayre, S.-W. Cha, W. Colella and F. B. Prinz, Fuel Cell Fundamentals. Hoboken, NJ: John Wiley & Sons, 2009.

TABLE I. Fuel Cell System Model Parameters

| Parameter | Description | Value |
|---|---|---|
| $M_{H2O}$ | molar mass of water | 18.0153 g/mol |
| $M_{O2}$ | molar mass of oxygen | 31.9988 g/mol |
| $M_{N2}$ | molar mass of nitrogen | 28.0134 g/mol |
| $M_{H2}$ | molar mass of hydrogen | 2.01588 g/mol |
| $R_v$ | gas constant of vapor | 461.5 J/kg/K |
| $R_{O2}$ | gas constant of oxygen | 259.8 J/kg/K |
| $R_{N2}$ | gas constant of nitrogen | 296.8 J/kg/K |
| $R_{H2}$ | gas constant of hydrogen | 4124.49 J/kg/K |
| $R_{air}$, $R_{oda}$ | gas constant of air, ODA | 287.058 J/kg/K |
| $F$ | Faraday’s constant | 96485.3 C/mol |
| $c_{air}$, $c_{oda}$ | specific heat capacity of air, ODA | 1004.7 J/kg/K |
| $c_l$ | specific heat capacity of water | 4181.9 J/kg/K |
| $c_v$ | specific heat capacity of vapor | 1864.6 J/kg/K |
| $h_o$ | enthalpy of evaporation | 2500.9×10$^3$ J/kg |
| $n_{cells}$, $A_{sfc}$ | number of cells in the stack, active surface area | |
| $d_m$ | membrane thickness | |
| $V_{im}$, $V_{om}$, $V_{ca}$, $V_{an}$ | volume of inlet and outlet manifold, stack cathode and anode | |
| $C_{st}$ | fuel cell stack heat capacity | |
| $c_{int}$, $c_{ext}$ | coolant specific heat capacity in the inner and outer cooling loop | |
| $W_i$ | mass flow leaving cathode and anode (*i=ca, an*), in- and outlet manifold (*i=im,om*), mass flow controller (*i=MFC*) | |
| $W_{i,oda}$, $W_{i,l}$, $W_{i,v}$ | mass flow of ODA, liquid water and vapor in cathode and outlet manifold (*i=ca, om)* | |
| $W_{int}$, $W_{ext}$ | coolant mass flow inner and outer cooling loop | |
| $m_{pipe,i}$, $m_{cooler}$ | mass of coolant in pipe *i* (*i=IL1, IL2, OL1, OL2*), mass of coolant in the cooler | |
| $p_i$ | pressure in volume *i* (*i=im, ca, an, om*) | |
| $m_{i,H2O}$ | mass of total water in volume *i* (*i=ca, an, om*) | |
| $m_{i,l}$ | condensate mass in *i* (*i=ca, an, om*) | |
| $m_{O2}$, $m_{N2}$ | mass of oxygen, nitrogen in cathode | |
| $m_{pipe,i}$,$m_{cooler}$ | coolant mass in *i* (*i=IL1, IL2, OL1, OL2)* and cooler | |
| $p_{O2}$, $p_{N2}$, $p_{H2}$, $p_{i,v}$ | partial pressure of oxygen, nitrogen, hydrogen and vapor in cathode, anode and outlet manifold (*i=ca, an, om*) | |
| $p_{amb}$, $T_{amb}$ | ambient pressure and temperature | |
| $T_i$ | temperature in volume *i* (*i=im, stack, om*) | |
| $T_{pipe,i}$, $T_{cooler}$ | temperature in pipe *i* (*i=IL1, IL2, OL1, OL2)* and cooler | |
| $T_{MFC}$, $T_L$ | time constant of mass flow controller, electrical load | |
| $k_{cooler}$ | heat transfer coefficient of the cooler | |
| $k_{im}$, $k_{om}$, $k_{ca}$, $k_{an}$ | mass flow coefficients for inlet- and outlet manifold, cathode and anode | |
| $\zeta_1,\ldots\zeta_4$ | coefficients for stack voltage activation loss | |
| $b_1,\ldots b_3$ | coefficients for stack voltage ohmic loss | |

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# Counter-examples Design to Global Convergence of Maximum Likelihood Estimators

Yiqun Zou, Xiafei Tang and Zhengtao Ding

***Abstract*— MLE(Maximum Likelihood Estimation) is widely applied in system identification because of its consistency, asymptotic efficiency and sufficiency. However gradient-based optimization of the likelihood function might end up in local convergence. To overcome this difficulty, the non-local-minimum conditions are very useful. Here we suggest a heuristic method of constructing local minimum examples for ARMAX, ARARMAX and BJ models. Based on them the derivation of non-local-minimum conditions can be inspired by analyzing these examples.**

## I. INTRODUCTION

Many methods have been presented in the area of system identification, such as **MLE**[1][9], frequency domain analysis[11] and subspace method[12]. Amongst them, maximum likelihood estimation is one of the most popular approaches.

The idea of **MLE** introduced by [7] and further proven by [14] is to obtain the maximum likelihood estimate $\hat{\theta}_{ML}$ through maximizing the likelihood function or minimizing its corresponding natural negative logarithmic form. An efficient method is gradient descent search [9] which is applied extensively in optimization. However the if the landscape of the objective function has at least one local minimum, the gradient search may get stuck in local convergence when badly initialized. In this case, the **MLE** will produce wrong system information. Hence so-called non-local-minimum conditions[15] have been developed to judge whether there exists any local minimum.

**N.B.** To clarify the difference between the global and local minimum, here we refer the local minimum to the "false" non-global minimum as in [9].

An innovative method to derive non-local-minimum conditions can be described as follows. First of all, we design so-called "local minimum examples", i.e. the particular model structures with local minima. Secondly, via tuning the model dynamics or the input signals etc of such examples in simulation, the condition which affects the local minimum existence can be tested. At last we analyze such condition theoretically in order to derive the corresponding non-local-minimum condition. In this paper we only look at the first step and make some suggestions on the design of local minimum examples.

The structure of this paper is organized as follows. Next section explains the background of **MLE**. In section 3 the general methodology of the construction of local minimum examples is provided. Section 4 shows details of local minimum examples construction for open loop **OE**, **ARMAX**, **ARARMAX** and **BJ** models respectively. In section 5, simulation examples for each model above are given. Section 6 summarizes our contributions and points out the future works.

## II. BACKGROUND OF MAXIMUM LIKELIHOOD ESTIMATION

To illustrate the concept of **MLE**, Ljung [9] lets the observations represented by the random variable $y^N = (y(1),\, y(2),\, \ldots,\, y(t),\, \ldots,\, y(N))$ which takes values in $R^N$ and the **PDF** (**P**robability **D**ensity **F**unction) of $y^N$ by $f_y(\hat{\theta}, y^N)$. If the observed value of $y^N$ is $y_*^N$, the probability that the observation should take the value $y_*^N$ is proportional to

$$f_y(\hat{\theta}, y_*^N) \tag{1}$$

This is a deterministic function of $\hat{\theta}$ which is known as *the likelihood function*. A reasonable estimator $\hat{\theta}$ or explicitly $\hat{\theta}_{ML}$ can be chosen so that the observed event becomes "as likely as possible". That is

$$\hat{\theta}_{ML}(y_*^N) = \arg\max_{\hat{\theta}} f_y(\hat{\theta}, y_*^N) \tag{2}$$

where the maximization is performed for fixed $y_*^N$. This function is known as *maximum likelihood estimator*.

Such an estimator is reasonable because of its three advantages: firstly it provides a consistent estimate asymptotically, i.e.

$$\hat{\theta}_{ML} \to \theta \quad \text{with probability } 1 \tag{3}$$

for different model structures, e.g. [2], [4]. This property is known as consistency [14]. Secondly the covariance of $\hat{\theta}_{ML}$ is lower bounded by the inverse of Fisher information matrix, i.e.

$$E[\hat{\theta}_{ML} - \theta][\hat{\theta}_{ML} - \theta]^T \geq (-E\left[\frac{d^2}{d\hat{\theta}^2} \log f_y(\hat{\theta}, y^N)|_{\hat{\theta}=\theta}\right])^{-1} \tag{4}$$

Y. Zou is with Department of Intelligence Science and Technology, School of Information Science and Engineering, Central South University, 410083, China Email: yiqunzou@gmail.com

X. Tang is with Control Systems Centre, School of Electrical and Electronic Engineering, University of Manchester, M13 9PL, UK Email: xiafei.tang@postgrad.manchester.ac.uk

Z. Ding is with Control Systems Centre, School of Electrical and Electronic Engineering, University of Manchester, M13 9PL, UK Email: zhengtao.ding@postgrad.manchester.ac.uk

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

where the equality holds asymptotically. This property is known as asymptotic efficiency [5]. Here $E$ represents the mathematical expectation. Thirdly assume $S(y^N)$ is a sufficient statistic. According to the **Factorisation Theorem**, (1) can be rewritten to

$$f_y(\hat{\theta}, y_*^N) = \Psi(S(y_*^N), \hat{\theta})h(y_*^N) \quad (5)$$

and further transformed into its logarithmic form

$$\log f_y(\hat{\theta}, y_*^N) = \log\Psi(S(y_*^N), \hat{\theta}) + \log h(y_*^N) \quad (6)$$

Maximizing $f_y(\hat{\theta}, y_*^N)$ with respect to $\hat{\theta}$ is equivalent to maximising $\log\Psi(S(y_*^N), \hat{\theta})$ with respect to $\hat{\theta}$. This implies $f_y(\hat{\theta}, y^N)$ depends on $y^N$ through every sufficient statistic $S(y^N)$. This property of **MLE** is sufficiency [7].

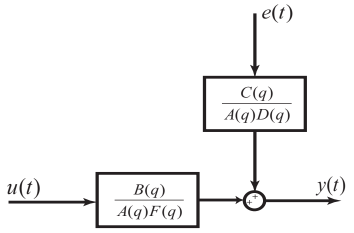


Fig. 1. General Linear Model Structure of **SISO** Open Loop Systems

In the scope of this paper only **MLE** in open loop is considered. In Fig.1, a general open loop process is shown. We assume the system dynamics can be described by the common family of model structures[9]

$$A(q)y(t) = \frac{B(q)}{F(q)}u(t) + \frac{C(q)}{D(q)}e(t) \quad (7)$$

We also assume that the estimate model is governed by

$$\hat{A}(q)y(t) = \frac{\hat{B}(q)}{\hat{F}(q)}u(t) + \frac{\hat{C}(q)}{\hat{D}(q)}\varepsilon(t, \hat{\theta}) \quad (8)$$

where the polynomials $\hat{A}(q)$, $\hat{B}(q)$, $\hat{C}(q)$, $\hat{D}(q)$ and $\hat{F}(q)$ are rational functions characterized by

$$\hat{A}(q) = 1 + \hat{a}_1 q^{-1} + \ldots + \hat{a}_{n_a} q^{-n_a} \quad (9)$$

$$\hat{B}(q) = \hat{b}_1 q^{-1} + \ldots + \hat{b}_{n_b} q^{-n_b} \quad (10)$$

$$\hat{C}(q) = 1 + \hat{c}_1 q^{-1} + \ldots + \hat{c}_{n_c} q^{-n_c} \quad (11)$$

$$\hat{D}(q) = 1 + \hat{d}_1 q^{-1} + \ldots + \hat{d}_{n_d} q^{-n_d} \quad (12)$$

$$\hat{F}(q) = 1 + \hat{f}_1 q^{-1} + \ldots + \hat{f}_{n_f} q^{-n_f} \quad (13)$$

In the notations above, $u(t)$, $y(t)$ and $e(t)$ are the input, output and noise signal respectively. The super-index "^" represents the estimates. Combining (7) and (8), the one-step-ahead prediction error $\varepsilon(t, \hat{\theta})$ can be expressed as

$$\varepsilon(t, \hat{\theta}) = \frac{\hat{D}(q)}{\hat{C}(q)}(\frac{\hat{A}(q)B(q)}{A(q)F(q)} - \frac{\hat{B}(q)}{\hat{F}(q)})u(t) + \frac{\hat{A}(q)C(q)\hat{D}(q)}{A(q)\hat{C}(q)D(q)}e(t) \quad (14)$$

Its generation is shown in Fig. 2. The estimate coefficient vector $\hat{\theta} = [\hat{a}_1 \ldots \hat{a}_{n_a} \, \hat{b}_1 \ldots \hat{b}_{n_b} \, \hat{c}_1 \ldots \hat{c}_{n_c} \, \hat{d}_1 \ldots \hat{d}_{n_d} \, \hat{f}_1 \ldots \hat{f}_{n_f}]^T$ and the true parameter $\theta$ in a similar manner are defined. It is worthwhile to point out that if necessary we will also use the notations $[A(q)\, B(q)\, C(q)\, D(q)\, F(q)]$ and $[\hat{A}(q)\, \hat{B}(q)\, \hat{C}(q)\, \hat{D}(q)\, \hat{F}(q)]$ to represent $\theta$ and $\hat{\theta}$ respectively.

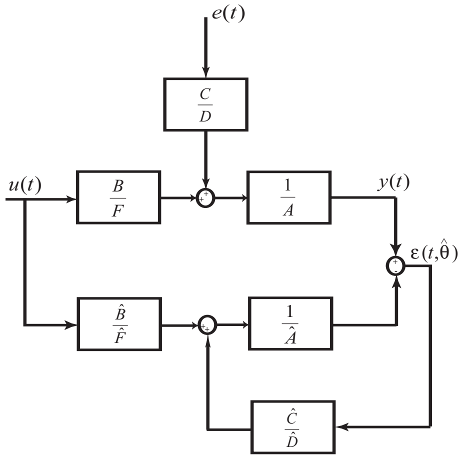


Fig. 2. Generation of Model Based One-step-ahead Prediction Error $\varepsilon(t, \hat{\theta})$

| Model Structures | Characteristics |
|---|---|
| **ARMAX**[2] | $D(q) = F(q) = 1$ |
| **ARARMAX**[9] | $F(q) = 1$ |
| **OE**[6] | $A(q) = C(q) = D(q) = 1$ |
| **BJ**[4] | $A(q) = 1$ |

TABLE I
DEFINITIONS OF MODEL STRUCTURES

To clarify different model structures used in this paper, their definitions are provided in TABLE I. According to it, **ARMAX** and **OE** models can be produced by **ARARMAX** and **BJ** models via choosing $D(q) = 1$ and $C(q) = D(q) = 1$ respectively. For reference convenience, we omit the forward shift operator $(q)$ and its frequency domain interpretation $(e^{j\omega})$ when there is no misunderstanding. In addition, we define an auxiliary notation

$$d(\hat{X}, \hat{Y}) \quad = \quad X(q)\hat{Y}(q) - \hat{X}(q)Y(q) \quad (15)$$

where $\hat{X}(q)$ and $\hat{Y}(q)$ are the corresponding estimate polynomials of $X(q)$ and $Y(q)$.

The following assumptions are postulated through out this paper:
(AP1). The input $u(t)$ persistently exciting of relevant system orders is deterministic and periodic or filtered white noise [13].
(AP2). The noise signal $e(t)$ and true prediction error $\varepsilon(t, \theta)$ are identical independent distributed subject to $N(0, \sigma^2)$ where the variance is known.
(AP3). The true polynomials $A$, $C$, $D$, $F$ and estimated

polynomials $\hat{A}$, $\hat{C}$, $\hat{D}$, $\hat{F}$ all have the roots inside the unit circle.
(AP4). The orders of estimate polynomials are equal to the true ones, i.e. $n_a = \hat{n}_a$, $n_b = \hat{n}_b$, $n_c = \hat{n}_c$, $n_d = \hat{n}_d$ and $n_f = \hat{n}_f$.
(AP5). The number of data $N$ goes to infinity.

Next let us derive the likelihood function for $y^N$ which is described by the estimate model (8). Assume $\varepsilon(t,\hat{\theta})$ has the **PDF** $f_e(\varepsilon(t,\hat{\theta}),t,\hat{\theta})$. According to the joint **PDF** for the observations $y^N$ provided in **Lemma 5.1** in [9], the likelihood function turns to be

$$f_y(\hat{\theta},y^N) = \prod_{t=1}^{N} f_e(\varepsilon(t,\hat{\theta}),t,\hat{\theta}) \tag{16}$$

Maximizing (16) is equivalent to minimizing

$$-\frac{1}{N}\log(f_y(\hat{\theta},y^N)) = -\frac{1}{N}\sum_{t=1}^{N}\log f_e(\varepsilon(t,\hat{\theta}),t,\hat{\theta}) \tag{17}$$

When $\hat{\theta} = \theta$ holds, i.e. $\varepsilon(t,\hat{\theta}) = e(t)$. Then the random function $f_e(\varepsilon(t,\hat{\theta}),t,\hat{\theta})$ turns to be a Gaussian density function. Thus (17) can be simplified into the loss function

$$V_N(t,\hat{\theta}) = \frac{1}{2N}\sum_{t=1}^{N}\varepsilon^2(t,\hat{\theta}) \tag{18}$$

The value of $V_N(t,\hat{\theta})$ is stochastic with fixed $\hat{\theta}$ and a small $N$. It is hard to analyze the property of $V_N(t,\hat{\theta})$ in this case. In order to avoid this difficulty, combining (AP5) we introduce the asymptotic loss function

$$V(\hat{\theta}) = \lim_{N\to\infty} V_N(t,\hat{\theta}) \quad \text{with probability 1} \tag{19}$$

in the following instead of $V_N(t,\hat{\theta})$. Note that (AP1), (AP2), (AP3) and (AP4) ensure the existence of $V(\hat{\theta})$ [13]. For convenience we replace $\lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}$ with the symbol $\bar{E}$. Then

$$V(\hat{\theta}) = \frac{1}{2}\bar{E}\varepsilon^2(t,\hat{\theta}) \tag{20}$$

holds. Since

$$\bar{E}\varepsilon^2(t,\hat{\theta}) \geq \bar{E}\varepsilon^2(t,\theta) \tag{21}$$

stands under (AP2), maximum likelihood estimator can be obtained by the minimization of (20).

## III. Methodology

The methodology of local minimum design originates from [8]. Goodwin *et al* define the **DEPEN**(**D**ecreasing **E**uclidean **P**arameter **E**rror **N**orm) region as $\Gamma$ in parameter space. Its elements $\hat{\theta}$ are those which will get closer to the true parameter $\theta$ in the Euclidean sense when an infinitesimal step is taken along the negative gradient direction of $V(\hat{\theta})$. For such region, it holds that

***Lemma 1***[8]. $\hat{\theta} \in \Gamma$ *if and only if the inner product between* $\tilde{\theta} = \hat{\theta} - \theta$ *and* $V'(\hat{\theta})$ *is positive, i.e.*

$$\tilde{\theta}^T V'(\hat{\theta}) > 0 \tag{22}$$

**Proof:** Let $\hat{\theta}_i$ denote the current estimate. The **SDS**(**S**teepest **D**escent **S**earch) can be expressed as

$$\hat{\theta}_{i+1} = \hat{\theta}_i - \zeta_i V(\hat{\theta}_i) \tag{23}$$

[9] where $\zeta_i$ is the step size. Subtracting $\theta$ from both sides gives

$$\tilde{\theta}_{i+1} = \tilde{\theta}_i - \zeta_i V(\hat{\theta}_i) \tag{24}$$

Squaring both sides provides

$$\tilde{\theta}_{i+1}^T\tilde{\theta}_{i+1} = \tilde{\theta}_i^T\tilde{\theta}_i - 2\zeta_i\tilde{\theta}_i^T V'(\hat{\theta}_i) + \zeta_i^2(V'(\hat{\theta}_i))^T V'(\hat{\theta}_i) \tag{25}$$

When $\zeta_i$ is sufficiently small, we could neglect the last term on the right side of (25) and prove the lemma. □

According to (22), those $\hat{\theta}$ circled by the dash-line belong to the **DEPEN** region in Fig. 3 since $\tilde{\theta}$ and $V'(\hat{\theta})$ point at the same direction. Conversely, **SDS** starting off at points like $\hat{\theta}_1$ which satisfies

$$\tilde{\theta}_1^T V'(\hat{\theta}_1) \leq 0 \tag{26}$$

could *possibly* converge to the nearest local minimum $\hat{\theta}_2$. We call such $\hat{\theta}_{ini}$ *point of interest* or **POI** and the dash-line circled landscape constituted by **POI** the *region of interest*. To sum up, the construction of local minimum examples can be transformed into the design of **POI** satisfying (26) followed used in **SDS** as the initial value.

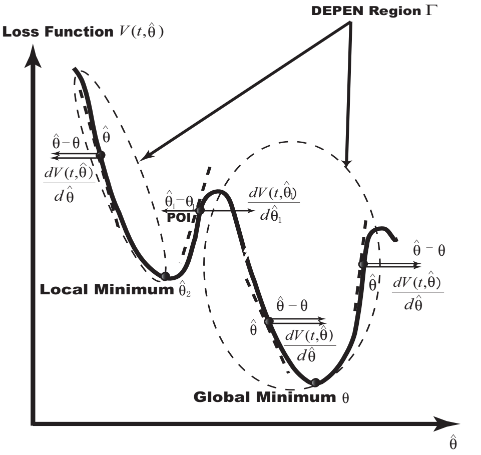

DEPEN Region Γ
Loss Function $V(t,\hat{\theta})$
$\hat{\theta}-\theta$
$\hat{\theta}$
$\frac{dV(t,\hat{\theta})}{d\hat{\theta}}$
$\hat{\theta}_1-\theta_1$
POI
$\frac{dV(t,\hat{\theta}_1)}{d\hat{\theta}_1}$
$\hat{\theta}-\theta$
$\hat{\theta}$
$\frac{dV(t,\hat{\theta})}{d\hat{\theta}}$
Local Minimum $\hat{\theta}_2$
$\hat{\theta}-\theta$
$\hat{\theta}$
$\frac{dV(t,\hat{\theta})}{d\hat{\theta}}$
Global Minimum $\theta$
$\hat{\theta}$

Fig. 3. Illustration of **DEPEN** Region and **POI**

**Remarks:**
(1). In Fig. 3, the directions of the two arrows starting at $\hat{\theta}$ represent the sign of $\tilde{\theta}$ and $V'(\hat{\theta})$ respectively which are both scalars on the two-dimensional plot.

(2). The search may also possible converge to other kinds of stationary point, such as the saddle point.

## IV. Local Minimum Examples Design

In this section we review how to construct local minimum examples for **OE** models[13] first and then implement the method described in section 3 to design local minimum examples for **ARMAX**, **ARARMAX** and **BJ** models.

For open loop **OE** models

$$y(t) = \frac{B}{F}u(t) + e(t) \tag{27}$$

according to the form of the prediction error $\varepsilon(t,\hat{\theta})$ in (14), its asymptotic loss function can be derived as

$$V(\hat{\theta}) = \frac{1}{2}\bar{E}\left[\left(\frac{B}{F} - \frac{\hat{B}}{\hat{F}}\right)u(t)\right]^2 + \frac{1}{2}\bar{E}[e(t)]^2 \tag{28}$$

Söderström suggested that optimizing (28) with respect to $\hat{b}$ firstly, $V(t,\hat{\theta})$ can be transformed into the pseudo-asymptotic function[**?**]

$$\tilde{V}(\hat{f}) = \frac{1}{2}(V_0 - V_1(\hat{f})^T V_2(\hat{f})^{-1} V_1(\hat{f})) \tag{29}$$

where

$$V_0 = \bar{E}\left[\frac{B}{F}u(t)\right]^2 + \bar{E}[e(t)]^2 \tag{30}$$

$$V_1(\hat{f}) = \bar{E}[\frac{B}{F}u(t)]\left[\frac{1}{\hat{F}}u(t-i)\right] \qquad 1 \le i \le n_b \tag{31}$$

$$V_2(\hat{f}) = \bar{E}\left[\frac{1}{\hat{F}}u(t-i)\right]\left[\frac{1}{\hat{F}}u(t-j)\right]^T \qquad 1 \le i,j \le n_b \tag{32}$$

Here $\hat{f} = [\hat{f}_1 \; \dots \; \hat{f}_{n_f}]^T$. $\left[\frac{1}{\hat{F}}u(t-i)\right]$ and $\left[\frac{1}{\hat{F}}u(t-j)\right]$ are $n_f$-column vectors. Since $V_2(\hat{f})$ is positive definite, the following inequality

$$\tilde{V}(\hat{f}) \le \frac{1}{2}V_0 \tag{33}$$

always holds. The equality only holds when $V_1(\hat{f}) = 0$. If the curve of $\tilde{V}(\hat{f}) = \frac{1}{2}V_0$ bisects the stability region of $\hat{f}$, normally there is at least one minimum at each side of the curve.

The **POI** of **ARMAX** and **BJ** models are more or less related to the local minimum of the **OE** models. This can be seen as follows.

*Design of **POI** for **ARMAX** models*: Assume the following **OE** model

$$y(t) = \frac{B(q)}{F(q)}u_1(t) + e(t) \tag{34}$$

has a local minimum at $[\hat{B}_1 \; \hat{F}_1]$. Then we design the provisional **ARMAX** model

$$A_p(q)y(t) = B_p(q)u_p(t) + C_p(q)e(t) \tag{35}$$

where

$$\begin{cases} A_p(q) = F(q) + \delta(q), \\ B_p(q) = B(q), \\ C_p(q) = F(q), \\ u_p(t) = \alpha u_1(t). \end{cases} \tag{36}$$

Here $\alpha$ is a positive coefficient attached before the input signal to adjust the **SNR**(**S**ignal-to-**N**oise-**R**atio)[**?**][**?**] and $\delta(q)$ is a deviation polynomial

$$\delta_1 q^{-1} + \delta_2 q^{-2} \dots + \delta_{n_a} q^{-n_a} \tag{37}$$

introduced to avoid the overlap between **OE** and **ARMAX** models. The selection of it yields to

$$\forall\sqrt{\delta_1^2 + \dots + \delta_{n_a}^2} < \xi \;\; \text{s.t.} \;\; \tilde{\theta}^T V'(t,\hat{\theta}) \le 0 \, \text{at} \; [\hat{F}_1 \; \hat{B}_1 \; \hat{F}_1] \tag{38}$$

where $\xi$ is a small positive scalar. **SDS** initialized at $\hat{\theta} = [\hat{F}_1 \; \hat{B}_1 \; \hat{F}_1]$ probably converges to the nearby local minimum $[\hat{A}_2 \; \hat{B}_2 \; \hat{C}_2]$.

*Design of **POI** for **ARARMAX** models:* Suppose the **ARMAX** model (35) has a local minimum $[\hat{A}_2 \; \hat{B}_2 \; \hat{C}_2]$. For the following **ARARMAX** model

$$A_p y(t) = B_p u_3(t) + \frac{C_p}{D_p}e(t), \tag{39}$$

where the input signal is

$$u_3(t) = \frac{1}{D_p}u(t), \tag{40}$$

the equalities

$$\begin{aligned} \tilde{\theta}^T V'(\hat{\theta}) &= \frac{1}{\pi}\int_0^{\pi} \Re(\frac{C}{\hat{C}})(|G_1|^2 \Phi_{uu}(\omega) + |G_2|^2\sigma^2)d\omega \\ &= 0 \end{aligned} \tag{41}$$

[8][**?**] hold at $\hat{\theta} = [\hat{A}_2 \; \hat{B}_2 \; \hat{C}_2 \; D_p]$. Here

$$G_1 = \frac{d(\hat{B},\hat{A})}{A\hat{C}} \qquad G_2 = \frac{d(\hat{C},\hat{A})}{A\hat{C}} \tag{42}$$

Therefore **SDS** initialized at such point converges to the nearby stationary point $[\hat{A}_3 \; \hat{B}_3 \; \hat{C}_3 \; \hat{D}_3]$. To ensure it is a local minimum, we only apply those $D_p$ which let the Hessian matrix of the stationary point positive definite.

*Design of **POI** for **BJ** models:* It also starts from (34) which has the local minimum $[\hat{B}_1 \; \hat{F}_1]$. For the following **BJ** model

$$y(t) = \frac{B}{F}u_2(t) + \frac{C}{D}e(t), \tag{43}$$

where $B$ and $F$ are the same polynomials as in (34) and the input signal

$$u_2(t) = \frac{C}{D}u_1(t). \tag{44}$$

The inner product of $\tilde{\theta}$ and $V'(\hat{\theta})$ is equal to zero at $[\hat{B}_1 \; C \; D \; \hat{F}_1]$. Hence **SDS** initialized at this point $\hat{\theta} = [\hat{B}_1 \; C \; D \; \hat{F}_1]$ converges to the nearby stationary point $[\hat{B}_2 \; \hat{C}_2 \; \hat{D}_2 \; \hat{F}_2]$ differing from the global minimum. To ensure it is a local minimum point, again we only adopt those polynomials $C$ and $D$ which make the Hessian matrix of $[\hat{B}_2 \; \hat{C}_2 \; \hat{D}_2 \; \hat{F}_2]$ positive definite.

## V. Simulation Examples

In this section we give a local minimum example for each model above. In all examples, the length of all data points $N$ is 10000. Noise signal $e(t)$ has unit variance. For **ARMAX**, **ARARMAX** and **BJ** example, **SDS** based on (23) initialized at the **POI** is applied iteratively until the condition

$$\frac{|V(\hat{\theta}_i) - V(\hat{\theta}_{i-1})|}{V(\hat{\theta}_i)} < 0.0005 \tag{45}$$

is met or a maximum iteration number 20 is reached.

**Example 1:** The dynamics of the **OE** model is given as

$$\begin{cases} B = q^{-1} \\ F = 1 - 1.2q^{-1} + 0.36q^{-2} \end{cases} \tag{46}$$

The input signal is

$$u_1(t) = (1 - 0.72q^{-2} + 0.1296q^{-4})v_1(t) \tag{47}$$

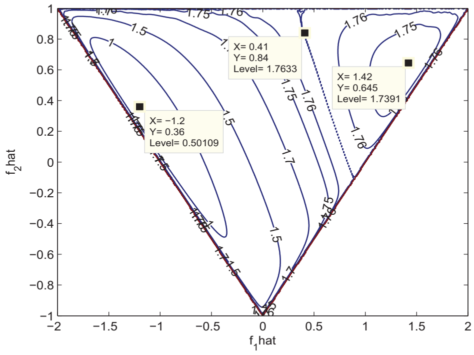


Fig. 4. The Contour of $\tilde{V}(\hat{f})$ for Example 1

Here $v_1(t)$ is i.i.d Gaussian signal with unit variance. The contour of $\tilde{V}(\hat{f})$ is shown in Fig. 4. The curve $\tilde{V}(\hat{f}) =$ 1.76619 bisects the stability triangle where the coefficients $\hat{f}_1$ and $\hat{f}_2$ satisfy [10]

$$-1 < \hat{f}_2 < 1 \tag{48a}$$

$$\hat{f}_1 - 1 < \hat{f}_2 \tag{48b}$$

$$-\hat{f}_1 - 1 < \hat{f}_2 \tag{48c}$$

into two subsets. Each subset has one minimum. They are

$$\theta = [1 \ -1.200 \ 0.360]^T \quad \text{(global minimum)} \tag{49a}$$

$$\hat{\theta}_1 = [-0.121 \ 1.419 \ 0.670]^T \quad \text{(local minimum)} \tag{49b}$$

**Example 2:** Given an **ARMAX** model where

$$\begin{cases} A_p(q) = (1 - 1.2q^{-1} + 0.36q^{-2}) + \delta(q) \\ B_p(q) = q^{-1} \\ C_p(q) = 1 - 1.2q^{-1} + 0.36q^{-2} \end{cases} \tag{50}$$

The input signal of the system is

$$u_p(t) = 0.3u_1(t) \tag{51}$$

where $u_1(t)$ is the input signal used in example 1. At

$$\hat{\theta} = [1.419 \ 0.670 \ -0.121 \ 1.419 \ 0.670]^T \tag{52}$$

selecting

$$\delta(q) = 0.2q^{-1} + 0.09q^{-2} \tag{53}$$

makes

$$\tilde{\theta}^T V'(\hat{\theta}) = -0.0305 \tag{54}$$

which implies (38) satisfied. Starting from (52), **SDS** converges to the following stationary point

$$\begin{cases} \hat{A}_2 = 1 + 1.304q^{-1} + 0.5435q^{-2} \\ \hat{B}_2 = -0.1035q^{-1} \\ \hat{C}_2 = 1 + 1.386q^{-1} + 0.6022q^{-2} \end{cases} \tag{55}$$

after four iterations. Its Hessian matrix

$$\begin{bmatrix} 7.07 & -5.99 & 0.14 & -7.94 & 6.50 \\ -5.99 & 7.07 & -0.05 & 7.16 & -7.94 \\ 0.14 & -0.05 & 0.33 & -0.16 & 0.02 \\ -7.94 & 7.16 & -0.16 & 9.11 & -7.92 \\ 6.50 & -7.94 & 0.02 & -7.92 & 9.11 \end{bmatrix}$$

is positive definite. The trace of loss function in **SDS** is shown in Fig. 5.

**Example 3:** For such **ARARMAX** model which has the same $A_p$, $B_p$ and $C_p$ with in example 2, we assign its $D_p$ and input signal as

$$\begin{cases} D_p = 1 + 0.8257q^{-1} \\ u_3(t) = \dfrac{1}{D_p} u_p(t) \end{cases} \tag{56}$$

**SDS** starting at $\hat{\theta} = [\hat{A}_2 \ \hat{B}_2 \ \hat{C}_2 \ D_p]$ eventually ends at the stationary point

$$\begin{cases} \hat{A}_3 = 1 + 1.312q^{-1} + 0.5464q^{-2} \\ \hat{B}_3 = -0.1035q^{-1} \\ \hat{C}_3 = 1 + 1.377q^{-1} + 0.6009q^{-2} \\ \hat{D}_3 = 1 + 0.815q^{-1} \end{cases} \tag{57}$$

after five iterations. The Hessian matrix of this point

$$\begin{bmatrix} 7.06 & -5.98 & 0.14 & -7.67 & 6.37 & 4.72 \\ -5.98 & 7.06 & -0.05 & 6.73 & -7.67 & -3.85 \\ 0.14 & -0.05 & 0.32 & -0.16 & 0.03 & 0.12 \\ -7.67 & 6.73 & -0.16 & 8.43 & -7.25 & -5.03 \\ 6.37 & -7.67 & 0.03 & -7.25 & 8.43 & 4.11 \\ 4.72 & -3.85 & 0.12 & -5.03 & 4.11 & 4.13 \end{bmatrix}$$

is positive definite. The trace of loss function in **SDS** is shown in Fig. 6.

**Example 4**: The true polynomials of this **BJ** model are

$$\begin{cases} B = q^{-1} \\ C = 1 + 0.6428q^{-1} \\ D = 1 - 0.6616q^{-1} + 0.1792q^{-2} \\ F = 1 - 1.2q^{-1} + 0.36q^{-2} \end{cases} \tag{58}$$

Its input signal is

$$u_2(t) = \frac{C}{D} u_1(t) \tag{59}$$

Let **SDS** begins at $\hat{\theta} = [\hat{B}_1 \; C \; D \; \hat{F}_1]$. After eight steps of iteration, **SDS** ends up at the following local minimum point

$$\begin{cases} \hat{B}_2 = -0.0804q^{-1} \\ \hat{C}_2 = 1 + 0.6927q^{-1} \\ \hat{D}_2 = 1 - 1.062q^{-1} + 0.4294q^{-2} \\ \hat{F}_2 = 1 + 1.398q^{-1} + 0.6252q^{-2} \end{cases} \tag{60}$$

Its Hessian matrix

$$\begin{bmatrix} 7.13 & -1.95 & -0.37 & -0.94 & -0.41 & -0.19 \\ -1.95 & 6.09 & -1.39 & 0.99 & 0.32 & -0.17 \\ -0.37 & -1.39 & 7.57 & 5.61 & 0.29 & 0.11 \\ -0.94 & 0.99 & 5.61 & 7.57 & 0.09 & -0.44 \\ -0.41 & 0.32 & 0.29 & 0.09 & 0.24 & -0.23 \\ -0.19 & -0.17 & 0.11 & -0.44 & -0.23 & 0.66 \end{bmatrix}$$

is positive definite. The trace of loss function in **SDS** is shown in Fig. 7.

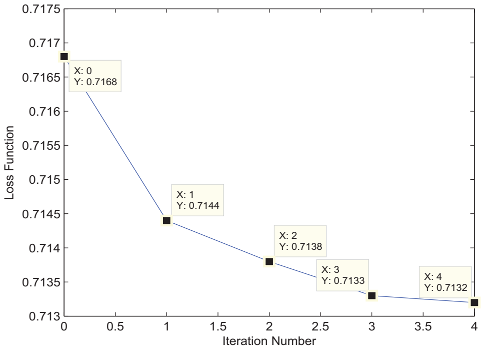


Fig. 5. Evaluation of Asymptotic Loss Function $V(\hat{\theta})$ in **SDS** for Example 2

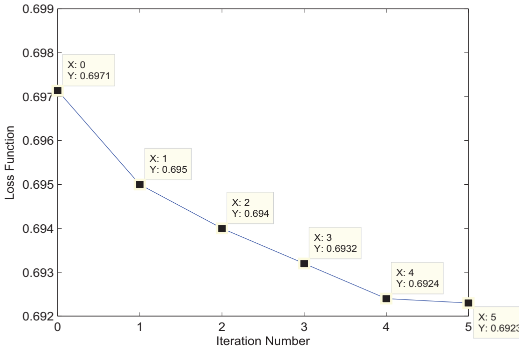


Fig. 6. Evaluation of $V(\hat{\theta})$ in **SDS** for Example 3

## VI. CONCLUSIONS AND FUTURE WORKS

In this paper, we design the **POI** followed by **SDS** to construct local minimum examples for open loop **ARMAX**, **ARARMAX** and **BJ** models. In particular, the **POI**s for **ARMAX** and **BJ** models have strong links to the local minimum in the corresponding **OE** models. Furthermore, simulation examples are also provided for each model structure above. These examples play as a key to the development of the non-local-minimum conditions in **MLE**. In the future, we will investigate the non-local-minimum conditions for these models starting by changing the examples dynamics etc.

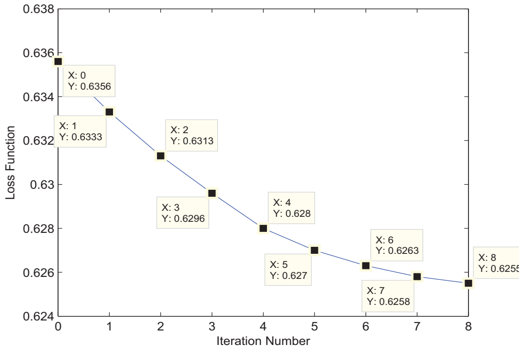


Fig. 7. Evaluation of $V(\hat{\theta})$ in **SDS** for Example 4

## VII. ACKNOWLEDGMENTS

The authors would like to thank Dr. Daniel Coca and Dr. Tim Breikin for their constructive comments on the way of discussing the simulation examples.

## REFERENCES

[1] Åström, K.J. *Maximum Likelihood and Prediction Error Methods*. Automatica, 16:551–574, 1980.

[2] Åström, K.J. and Bohlin, T. *Numerical Identification of Linear Dynamic Systems from Normal Operating Records*. In IFAC Symposium on Self-Adaptive Systems, Teddington, UK, 1965.

[3] Åström, K.J. and Söderström, T. *Uniqueness of Maximum Likelihood Estimates of the Parameters of an ARMA model*. IEEE Transactions on Automatic Control, 19(6):769-773, 1974

[4] Box, G.E.P. and Jenkins, G.M. *Time Series Analysis, Forecasting and Control(3rd ed. 1994)*. Holden-Day, 1970.

[5] Cramér, H. *Mathematical Methods of Statistics*. Princeton University Press, Princeton, 1946.

[6] Dugard, L. and Landau, I.D. *Recursive Output Error Identification Algorithms*. Automatica, 16:443-462, 1980. Messenger of Mathmatics, 41:155–160, 1912.

[7] Fisher, R.A. *On the Mathematical Foundations of Theoretical Statistics*. Philosophical Transactions of the Royal Society of London. Series A, 222:309–368, 1921.

[8] Goodwin, G.C., Carlos, J.A. and Skelton, R.E. *Conditions for Local Convergence of Maximum Likelihood Estimation for ARMAX Models*. In 13th IFAC Symposium on System Identification, Rotterdam, The Netherland, 2003.

[9] Ljung, L. *System Idenfication:Theory for the User, 2nd Edition*. Prentice Hall, 1999.

[10] Ogata, K. *Modern Control Engineering, 3rd Edition*. Prentice Hall, 1996.

[11] Pintelon, R. and Schoukens, J. *System Identification: A Frequency Domain Approach*. IEEE Press, 2001.

[12] van Overschee, P. and DeMoor, B. *Subspace Identification for Linear Systems: Theory, Implementation, Applications*. Kluwer Academic Publishers, 1996.

[13] Söderström, T. *On the Uniqueness of Maximum Likelihood Identification*. Automatica, 11:193–197, 1975.

[14] Wald, A. *Note on the Consistency of the Maximum Likelihood Estimate*. The Annals of Mathematical Statistics, 20(4):595–601, 1949.

[15] Zou, Y. and Heath, W.P. *Global Convergence Conditions in Maximum Likelihood Estimation*. International Journal of Control, DOI: 10.1080/00207179.2012.658085, 2012.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Identification of Structural Parameters Using Damped Transfer Matrix and State Vector

P. Nandakumar, K. Shankar
Department of Mechanical Engineering
Indian Institute of Technology Madras
Chennai 600036, INDIA
Email: ppnkumar74@gmail.com
Telephone: (+91)44-22575742
Fax: (+91)44-22574701

***Abstract*—A new method for identification of structural parameters is proposed using Damped Transfer Matrices (DTM) and state vectors. A new transfer matrix is derived for continuous mass systems including the damping parameters. The state vector at a location is the sum of the internal and external contributions of displacements, forces and moments at that point, when it is multiplied with the transfer matrix, state vector at the adjacent location is obtained. The structural identification algorithm proposed here involves prediction of displacement responses at selected locations of the structure using Damped Transfer Matrix and compares them with the measured responses at the respective locations. The mean square deviations between the measured and predicted responses at all locations are minimized using a non-classical optimization algorithm, and the optimization variables are the unknown stiffness and damping parameters in the DTM. A non-classical heuristic Particle Swarm Optimization algorithm (PSO) is used, since it is especially suited for global search. This DTM algorithm with successive identification strategy is applied on one element or substructure of a structure at a time and identifies all the parameters of adjacent elements successively. The algorithm is applied on numerically simulated experiments of structures such as a cantilever and one sub-structure of a nine member frame structure. Also this algorithm is verified experimentally on a sub-structure of a fixed beam. The main advantage of this algorithm is that it can be used for the local identification in a zone in a structure without modelling the entire global structure.**

***Keywords*— Damped Tranfer Matrix; State Vectors; Successive Identification; Particle Swarm Optimization**

## I. Introduction

Structural identification (SI) problems typically deal with the estimation of mass, stiffness and damping properties of a structure from input/output measurements. It plays an important role in model updating and structural health monitoring. From a computational point of view, structural identification presents a challenging problem particularly when the system involves a large number of unknown parameters. SI algorithms are generally classified into frequency domain and time domain algorithms. Frequency domain SI algorithms have been developed more widely. Maia and Silva [1] presented some modal analysis techniques for identification. Ge and Lui [2] identified damage on structures like cantilever, ten story steel frame and plates by comparing natural frequencies of the undamaged and damaged structures.

Time domain algorithms are usually categorized as Classical or Non-classical methods. Ghanem and Shinozuka [3] reported few classical SI time domain algorithms such as Recursive Least Square method (RLS), Extended Kalman Filter method (EKF), maximum likelihood method, recursive instrumental variable method. Juang and Pappa [4] presented a deterministic SI algorithm based on state space model of second order system using Observer Kalman Filter Identification and Eigen Realisation Algorithm (OKID/ERA) by which all the structural properties such as mass, damping coefficient, stiffness can be identified. Some of the shortcomings of the classical methods are requirement of the calculation of derivatives; difficulty of converging to the global optima, requirement of initial values, and inability to deal with large number of variables. To overcome these drawbacks, Non-classical SI algorithms are used. A non-classical method is usually based on heuristic concepts such as Evolutionary principle (GA) or behavioural principle (PSO). Koh *et al* [5]. identified a maximum of 52 structural parameters including damping using GA with a hybrid local search method. GA directs the search toward the global optima and the local search improves the convergence. Kennedy and Eberhart [6] developed a new stochastic optimization algorithm PSO which was proved that much superior to GA and easy to configure [7]. Perez and Behdinan [8] also used PSO for a structural identification problem of 72 bar truss with good accuracy.

The computational effort of identifying a $n$ DOF system is of the order of $n^2$. Even for a modern computer, the computational speed for solving large matrices is challenging. As an alternative for this problem, transfer matrices and state vectors are used for SI algorithm. Steidel [9] derived the transfer matrix for a spring mass system and a beam element. The transfer matrix for the beam element is derived by assuming that the mass is concentrated only at end nodes and the beam element is mass less throughout its length. Meirovitch [10] determined the natural frequencies and mode shapes of a non-uniform pinned-pinned beam with ten elements using transfer matrices. Nandakumar and Shankar

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

[11] used the transfer matrices derived by Steidel [9] and state vectors first in SI problem and identified successfully the parameters of cantilever and ten DOF lumped mass system. Later Tuma and Cheng [12] derived an improved transfer matrix for beam element with an assumption of the mass of the beam element is concentrated at its mass center. It is found that there is a good improvement in natural frequencies.

Nandakumar and Shankar [13] derived a transfer matrix from the consistent mass matrix of the beam element and determined higher order natural frequencies with much better accuracy than the existing lumped mass based transfer matrices. Using the same transfer matrix, stiffness parameters of structures were identified with better accuracy. However all these transfer matrices discussed have a limitation in that they can be used only for lightly damped structures/materials by ignoring its damping effect. To identify properties of highly damped structures which have significant damping, a new damped transfer matrix (DTM) including damping parameters is derived. In this paper structural parameters including damping parameters were identified using DTM by Successive Structural Identification strategy.

## II. Transfer Matrices and State Vectors

A state vector at a point in the structure is the summation of the internal response vector and external force vector. The former contains the output responses such as displacement, angular displacement and the internal forces and moments and the later contains the externally applied forces and moments. The Transfer Matrix (TM) is a square matrix which contains the structural parameters. When a state vector is multiplied with the TM, internal response vector at the adjacent location is obtained.

### *A. Transfer Matrix for Beam element*

In this section the transfer matrix for damped vibration of beams is derived. The equilibrium equation of two noded beam element is

$$[M]\ddot{x}(t) + [C]\dot{x}(t) + [K]x(t) = F(t) \tag{1}$$

where $\ddot{x}(t)$, $\dot{x}(t)$ and $x(t)$ are nodal acceleration, velocity and displacement responses vector respectively, $F(t)$ is nodal force vector. The state vector for a node on the beam element is $\{X\} = \{y(t), \theta(t), M(t), V(t)\}^T + \{0, 0, F(t), \mu(t)\}^T$, where $y(t)$ is translational displacement, $\theta(t)$ is angular displacement, $M(t)$ is bending moment, $V(t)$ is shear force, $F(t)$ is applied force and $\mu(t)$ is applied moment at that node. The damping in the beam element is modelled using Rayleigh's damping model.

$$[C] = \alpha[M] + \beta[K] \tag{2}$$

Also $\ddot{x}(t) = -\omega^2 x(t)$, $\dot{x}(t) = i\omega x(t)$. therefore, the Eq.(1) becomes,

$$F(t) = [Z]x(t) \tag{3}$$

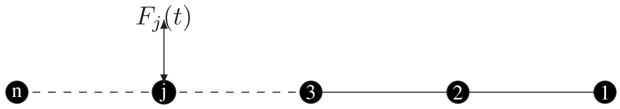


Fig. 1. Sub-structure with arbitrary point excitation

where $[Z] = (i\omega\alpha - \omega^2)[M] + (1 + i\omega\beta)[K]$. Since the beam element is in equilibrium, the Eq.(3) is written for one element

$$\begin{Bmatrix} -M_1(t) \\ -V_1(t) \\ -- \\ M_2(t) \\ V_2(t) \end{Bmatrix} = \begin{bmatrix} Z_{11} & Z_{12} & | & Z_{13} & Z_{14} \\ Z_{21} & Z_{22} & | & Z_{23} & Z_{24} \\ - & - & + & - & - \\ Z_{31} & Z_{32} & | & Z_{33} & Z_{34} \\ Z_{41} & Z_{42} & | & Z_{43} & Z_{44} \end{bmatrix} \begin{Bmatrix} y_1(t) \\ \theta_1(t) \\ -- \\ y_2(t) \\ \theta_2(t) \end{Bmatrix} \tag{4}$$

Assume the output responses and internal forces/moments at node 1 are known and external forces are zero, the force and DOF vectors in the Eq.(4) is rearranged to form state vectors $\{X\}$ in such away that $\{X_2\} = [T_d]\{X_1\}$, the transfer matrix is

$$[T_d] = \begin{bmatrix} -[Q] & [0] \\ -[S] & [I] \end{bmatrix}^{-1} \begin{bmatrix} [P] & [I] \\ [R] & [0] \end{bmatrix} \tag{5}$$

where $[P] = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix}$, $[Q] = \begin{bmatrix} Z_{13} & Z_{14} \\ Z_{23} & Z_{24} \end{bmatrix}$, $[R] = \begin{bmatrix} Z_{31} & Z_{32} \\ Z_{41} & Z_{42} \end{bmatrix}$, $[S] = \begin{bmatrix} Z_{33} & Z_{34} \\ Z_{43} & Z_{44} \end{bmatrix}$

### *B. Transfer matrix and state vector for the global structure*

Calculation of state vector at any node of a global structure, from one known initial state vector at a given node using transfer matrices is illustrated here. For example, a portion of a structure is considered with $n$ nodes subjected to an arbitrary point excitation as shown in Fig.1. It is assumed that all elements in the state vector at the node 1 are known. The state vectors at other nodes can be calculated by successive multiplication of elemental transfer matrices. The equation to calculate internal response vector for $n^{th}$ node from internal response vector at node 1 and external force vector is obtained.

$$\{X_{ni}\} = \left(\prod_{k=1}^{n-1} [T_{(n-k),(n+1-k)}]\right)\{X_{1i}\} + \sum_{j=1}^{n-1}\left(\prod_{k=1}^{n-j} [T_{(n-k),(n+1-k)}]\right)\{X_{je}\} \tag{6}$$

For free damped vibration i.e without any external forces, as a special case the above equation can be deduced as

$$\{X_n\} = \left(\prod_{k=1}^{n-1} [T_{(n-k),(n+1-k)}]\right)\{X_1\} \tag{7}$$

In the above equation, $[T^G_{1,n}] = \prod_{k=1}^{n-1}[T_{(n-k),(n+1-k)}]$ is known as global transfer matrix.

## III. Parameter Identification by Damped Transfer Matrix Method (DTM)

The proposed DTM algorithm is used for identifying the unknown flexural rigidities (EI) of the structure assuming the masses are known. The beam is excited with a known force at a point. The elements of initial state vector is measured at one location, from which it is possible to predict the displacement at any location in the structure using successive multiplication of the DTMs, as discussed in section II-B. The mean square deviation between the predicted and measured displacements at measured locations in the structure can be minimized by Particle Swarm Optimization algorithm (PSO) with the unknown elemental $EI$ values, $\alpha$ and $\beta$ in the DTM as the optimization variables. The Successive identification strategy [13] is adopted to identify unknown parameters since it is superior in speed and accuracy for the identification of a few adjacent elements.

Since the elements of DTM are complex numbers, the elements of all predicted state vectors are complex responses except the initial state vector. Its elements are all real signals since they are measured directly. The imaginary part of the responses in the predicted state vectors is proportional to the contribution of damping property of the structure. For the lightly damped structures/materials the imaginary part of the responses is very small and hence may be neglected and only real part of the response is considered for identification. But for significantly damped structures the imaginary part of the responses is considerable, hence the complex responses are converted from Cartesian form into Polar form. Let the polar form of the responses contain magnitude ($u_e$) and phase shift ($\phi$). The number of time steps are to be shifted is calculated by the relation $t_s = \frac{\phi f_s}{\omega}$ where, the $f_s$ is sampling frequency in Hz and $\omega$ is circular frequency of excitation in r/s. The error function between measured and predicted responses is given by

$$\varepsilon = \frac{\sum_{j=1}^{L} |u_m(j) - u_e(j+t_s)|^2}{L} \tag{8}$$

where $u_m(j)$ and $u_e(j)$ are measured and estimated displacement responses respectively at $j^{th}$ time step. $L$ is the number of time steps. Then the cycle is repeated for all the pairs of adjacent measured responses and identify all unknown parameters successively. Since, the number of unknown parameters to be identified is one or few for one identification cycle, the convergence is very fast, the overall computational time is very small. This strategy is promising in the identification of local parameters in a structural member.

## IV. Numerical Examples and Results

The SI algorithm using damped transfer matrix (DTM) is applied on two numerically simulated experiments. i.e a uniform cantilever and a sub-structure of a nine member frame structure. The structure is excited by a harmonic force at a node and the acceleration responses are measured at selected nodes and converted to displacement responses by numerical integration. In all examples, measured responses are numerically simulated using Newmark's constant acceleration method. The unknown stiffnesses and damping parameters are searched by PSO algorithm within the search range of 50% to 150% of the exact values. In order to simulate the effect of noise in experiments, Gaussian random noise of 3% is added to the measured signals.

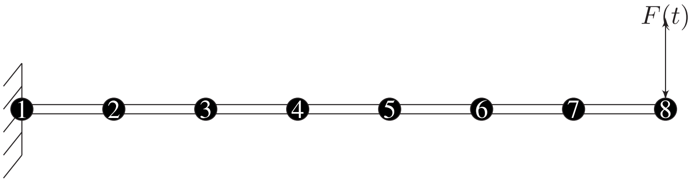


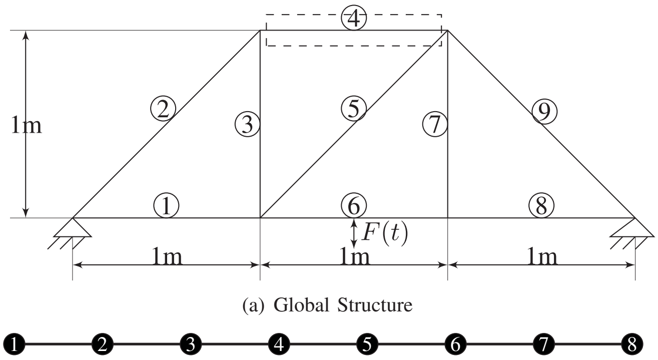
Fig. 2. Finite Element model of cantilever

### A. Example-1: Cantilever

A steel cantilever of dimension 24.6 × 5.7 × 350 mm is fixed at its one of the end as shown in Fig.2. The Young's modulus of cantilever material is 200 GPa and its density is 7691 kg/m$^3$. It is divided into seven finite elements of length 50mm each. The flexural rigidity ($EI$) of each element is 75.93 N.m$^2$. The damping constants $\alpha$ and $\beta$ are 20.77 and 5.71×10$^{-5}$ respectively. The free end is excited by a harmonic input force of $1.5sin(2\pi 10t)$ N. The first natural frequency of the beam is 38.33 Hz. The effect of the damping was accounted by Rayleigh's damping with modal damping ratio of 5% at its first two modes. Since the bending moment and shear force responses are zero at the free end, the initial state vector is formed at the free end. The translational responses are measured at all nodes and angular response is measured at the free end only. From the initial state vector the unknown parameters are identified using DTM successively with PSO parameters of swarm size 50 and 50 iterations in each identification cycle with variable inertia weight varies from 0.9 to 0.4. The identification algorithm is repeated with translational responses measured at nodes 3, 5, 7 and 8. The cantilever is divided into four substructures from nodes (1-3), (3-5), (5-7) and (7-8) and parameters are identified in each substructure successively. This problem was repeated with 3%

TABLE I
Percentage of error in identified results of cantilever

| Element | % of Error | | | |
|---|---|---|---|---|
| | Complete Measurement | | Incomplete Measurement | |
| | Noise free | 3% Noise | Noise free | 3% Noise |
| 1 | -0.19 | 0.98 | 3.57 | 4.95 |
| 2 | 0.17 | 1.88 | -1.19 | -1.54 |
| 3 | 0.01 | -0.45 | -0.15 | 0.79 |
| 4 | -0.01 | -1.96 | 3.68 | -3.21 |
| 5 | -0.02 | -0.56 | -1.06 | 4.99 |
| 6 | -0.04 | 0.64 | 0.52 | 1.19 |
| 7 | -0.05 | 2.51 | -0.05 | 0.27 |
| MAE | 0.07 | 1.28 | 1.46 | 2.42 |
| $\alpha$ | -0.71 | 5.83 | -1.56 | 7.40 |
| $\beta$ | -18.21 | -21.43 | -21.35 | 24.41 |



(a) Global Structure

(b) Sub-structure-Member 4

Fig. 3. Frame Structure

noise at all measured data. The total computational time is 48.5s with complete measurements and the computational time is 27.7s with incomplete measurements. The percentage of mean absolute error in identified values of $EI$ with complete and incomplete measurements is tabulated in Table.I. In this example also the mean absolute error in identified $EI$ values with complete measurements (0.07%) is less than that with incomplete measurement (1.46%) of responses. Similarly the percentage of error in identified damping constants ($\alpha$ and $\beta$) is also more in the case of incomplete measurement (-0.71% and -18.21%) than that of complete measurement case (-1.56% and -21.35%). Hence the DTM derived for beam element is satisfactorily identifying stiffness and damping properties of the beam.

*1) Comparison of results with other time domain methods:* Sandesh and Shankar [14] identified stiffness parameters of a substructure in a similar cantilever with six elements using a time domain Least Square technique with damping included with mean absolute error of 4.51% at 3% noise level which requires measurement at all DOF including rotation. Whereas the DTM algorithm identified results with mean absolute error of 2.42% at 3% noise level with only four translational response measurement and only one angular response measurement. This shows that the DTM algorithm performs well when compared to other SI algorithms.

### *B. Example-2: Sub-structural Identification of frame structure*

A frame steel structure made of nine members is fixed at two supports as shown in Fig.3(a) and it has taken from [15]. Each member has a cross section of 12 × 6mm and a flexural rigidity ($EI$) of 43.2 N.m$^2$. The first natural frequency is 11.9 Hz. The damping effect is taken into account by adopting Rayleigh damping with the modal damping ratio of 5%. The damping constants $\alpha$ and $\beta$ are calculated as 3.919 and 6.36×10$^{-4}$. It is proposed to identify the properties of the top horizontal member 4, which has a length of 1m. The properties of substructure to be identified is indicated by box in Fig.3(a). which has a length of 0.875 m. It is divided into seven elements as shown in Fig.3(b). The structure is excited by a sinusoidal input force of $10sin(2\pi 10t)$ N at the midpoint of the member 6. Since the boundary conditions are unknown, it is necessary to measure translational and angular responses, shear force and bending moment responses of any arbitrary node to define initial state vector. The first two responses can be measured directly by accelerometers and last two responses have to be measured by strain gauges. The initial state vector is formed at the node 8 which is $\{X_8\} = \{y_8(t), \theta_8(t), M_8(t), V_8(t)\}^T$. Since the external excitation force is not applied on the substructure, it is not necessary to measure.

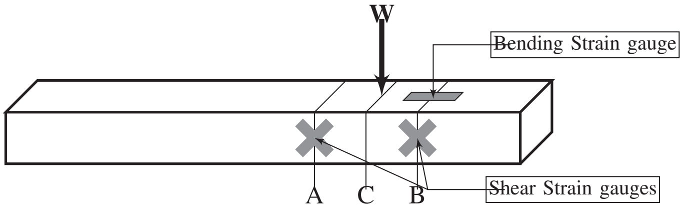


Fig. 4. Strain gauge arrangements

*1) Measurement of shear force and bending moment responses:* For a rectangular section beam, the bending moment response is given by

$$M(t) = \frac{2EI\epsilon_B(t)}{h} \tag{9}$$

The shear force in the section is given by

$$V(t) = \frac{4EI\epsilon_S(t)}{h^2(1+\nu)} \tag{10}$$

where $EI$ is flexural rigidity of the section and $y = h/2$, $h$ is thickness of the section, $\nu$ is Poisson's ratio. The bending strain response can be measured using strain gauge, From the above formulae, to calculate the bending moment and shear force responses at a node, the knowledge of the flexural rigidity ($EI$) at that node is required. The estimation of the $EI$ value at the starting node using a simple shear strain test as is presented here. The strain gauges are fixed to measure bending and shear strain as shown in Fig.4. At a point C in between the nodes A and B, a static load of $W$=10 kN is applied and the corresponding strain at the nodes A and B are measured. Let the static strain measured at the nodes A and B be $\epsilon_{SA}$ and $\epsilon_{SB}$ respectively. The change in shear force at the nodes A and B is equal to the applied load $W$ at C since the self weight of the portion AB is very small. The $EI$ at the initial node B is given by

$$EI = \frac{Wh^2(1+\nu)}{4(\epsilon_{SA} - \epsilon_{SB})} \tag{11}$$

for the measured values $\epsilon_{SA}$=9.8571 × 10$^{-5}$, $\epsilon_{SB}$=-0.0013, $\nu$=0.37 and $h$=6 mm, the $EI$ at the node 8 is obtained as $EI_8$=43.2 N.m$^2$. The bending strain gauges and shear strain gauges fixed at only B are further required for the dynamic strain measurement. The bending moment and shear force responses at the initial node 8 is calculated from measured strain responses using Eq.(9) and Eq.(10).

*2) Parameter Identification:* The initial state vector is formed at the node 8. In case of complete measurement, translational responses are measured at all eight nodes and angular response is measured at initial node only. The parameters were identified by PSO with parameters of swarm size 50 and 50 iterations in each cycle. The total time of identification for seven EI values with complete measurement was 43.84s. The same example is identified with incomplete measurements also. Translational displacement measurements at nodes 1, 3, 5 and 8 are used here. The structure is divided into three portions between nodes (1-3), (3-5) and (5-8). The SI algorithm starts from the initial state vector at the node 8 and identifies parameters of each portion successively. PSO used a swarm size of 50 and 100 iterations in each cycle. The total computational time for convergence was 40.98s. The percentage of error variation in identification of parameters is shown in Table.II. The mean absolute error in identified results of EI is 1.01% with complete measurements and is 4.07% with only four sensors. The DTM performs satisfactorily at 3% Noise level at all measured responses. Also the damping constants are identified with maximum percentage error of -12.84% at incomplete noisy measurement. The input force response is not needed for computation and no need to measure the same. The main advantage of the transfer matrix method is that it is more suitable for the identification of local parameters of complex structure without analysing the entire structure.

TABLE II
PERCENTAGE OF ERROR IN IDENTIFIED RESULTS OF SUBSTRUCTURE OF FRAME

| Element | % of Error | | | |
|---|---|---|---|---|
| | Complete Measurement | | Incomplete Measurement | |
| | Noise free | 3% Noise | Noise free | 3% Noise |
| 1 | 1.34 | -1.40 | 4.99 | 9.99 |
| 2 | 0.29 | 2.99 | -1.84 | -7.71 |
| 3 | 0.96 | 0.05 | 5.02 | 4.99 |
| 4 | -0.16 | -0.33 | 4.11 | -4.38 |
| 5 | -0.41 | 1.53 | -4.54 | -3.26 |
| 6 | -0.14 | -0.74 | -0.58 | -1.84 |
| 7 | -0.01 | -0.04 | 0.23 | 0.74 |
| MAE | 0.47 | 1.01 | 3.05 | 4.07 |
| $\alpha$ | -1.84 | -6.97 | -3.18 | -9.45 |
| $\beta$ | -5.56 | -7.14 | 9.98 | -12.84 |

*3) Comparison of results with other Time domain methods:* It may be noted that Prashanth and Shankar [15] had identified this problem with a 2 stage neural network trained with time domain acceleration signals at two nodes. The error of identification there is expressed as a non-dimensional damage index based on the change in modulus of elasticity of an element. The mean error incurred in that method. was about 0.99% for non-noisy signal and 2.1% for signal with 5% noise but the computational effort of training the network and the complexity of using a two stage network has to be contrasted with the simplicity of the DTM method.

## V. EXPERIMENTAL VERIFICATION OF SUB-STRUCTURE OF A FIXED BEAM

A beam made of acrylic material with cross sectional dimension of 25 ×12 mm and length of 660 mm was fixed at both ends as shown in Fig 5. The modulus of Elasticity ($E$) and density were estimated as 3.7GPa and 1190kg/m$^3$ respectively by simple experiments. The actual flexural rigidity ($EI$) of the beam is 13.32 N.m$^2$. The damping ratio($\zeta$) was calculated from a simple free vibration decay test and estimated as 7%. The natural frequencies for the first two modes are 49.04Hz and 135.45Hz. Assuming Rayleigh's damping model, the exact values of damping constants $\alpha$ and $\beta$ were calculated as 31.67 and 1.21×10$^{-4}$ respectively. The beam was divided into seven elements. A substructure of length 320mm, is shown inside the dotted rectangle in Fig.5 was considered for structural identification. The sub-structure has four elements of length 80 mm each. The node 6 was taken as starting node and the $EI$ value at that node is required to form state vector at node 6. The flexural rigidity ($EI_6$) at the starting node was identified by conducting a static experiment as explained in the section IV-B1. A static load of W=5.045kgf (49.49N) was applied at a point C in between nodes 5 and 6 at a distance 20 mm from the node 6. Strain gauges were fixed as shown in Fig 4. Five sets of readings were taken and the mean values of the shear strains at the points A and B are $\epsilon_{SA} = 9.625\mu$ strain and $\epsilon_{SB} = -170\mu$ strain respectively. The Poisson's ratio ($\nu$) of the beam material is 0.37. Substituting the values in the Eq.(11), the flexural rigidity at the starting node $EI_6$ was obtained which is 13.48 N.m$^2$.

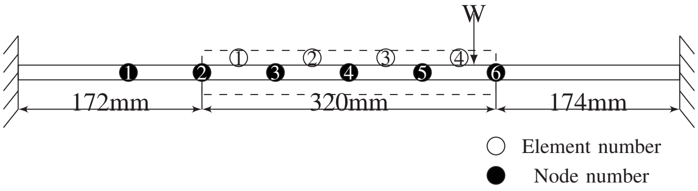


Fig. 5. Substructure of fixed beam

To measure dynamic response, one DYTRAN miniature accelerometer of 2gm mass with sensitivity of 107 mV/g and acceleration range of 50g was fixed at each node to measure translational acceleration. Two accelerometers were fixed very close to each other at a distance of $dx$=7 mm at the starting node 6. The experimental set up is shown in Fig.6. The structure is excited by a sinusoidal force of $3.4sin(2\pi 80t)$ N at the node 1 by a LDS permanent magnet 20 N modal shaker with a maximum displacement of 5 mm with an operating frequency range of 5 Hz-13 kHz. The measurement of input force was not required for this problem, since it was applied outside of the substructure. The strain and acceleration responses were acquired by 16 channel DEWE 1201 data acquisition card(DAC) at a sampling frequency of 1000 Hz. The angular acceleration at the starting node 6 ($\alpha_6$) was calculated by central difference formula. The translational acceleration at the starting node is the mean value of acceleration measured by two accelerometers. Both translational and angular accelerations were converted into respective displacement responses.

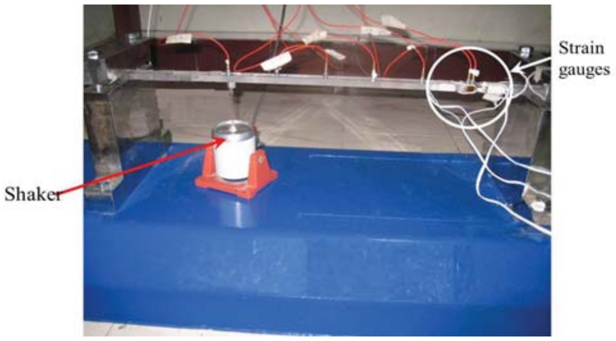


Fig. 6. Experimental set up of Fixed beam

The bending moment and shear force responses at the starting node 6 were calculated from the measured bending and shear strain responses using Eq.(9) and Eq.(10) and state vector at the node 6 was formed. The responses at the other nodes were determined from the starting state vector using DTM with predicted values of structural parameters by PSO. The mean square error between the measured and predicted values of responses were minimized using the Eq.(8). Since there are uncertainty in the experimental data, the parameters were searched between 50% and 200% of their exact value by PSO using damped transfer matrix. The PSO parameters are 50 swarm size, 100 iterations for each identification cycle. The identified parameters are shown in the Table.III. The total computational time taken for convergence is 80.4s. The farthest element 1 was identified with least accuracy, since the error in each identification cycle is accumulated in the state vectors of the succeeding nodes. The DTM algorithm identified the damping properties of the beam with 24.7% and 13.5% mean absolute error. Hence it is clear that the DTM algorithm works on any sub-structure without considering the global model of the complete structure and identifies its local parameters with good accuracy.

TABLE III
IDENTIFIED PARAMETERS OF SUB-STRUCTURE OF FIXED BEAM

| Parameter | Exact | Identified | % of Error |
|---|---|---|---|
| | N.m$^2$ | N.m$^2$ | |
| $EI_1$ | 13.32 | 11.05 | -17.04 |
| $EI_2$ | 13.32 | 12.90 | -3.16 |
| $EI_3$ | 13.32 | 14.43 | 8.39 |
| $EI_4$ | 13.32 | 12.33 | -7.42 |
| $\alpha$ | 31.67 | 39.49 | 24.7 |
| $\beta$ | $1.21 \times 10^{-4}$ | $1.37 \times 10^{-4}$ | 13.50 |

## VI. CONCLUSION

A new SI method based on damped transfer matrix is presented here. The initial state vector has to be provided, and displacement at any point in the structure is predicted using transfer matrix which contains all the structural properties. Using PSO algorithm, the mean square error between measured and predicted responses can be minimized with the unknown structural parameters as the optimization variables. Since the size of the transfer matrix does not increase whatever be the model size, computational effort is reduced. However successive operations are required to identify the unknown parameters. The successive identification method of this algorithm works fast and identifies the structural parameters with good accuracy. Numerical and experimental studies have been made on both global structures with known boundary conditions and sub-structures of unknown boundary conditions. Since the responses measured in the experimental study contains realistic noise and errors, the results have less accuracy when compared with numerically simulated study with artificially added noise. The accuracy of the method is good compared to existing time domain SI methods. A main advantage of this algorithm is the local identification of parameters of structures without the need to model the entire global structure.

## ACKNOWLEDGEMENT

The authors gratefully acknowledge the research grant provided by Council for Scientific and Industrial Research (CSIR), New Delhi, India.

## REFERENCES

[1] N. M. M. Maia and J. M. M. Silva, “Modal analysis identification techniques,” *Philosophical Transaction of the Royal Society A*, vol. 359, pp. 29–40, 2001.

[2] M.Ge and E.M.Lui, “Structural damage identification using systen dynamic properties,” *Computers and Structures*, vol. 83, pp. 2185–2196, 2005.

[3] R. Ghanem and M. Shinozuka, “Structural system identification -1: Theory,” *Journal of Engineering Mechanics*, vol. 121, pp. 255–264, 1995.

[4] J. N. Juang and R. S. Pappa, “An eigensystem realization algorithm for modal parameter identification and model reduction,” *Journal of Guidance Control and Dynamics*, vol. 8, no. 5, pp. 620–627, 1985.

[5] C. G. Koh, Y. Chen, and C. Liaw, “A hybrid computational strategy for identification of structural parameters,” *Computers and Structures*, vol. 81, no. 2, pp. 107–117, 2003.

[6] J. Kennedy and R. Eberhart, “Particle swarm optimization,” in *Proc. IEEE International Conference on Neural Networks*, vol. 4, nov/dec 1995, pp. 1942 –1948.

[7] C. R. Mouser and S. A. Dunn, “Comparing genetic algorithm and particle swarm optimization for inverse problem,” *ANZIAM Journal*, vol. 46, pp. 89–101, 2005.

[8] R. E. Perez and K. Behdinan, “Particle swarm approach for structural design optimization,” *Computers and Structures*, vol. 85, pp. 1579–1588, 2007.

[9] R. F. Steidel., *An Introduction to Mechanical Vibrations*, 2nd ed. John Wiley and Sons, 1978.

[10] L. Meirovitch, *Fundamentals of Vibrations*, 1st ed. McGraw-Hill Book Company, 2001.

[11] P. Nandakumar and K. Shankar, “Identification of structural parameters using transfer matrix andstate vectors in time domain,” in *Proc. 5th International Conference of Advances in Mechanical Engineering*. SVNIT, Surat, India, June 2011, pp. 175–179.

[12] J. J. Tuma. and F. Y. Cheng, *Theory and Problems of Dynamic structural analysis*. McGraw-Hill Book Company, 1982.

[13] P. Nandakumar and K. Shankar, “Particle swarm based structural identification using consistent mass transfer matrix method,” in *Proc.International Conference on Advances in control and Optimization of Dynamical Systems*. Indian Institute of Science, Bangalore, India, February 2012.

[14] S.Sandesh and K.Shankar, “Time domain identification of structural parameters and input time history,” *International Journal of Structural Stability and Dynamics*, vol. 9, no. 2, pp. 243–265, 2009.

[15] P. Prashanth and K. Shankar, “A hybrid neural network strategy for the identification of structural damage using time domain responses,” *IES Journal Part A: Civil and Structural Engineering*, vol. 1, no. 4, pp. 17–34, 2008.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Model-based Fault Detection and Isolation for Wind Turbine

Abdulhamed Hwas* and Reza Katebi**

*PhD student at University of Strathclyde, Glasgow, UK, email: Abdulhamed.hwas@eee.strath.ac.uk

** Industrial Control Centre, University of Strathclyde, Glasgow, UK, email: r.katebi@eee.strath.ac.uk

***Abstract*—** In this paper, a quantitative model based method is proposed for early fault detection and diagnosis of wind turbines. The method is based on designing an observer using a model of the system. The observer innovation signal is monitored to detect faults. For application to the wind turbines, a first principles nonlinear model with pitch angle and torque controllers is developed for simulation and then a simplified state space version of the model is derived for design. The fault detection system is designed and optimized to be most sensitive to system faults and least sensitive to system disturbances and noises. A multi-objective optimization method is then employed to solve this dual problem. Simulation results are presented to demonstrate the performance of the proposed method.

***Keywords***- *fault detection; observer; wind turbine; sensor monitoring*

## I. Introduction

Faulty components in wind turbine can cause high loses in energy production and possible damage to the wind turbines. The losses may be higher for offshore wind farms. This decreases the reliability and increases the cost of maintenance of the wind turbines. Figure 1 shows the percentage breakdown of the number of failures that occurred during the years 2000-2004 [1]. Most failures were linked to the electrical system followed by sensors and blades/pitch components.

Wind turbine fault monitoring that is a means to avoid abnormal event progression and reduces productivity loss, system breakdowns and damage. It increases safety and reliability of the system to achieve higher performance.

Fault diagnosis methods surveyed in literature can be classified into two general categories, quantitative and qualitative methods. In quantitative method, the understanding is expressed in terms of mathematical functional relationships between the inputs and outputs of the system in the form of system descriptions. In qualitative method, the relationships are expressed in terms of qualitative functions between different components of the system. This approach usually depends upon the knowledge from experts in both the normal and fault cases.

This paper proposes an observer-based Fault Detection and Isolation (FDI) method using a multi-objective optimisation procedure. The paper is organized as follows: Next section is concerned with the requirements for designing an observer and theoretical residual generation. Linear state space model of the wind turbine is presented in section III. Multi-objective optimization genetic algorithm is briefly described in section IV. Fault modelling, observer-based FDI and simulation results are demonstrated in section V. Finally, the conclusion is drawn in section VI.

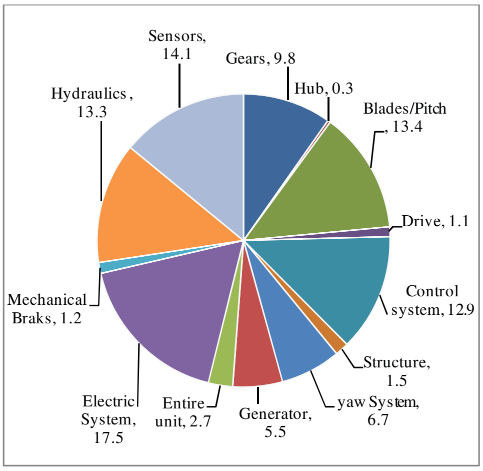


Figure 1 The distribution of a number of failures for Swedish wind turbines between 2000-2004 [1]

## II. Observer Based Fault Detection

### A. *Observer Design*

The system described by equation (1) is used to design an observer. The mathematical description of the observer is the same as the system except that the observer has an additional term, the gain K(t), which is continuously correcting the system output and improves the state estimates. The observer is defined in equations (2).

$$\begin{aligned}\dot{x}(t)&=Ax(t)+Bu(t)\\ y(t)&=Cx(t)+Du(t)\end{aligned}\tag{1}$$

$$\begin{aligned}\dot{\hat{x}}(t)&=A\hat{x}(t)+Bu(t)+K(t)[y(t)-\hat{y}(t)]\\ \hat{y}(t)&=C\hat{x}(t)+Du(t)\end{aligned}\tag{2}$$

where x, u, y are the state, input and output of the system of dimension n, m and r and A, B, C and D are system matrices of appropriate dimensions. Define the difference between $x(t)$ and $\hat{x}(t)$ as the state error vector (residual), e(t), thus the dynamic error can be written as:

$$\dot{e}=\dot{x}(t)-\dot{\hat{x}}(t)=(A-KC)e(t)\tag{3}$$

For fault monitroing purposes, a weighted residual is defined as follows:

$$r(t)=QCe(t) \tag{4}$$

where Q is residual weighting matrix.

Equation (3) illustrates the dynamic behaviour of the innovation siganl and this is governed by the eigenvalues of the matrix (*A-KC)*. If the matrix A-KC is stable, the error will tend to zero or a constant. If the eigenvalues are chosen in such a way that the dynamic behaviour of the error is asymptotically stable and adequately fast, then any error will tend to zero with sufficient speed. This is possible by choosing an appropriate value for K to achieve the stability, when the system is completely observable [2].

### *B. Residual Generations*

Assume the system is fully observable. The system dynamics with faults and disturbance modelscan be written as:

$$\begin{aligned} \dot{x}(t)&=Ax(t)+Bu(t)+R_1 f(t)+d(t) \\ y(t)&=Cx(t)+Du(t)+R_2 f(t) \end{aligned} \tag{5}$$

where *f(t)* represents the fault vector and considered to be an unknown function of time. The vector *d(t)* is the disturbance vector. The matrices $R_1$ and $R_2$ are the fault distribution matrices.

Using equation (2) and (5), the estimation error and the residual can be written as:

$$\dot{e}=(A-KC)e(t)+d(t)+R_1 f(t)-KR_2 f(t) \tag{6}$$

$$r(t)=Q\ [Ce(t)+R_2 f(t)] \tag{7}$$

Taking Laplace transform of Equation (7) gives:

$$\begin{aligned} r(s)&=Q[R_2+C(sI-A+KC)^{-1}(R_1-KR_2)]\mathrm{f(s)} \\ &+QC(sI-A+KC)^{-1}[d(s)+e(0)] \\ &=G_{rf}(s,K,Q)f(s)+G_{rd}(s,K,Q)[\mathrm{d(s)+e(0)}] \end{aligned} \tag{8}$$

where e(0) is the initial value of the state estimation error.

The effect of the faults on the signal r(t) can be maximized using the following performance index in frequency domain [3]:

$$\begin{aligned} J_f(K,Q)= \sup_{\omega\in[\omega_1,\omega_2]} \bar{\sigma}\{[QR_2+QC(j\omega I-A \\ +KC)^{-1}(R_1-KR_2)]^{-1}\} \end{aligned} \tag{9}$$

where $\bar{\sigma}\{.\}$ denotes the maximal singular value.

The disturbance effects on the residual can be minimised using the following performance index:

$$J_d(K)=\|(A-KC)^{-1}\| \tag{10}$$

Sensor, actuator and component fault matrices can be represented as:

$$\begin{aligned} R_1&=\left|\begin{array}{l|l} 0 & senseor\ fault \\ B & actuator\ fault \\ I & component\ fault \end{array}\right. \\ R_2&=\left|\begin{array}{l|l} I & senseor\ fault \\ D & actuator\ fault \\ 0 & component\ fault \end{array}\right. \end{aligned} \tag{11}$$

Thus we can rewrite fault indices

$$\begin{aligned} J_{af}(K,Q)= \sup_{\omega\in[\omega_1,\omega_2]} \bar{\sigma}\{-[Q_2D+QC(j\omega I \\ -A+KC)^{-1}(B-KD)]\ \} \end{aligned} \tag{12}$$

$$\begin{aligned} J_{sf}(K,Q)= \sup_{\omega\in[\omega_1,\omega_2]} \bar{\sigma}\{[QI+QC(j\omega I-A \\ +KC)^{-1}(-KI)]^{-1}\} \end{aligned} \tag{13}$$

$$\begin{aligned} J_{cf}(K,Q)= \sup_{\omega\in[\omega_1,\omega_2]} \bar{\sigma}\{-[QC(j\omega I-A \\ +KC)^{-1})]\} \end{aligned} \tag{14}$$

The problem can now be stated as minimising the criteria in (12), (13) and (14) subject to the system dynamics in (5). This is a multi-objective optimisation problem and hence genetic algorithm is proposed to solve the problem.

## III. State Space Model of the wind turbine

Mathematical models for the main components of nonlinear 5MW wind turbine system, particularly aerodynamic, two-mass drive train, DFIG generator and their controllers are developed and validated in previous work [4]. The linear state space matrices for a 5 MW wind turbine defined at wind speed 10 m/s is are follows :

Table I description of parameters of the wind turbine [4]

| DESCRIPTION | SYMBOL | DESCRIPTION | SYMBOL |
|---|---|---|---|
| *Turbine inertia* | $J_T$ | *Leakage coefficient* | $\sigma$ |
| *Gearbox ratio* | $n_g$ | *Stator current* | $i_d$, $i_q$ |
| *Generator inertia* | $J_G$ | *Pitch angle* | $\beta$ |
| *Torsional stiffness* | $K_s$ | *Desired pitch angle* | $\beta_d$ |
| *Torsional damping* | $C_s$ | *Mechanical torque* | $T_{wt}$ |
| *Synchronous speed* | $\omega_s$ | *Electrical torque* | $T_e$ |
| *Stator resistance* | $R_s$ | *Control torque* | $T_e^c$ |
| *Rotor resistance* | $R_r$ | *Control rotor voltages* | $v_{dr}$,$v_{qr}$ |
| *Stator inductance* | $L_s$ | *Wind turbine speed* | $\omega_{wt}$ |
| *Rotor inductance* | $L_s$ | *Generator speed* | $\omega_m$ |
| *Mutual inductance* | *Lm* | *Stator voltage* | $v_s$ |
|  |  | *Gearbox ratio* | $n_g$ |

$$A=\begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & \frac{-1}{n_g} & 0 & 0 \\ 0 & -\frac{K_S}{J_T} & \frac{-C_S}{J_T} & \frac{C_S}{J_T n_g} & 0 & 0 \\ 0 & \frac{K_S}{J_G n_g} & \frac{C_S}{J_G n_g} & \frac{-C_S}{J_G n_g^2} & 0 & 0 \\ 0 & 0 & 0 & -i_q & \frac{-R_S}{\sigma Lr} & (\omega_S-\omega_m) \\ 0 & 0 & 0 & \frac{i_d+L_m u_{sq}}{L_s \omega_s} & -(\omega_S-\omega_m) & \frac{-R_r}{\sigma L_r} \end{bmatrix} \quad (15)$$

$$B=\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{J_T} & 0 & 0 & 0 \\ 0 & 0 & \frac{-1}{J_G} & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (16)$$

$$C=\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{-\sqrt{3} n_p L_m V_s K_c}{\sigma L_r L_s \omega_s} \end{bmatrix} \quad (17)$$

where $\dot{\theta}=\omega_{wt}-\frac{\omega_m}{n_g}$ and $K_c$=0.8383. D is zeros matrix.

The states x, inputs u and outputs y are defined as:

$$x=[\beta,\theta_K,\omega_{wt},\omega_m,i_{dr},i_{qr}]^T,$$

$$y=[\beta,\omega_{wt},\omega_m,T_e],$$

$$u=[\beta_d,T_{wt},T_e^c,v_{dr},v_{qr}]^T$$

The physical variables and parameters are described in table I.

## IV. Multi-objective Optimization via Multi Objective Genetic Algorithm Approach

Multi Objective Genetic Algorithm (MOGA) is used to minimise the objective functions. It is more suitable than other approaches such as genetic algorithm because an equality constraint is not required to apply with MOGA [5]. The method finds the solution of problems with two or more objectives to be satisfied all together. Often, such objectives are in conflict with each other, and are expressed in different units. Because of their nature, multi-objective optimization problems normally have not one but a set of solutions, which are called Pareto points or pareto optimal solutions [8].

In order to design the observer, MOGA is used to solve the multi-objective optimization problem defined as follows. Here we need to minimise two objective functions Jsf(K,Q) and Jd(K), with *nxm* decision variable. Mathematically, the problem can be written as:

*Define:*

F(X) = [$F_1$(X); $F_2$(X)]

where $F_1$(X)= 1/$J_{sf}$(K,Q) and $F_2$(x)=$J_d$(K), given that max $J_{sf}$(K,Q)= min (1/$J_{sf}$(K,Q) and X={*x1,......, $x_{nxm}$*} is a vector of decision variables.

The problem is to minimise F(X) subject to:

$F_1^j$(X) ≤ 0 and $F_2^k$(X) ≤ 0.

$F_1^j$ (X): *j*th inequality constraint evaluated at X

$F_2^k$(X) : *k*th equality constraint evaluated at X

In the vector function F(x), some of the objectives may be in conflict with others, and some have to be minimized while others are maximized. The constraints define the feasible region X, and any point *x*∈X is a feasible solution. There is rarely a situation in which all F(X) have an optimum in X at a common point. Therefore, in the absence of preference information, solutions to multi-objective problems are compared using the notion of Pareto dominance.

Without loss of generality, in a minimization problem for all objectives, a solution $X_1$ dominates a solution, $X_2$ if the two following conditions are true:

- $X_1$ is no worse than $X_2$ in all objectives, i.e., $f_i(X_1) \leq f_i(X_2)$
- $X_1$ is strictly better than $X_2$ for at least one objective, i.e., $f_i(X_1) < f_i(X_2)$.

Then, a solution is said to be Pareto-optimal if it is not dominated by any other possible solution, as described above. Thus, the Pareto-optimal solutions to a multi-objective optimization problem form the Pareto front or Pareto-optimal set [9].

The performance indices $J_{sf}$(K,Q) and $J_d$(K) are functions in K and Q. Therefore, the parameters set to be designed are the observer gain matrix and residual weighting factor matrix. The matrix K must achieve the stability of the observer and optimisation of the performance indices. Ackermann's formula is used to parameterize the matrix K [2].

$$K=\left[A^n+\alpha_1 A^{n-1}+\alpha_2 A^{n-2}+\cdots+\alpha_{n-1}A+\alpha_n I\right]\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix}\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \quad (18)$$

where desired eigenvalues are defined as:

$$(s-p_1)(s-p_2)\cdots(s-p_n)=s^n+\alpha_1 s^{n-1}+\alpha_2 s^{n-2}+\cdots+\alpha_{n-1}s+\alpha_n$$

To improve the design desired eigenvalues are assigned in predefined regions to meet stability and response requirements as in equations[10]:

$$p_i = L_i + (U_i - L_i)\sin^2(x_i),\ i=1,...,n$$

$L_i \leq p_i \leq U_i$

where

$U_i$=[-6 -10 -1 -3 -8.5 -14]; $L_i$=[-8 -12 -2 -4 -9.5 -20];

where $U_i$ and $L_i$ are the lower and higher limits for the eigenvalues respectively. Index $x_i$ can be freely selected. Constrained performance indices have now been transformed into unconstrained as a function of X, where K($x_1$, $x_2$, $x_3$, $x_4$, $x_5$, $x_6$) and Q=[$x_7$ $x_{11}$ $x_{15}$ $x_{19}$; $x_8$ $x_{12}$ $x_{16}$ $x_{20}$; $x_9$ $x_{13}$ $x_{17}$ $x_{21}$; $x_{10}$ $x_{14}$ $x_{18}$ $x_{22}$].

The Matlab Genetic Algorithm Toolbox was utilised. The tuning parameters were set as follows: Initial range of variables= [0; 10000], population size: 75, number of generations: 200.

MOGA solver can accept one or more plot functions through the options argument. This feature is useful for visualizing the performance of the solver at run time. Figure 2(b) is the Pareto front, which plots the Pareto front at every generation. Figure 2(a) is the score diversity for each objective.

From Figure 2, there is only one optimal value on the Pareto front figure, that gives J$_{sf}$(K,Q)=infinity and J$_d$(K) = 258. The eigenvalues at these indices P and matrix K, residual weighting factor matrix Q are:

$$K=\begin{bmatrix} 0.012 & -7.6e\text{-}06 & 3.7e\text{-}08 & 7.3e\text{-}10 \\ 0.0001 & -2. & -0.017 & -1.8e\text{-}14 \\ -0.0003 & 15.26 & -0.016 & 3.6e\text{-}14 \\ 0.182 & -4338.1 & 17.35 & -2.2e\text{-}11 \\ -2.2e\text{-}09 & 1.2e\text{-}15 & -672 & -0.40 \\ -2.2e\text{-}09 & 1.4e\text{-}15 & 11.03 & -7.32 \end{bmatrix}$$

$$Q=\begin{bmatrix} 4426 & 4495 & 7814 & 5446 \\ 3015 & 4094 & 6897 & 1503 \\ 3836 & 3872 & 4308 & 8549 \\ 5144 & 6240 & 6211 & 5673 \end{bmatrix}$$

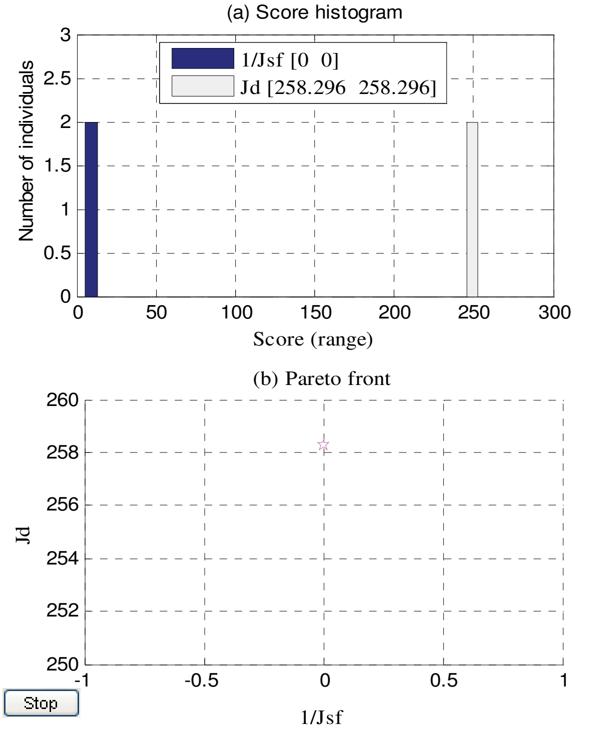


Fig. 2 Pareto front and the number of individuals for sensor and disturbance performances indices.

## V. Observer-based sensor FDI scheme

A successful FDI should be accompanied by a fault isolation procedure to isolate a particular fault from others. For example, to determine in which sensor, actuator or component the fault happened. Observer-based residual generator approach is suitable for detecting a fault [6], but to isolate the fault, a method is proposed here as follows..

### A. Fault Model

Faults are modelled as unknown change in signals as an additive fault. This fault can be classified according to their source as an actuator $\Delta u(t)$, sensor $\Delta y(t)$ or component $\Delta u_c(t)$ faults [3]. Figure 3 shows the effect of additive faults on the observed signals of the inputs and outputs. The component faults affect both the true output and the observed output ( $y_o^c$ ). The observed signals for the input and output can be rewritten as below.

$$u_o(t)=u(t)+\Delta u(t)+\delta u(t) \tag{19}$$

$$y_o^c(t)=u_o(t)+\Delta u_c(t)+\delta u_c(t) \tag{20}$$

$$y_o(t)=y_0^c(t)+\Delta y(t)+\delta y(t) \tag{21}$$

where $u_0(t)$, $y_0^c(t)$ and $y_o(t)$ are the actuator, component and sensor outputs respectively.

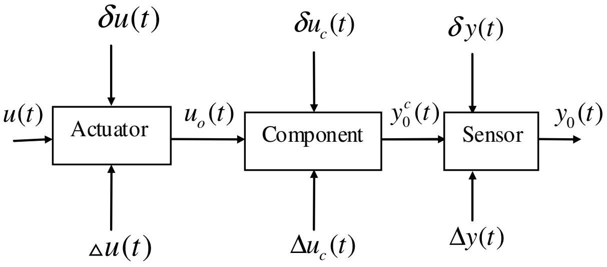


Fig. 3 A diagram for the additive faults, where $\delta u(t)$ and $\delta u_c(t)$ denote the actuator and component disturbance signals respectively. $\delta y(t)$ is a sensor noise signal.

### B. Design of Observer-based Sensor FDI

To design a robust observer-based sensor FDI, we assumed that only one sensor fault occurs, all actuators and components are fault free. Then from equations (5) and (11) the system equation can be expressed as:

$$\begin{aligned} \dot{x}(t)&=Ax(t)+Bu(t)+d(t) \\ y(t)&=Cx(t)+R_2 f(t) \end{aligned} \tag{22}$$

Then the residual generator can be created for each sensor as:

$$r_k(t)=Q[(Ce(t)+R_2 f(t))(C_m-C_k)] \tag{23}$$

where $k$ is the number of the measurement sensor, $C_k$ is obtained from the matrix $C$ by deleting zero columns and assuming $k_{\text{th}}$ row equal zero. $C_m$ is the matrix $C$ without zero columns.

From equation (23), it is obvious that each residual generator is driven making all other residuals equal zero. From above a robust and observer based sensor FDI schemes is designed as shown in figure 4. Each sensor residual ($r_k$) is separated from the output of the residual (r) by $r\ x\,(C_m - C_k)$

, and then dimension of $r_k$ is modified using $Q_k$. The advantage of this approach is that it uses only one observer comparing with other approaches, which use a bank of observers such as a structured residual set designed by a dedicated or a generalized observer scheme [6].

For simulation, the values of the residual weighting factors are selected as follows:

$$\frac{Q_1}{100}=\frac{Q_2}{1000}=Q_3=10\ \ Q_4=Q \qquad (24)$$

Q are then obtained using the Method of Multi Objective Genetic Algorithm. Q1, Q2, Q3 and Q4 are residual weighting factors for pitch angle, rotor speed, generator speed and torque sensors, respectively

### C. Simulation Results

Simulation Results are shown in Figs. 5 to 8. The faults are applied by multiplying the sensor signal by 1.05, i.e. an increase of 5% at 200s, and the transient period is neglected. Fig. 5 shows the residual norm of the pitch sensor increased, when there is a fault. Fig. 6 is for a fault occurred in rotor speed sensor; the residual norm of the rotor and generator rotational speed sensors detects the faults. Fig. 7 for a fault occurred in generator speed sensor and Fig. 8 for a fault occurred in the generator torque sensor. The speed of the fault detection is very fast. Consequently, from figures 5, 6, 7 and 8, we can construct the Boolean decision table as shown in Table II. If a fault occurs, we can compare the results with this fault signature table and decide the location of the fault as shown in the Fig. 4. Therefore, fault detection and isolation are achieved.

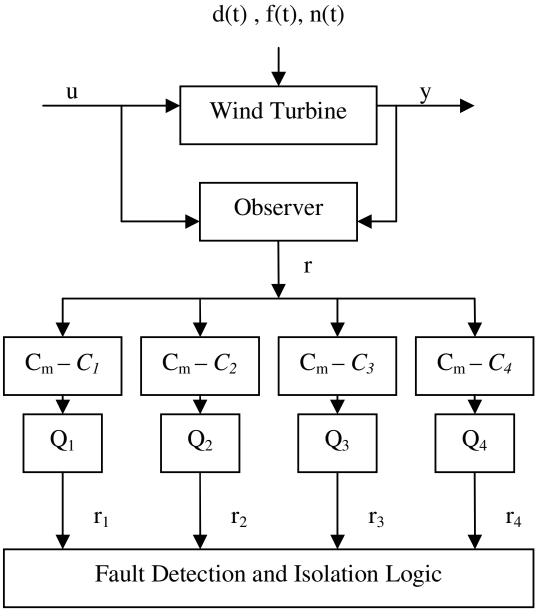


Figure 4 A robust sensor FDI scheme. $r_1$, $r_2$, $r_3$ and $r_4$ represent pitch angle, rotor speed, generator speed and generator torque residuals respectively. r contains residuals for all sensors.

For a specific fault, we can now find its location using the following thresholds:

$$\begin{aligned} r_k(t) &> \mid T_k \mid \Rightarrow f_k(t)=1 \\ r_k(t) &\leq \mid T_k \mid \Rightarrow f_k(t)=0 \end{aligned} \qquad (25)$$

where $T_k$ is the threshold. $f_k(t)$ sensor fault. $k$=1,2 ,3 ,4 (for pitch angle, rotor speed, generator speed and torque sensors).

Table II Boolean decision for sensor faults

| **Residual** | **Pitch residual** | **Rotor speed residual** | **Generator speed residual** | **Electrical torque residual** |
|---|---|---|---|---|
| ***Pitch fault*** | *1* | *0* | *0* | *0* |
| ***Rotor speed fault*** | *0* | *1* | *1* | *0* |
| ***Generator speed fault*** | *0* | *1* | *1* | *1* |
| ***Electrical torque fault*** | *0* | *0* | *0* | *1* |

## VI. Conclusions

A robust observer-based sensor fault detection and isolation scheme is developed for wind turbine. This scheme is systematic and easy to design and implement. Simulation results proved that it is very suitable for detection and isolation faults in sensors and simple to handle multiple faults.

The advantage of the proposed approach is that it depends on only one observer comparing with other approaches, which use a bank of observers such as a structured residual set designed by a dedicated or a generalized observer scheme.

## References

[1] Johan, R. and B. Lina Margareta, B. Survey of Failures in Wind Power Systems With Focus on Swedish Wind Power Plants During 1997& ndash;2005. Energy Conversion, IEEE Transactions on, 2007. 22(1): p. 167-173.

[2] Ogata, K., (2002). Modern Control Engineering. Pearson Education International, fourth edition.

[3] Chiang, L., Russell, E., and Braatz, R., (2001). Fault Detection and Diagnosis in Industrial Systems. Advanced Textbooks in Control and Signal Processing, Springer.

[4] Hwas, A. (2010). Wind energy conversion system model. Industrial Control Centre, Technical progress report first year, University of Strathclyde.

[5] Hur, Sung-ho and Katebi, M.R. and Taylor, A., (2010). Fault detection and diagnosis of a plastic film extrusion process. In: UKACC International Conference on CONTROL 2010, 2010-09-07 - 2010-09-10, Coventry, UK.

[6] Chen, J. and Patton, R.J., (1999). Robust model-based fault diagnosis for dynamic systems. Kluwer Academic Publishers.

[7] Gertler, J., (1991). Analytical redundancy methods in fault detection and isolation; survey and synthesis, IFAC Fault Detection, Supervision and Safety for Technical Processes, Baden-Baden, Germany, pp. 9–21.

[8] V. Chankong and Haimes, Y. Multi-objective decision making theory and methodology. New York: North-Holland, 1983.

[9] Zeleny, M. Multiple Criteria Decision Making, ser. Quantitative Methods for Management. : McGraw-Hill, 1982

[10] Burrows, S. P. and Patton, R. J. Design of a low sensitivity, minimum norm and structurally constrained control law using eigenstructure assignment. Optimal Control Appl. and Meth., 1991, 12, 131-140.

[11] Chen, W., Khan, A., Abid, M. and Ding, S. Integrated design of observer based fault detection for a class of uncertain nonlinear systems. - Int. J. Applied Mathematics and Computer. Science, 2011, vol.21, no.3

[12] Korbicz, J., Maquin, D. and Theiliol, D. Adances in Control and Fault-Tolerant Systems. Special issue of Int. J. Applied Mathematics and Computer. Science, 2012, vol.22, no.1.

## APPENDIX A: SIMULATION RESULTS OF THE ROBUST SENSOR FDI SCHEME.

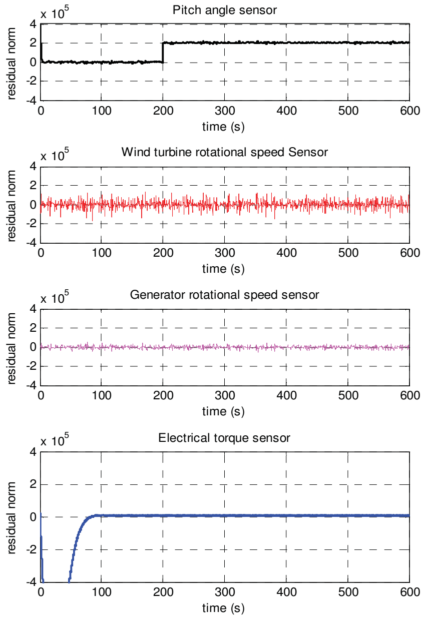


Fig. 5 Residual norm when a fault occurs in the Pitch angle sensor.

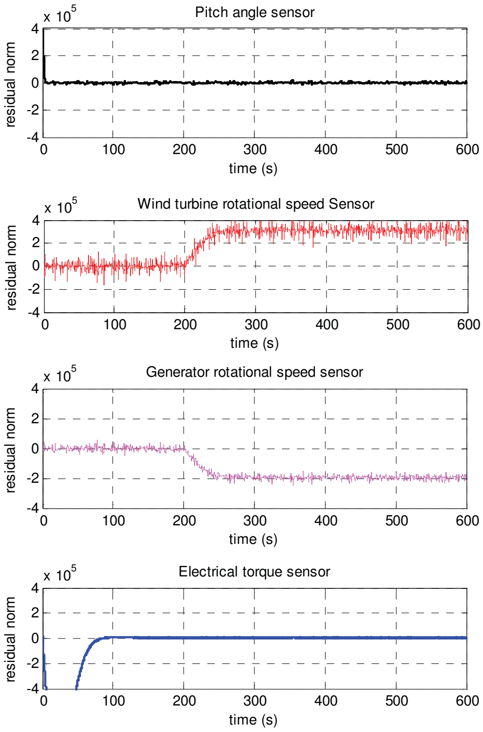


Fig. 6 Residual norm, when a fault occurs in rotor speed sensor

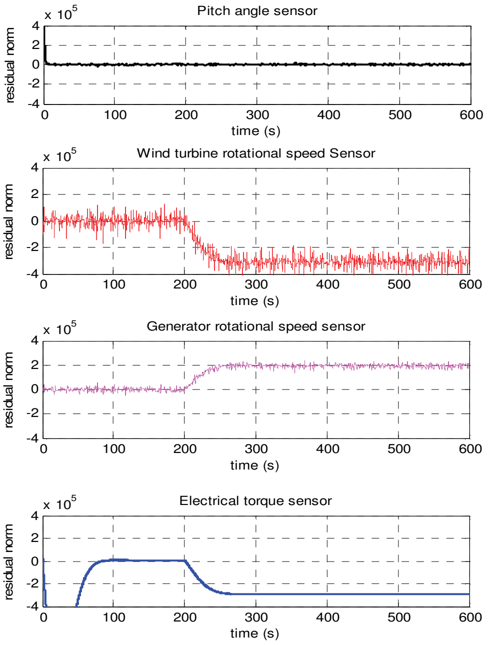


Fig. 7 Residual norm, when a fault occurs in generator speed sensor.

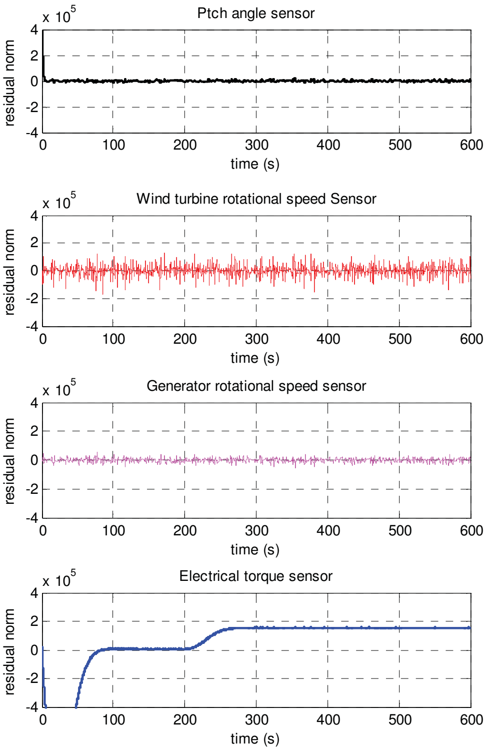


Fig. 8 Residual norm, when a fault occurs in the generator torque sensor.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Robust MPC algorithms using alternative parameterisations

Bilal Khan
Automatic Control and Systems Engineering
The University of Sheffield,UK
Email: b.khan@sheffield.ac.uk

John Anthony Rossiter
Automatic Control and Systems Engineering
The University of Sheffield,UK
Email: j.a.rossiter@sheffield.ac.uk

***Abstract*—This paper demonstrates the efficacy of alternative parameterisations of the degrees of freedom within a robust MPC algorithm. Alternative parameterisations have been shown to improve the feasible region when the number of degrees of freedom is limited for the nominal case. This paper extends that work to the robust scenario and shows that similar benefits accrue and moreover, the increase in complexity for the robust case as compared to the nominal case is much less than might be expected. The improvements, with respect to a conventional algorithm, are demonstrated by numerical examples.**

**Keywords: Robust MPC, Alternative parameterisations, Feasibility.**

## I. Introduction

Model Based Predictive Control (MPC) or receding horizon control (RHC) or embedded optimisation or moving horizon or predictive control [1], [2], [3], are general names for different computer control algorithms that use predictions as a basis for forming a control law. MPC has now reached a high level of maturity in academia and is widely used in industry. There is substantial interest in how to develop algorithms for the stochastic case and to deal with nonlinearity or uncertainty, in particular because formal consideration of these issues can lead to substantial computation and/or complexity. The main aim of this paper is to contribute to research which improves feasibility while tackling the robust case, perhaps at some small loss of optimality. Specifically, the focus is on the potential of alternative parameterisations of the degrees of freedom (e.g. [20]) to enable enlargement of feasible regions in the uncertain case without too much detriment to performance, optimality and the computational burden.

Two common approaches to robustness have been considered in MPC literature. The first makes use of the inherent robustness of nominal MPC algorithm design, i.e. MPC algorithm that is not specially designed for robust aspects (like stability and performance) [5], [6]. The second approach is the explicit inclusion of robustness requirements into the design of MPC algorithm and this has received considerable attention in the MPC literature. An MPC algorithm design based on including the uncertainty information in the model will be referred to as robust MPC [7], [8].

Traditionally robust MPC requires the solution of min-max optimisation problem, where optimisation is performed in order to minimize a worst-case cost over all possible uncertainty realisations [9]. Furthermore, this also guaranteed constraint satisfaction for all possible future trajectories. For a min-max optimisation problem, stability was proved in [10] by considering feedback within the horizon. In general, solving a min-max problem subject to constraints has to optimise over a sequence of control strategies rather than a sequence of fixed control moves, as the latter is computationally too demanding for practical implementation [11]. All these aspects contribute to make several variants of robust MPC intractable for on-line optimisation [7], [11].

However, various alternative robust MPC algorithms have been proposed to approximate solutions of the max-min problem but with a reduced computational burden. In [12] a classical result is presented by directly incorporating the plant uncertainty into the MPC formulation. The existence of a feedback law minimising an upper bound on the infinite horizon objective function and satisfying constraints is reduced to a convex optimisation using linear matrix inequalities (LMI). The main disadvantages are the use of LMI-based optimisation that can still be computationally demanding and moreover the methods use conservative constrained handling. Several authors reduced the online computational complexity [11], [13], [14] by performing more offline analysis using invariant ellipsoidal sets, however with a significant restriction to the volume of the feasible region. A robust triple mode MPC algorithm was proposed in [15] by introducing an additional mode in dual mode MPC with a large feasible region and good performance; in essence, the objective is to find suitable and fixed linear time varying (LTV) control law.

Of the numerous robust MPC algorithms, quite a few incorporate the notion of feedback in the prediction sequence over which the on-line optimisations take place as this reduces the divergence of the predictions which occurs in open-loop. This closed-loop paradigm improves the control performance but does not necessarily lead to an inexpensive optimisation strategy [7], [8], [9]. Reduced-complexity invariant sets are introduced in [7] for the case of quasi-infinite horizon closed-loop MPC. The reduced-complexity invariant sets may result in a decrease in the number of on-line optimisation variables [7]. This invariant set structure is used in the design of robust MPC and this paper pursues this type of approach to including uncertainty information in the model ([8], [16]).

In the nominal case, Laguerre, Kautz and generalised parameterisations are able to achieve large feasible regions while

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

maintaining local optimality and a relatively low computational complexity [17], [21], [18], [19], [20]. This paper extends the earlier studies in [18], [20] to the case of parameter uncertainty by developing the algorithm of [7] for constructing polyhedral robust positive invariant sets; this enables the online robust MPC algorithm to be based on a standard quadratic programme while adding the benefits of improved feasibility due to the change in parameterisation.

Section II gives the necessary background about nominal MPC, generalised parameterisations for an optimal MPC and robust MPC. Section III discusses alternative parameterisations within Robust MPC using a generalised parameterisation. An algorithm is proposed for Robust MPC using the generalised parameterisation. Comparisons between the existing Robust MPC (RMPC) and the new proposed algorithms are given in section IV using numerical examples. Finally conclusions and future work are in section V.

## II. Problem formulation for Robust MPC

Assume discrete time linear parameter varying (LPV) state space models of the form [8]

$$\begin{aligned} x_{k+1} = A(k)x_k + B(k)u_k, \\ (A(k),B(k)) \in Co\{[A_1B_1],...,[A_mB_m]\}, \end{aligned} \tag{1}$$

with the nominal model being $(A,B)$ and $x_k \in \mathbb{R}^{n_x}$ and $u_k \in \mathbb{R}^{n_u}$ being the state vectors and the plant input respectively. Assume that the states and inputs at all time instants should fulfill the following constraints:

$$L_x x + L_u u \leq l. \tag{2}$$

### A. Nominal MPC algorithm

The performance index [3], [22] to be minimised, at each sample instant, with respect to $u_k, u_{k+1}, \dots$ is:

$$\begin{aligned} J = \sum_{i=0}^{\infty}(x_{k+i+1})^T Q(x_{k+i+1}) + (u_{k+i})^T R(u_{k+i}) \\ s.t. \left\{ \begin{array}{ll} x_{k+1} = Ax_k + Bu_k, (2) & \forall k \geq 0, \\ u_k = -Kx_k & \forall k \geq n_c, \end{array} \right. \end{aligned} \tag{3}$$

with $Q$ and $R$ positive definite state and input cost weighting matrices and where $K$ is the optimal feedback gain minimizing $J$ in the absence of constraints (2). Practical limitations imply that only a finite number, that is $n_c$, of free control moves can be used [3]. For these cases, $u_k = -Kx_k$ is implemented by ensuring that the state $x_{n_c}$ must be contained in a polytopic control invariant set (often denoted as the maximal admissible set or MAS):

$$\begin{aligned} \chi_0 = \{x_0 \in R^{n_x} : L_x x - L_u K x_k \leq l, \\ x_{k+1} = Ax_k + Bu_k, \forall k \geq 0\}. \end{aligned} \tag{4}$$

For simplicity of notation, the MAS is described in the form $S_0 = \{x : Mx \leq b\}$ for suitable $M$ and $b$ and the d.o.f. can be reformulated in terms of a new variable $c_k$:

$$\begin{aligned} u_k = -Kx_k + c_k, \qquad & k = 0,...,n_c-1, \\ u_k = -Kx_k, \qquad & k \geq n_c. \end{aligned} \tag{5}$$

The MCAS (maximal controlled admissible set) is given as

$$\chi_c = \{x_k : \exists C, Mx_k + NC \leq b\}, \tag{6}$$

where $C = [c_k^T, \dots, c_{k+n_c-1}^T]^T$ and hence the equivalent optimisation to (3) is:

$$\min_C \ J_c = C^T S C \ s.t. \ Mx_k + NC \leq b. \tag{7}$$

The optimal MPC (OMPC) algorithm is given by solving the QP optimisation (7) at every sampling instant then implementing the first component of $C$, that is $c_k$ in the control law of (5). When the unconstrained control law is not predicted to violate constraints (i.e. $x_k \in \chi_0$), the optimising $C$ is zero so the control law is $u_k = -Kx_k$. The optimisation of (7) can require a large $n_c$ d.o.f. to obtain both good performance and a large feasible region.

### B. Generalised parameterisation for Optimal MPC

More recently different alternative parameterisation techniques have been developed to improve the feasible region in nominal case. Laguerre, Kautz and generalised parameterisations have been proposed in [17], [18], [20] as effective alternatives to the standard basis set for parameterising the d.o.f. within an optimal MPC. This section will discuss briefly Laguerre, Kautz and generalised optimal MPC.

The generalised parameterisation [20] is defined using a higher order discrete network such as:

$$\begin{aligned} \mathcal{G}_i(z) = \mathcal{G}_{i-1}(z)\frac{(z^{-1}-a_1)\dots(z^{-1}-a_n)}{(1-a_1z^{-1})\dots(1-a_nz^{-1})}, \\ 0 \leq a_k < 1, \qquad k = 1\dots n. \end{aligned} \tag{8}$$

With $\mathcal{G}_1(z) = \frac{\sqrt{(1-a_1^2)\dots(1-a_n^2)}}{(1-a_1z^{-1})\dots(1-a_nz^{-1})}$. The generalised function with $a_k, \forall k = 1,\dots,n$ gives [20]

$$\begin{aligned} Laguerre\ network: \qquad & \mathcal{G}_i = \mathcal{L}_i, \quad if \ \ a_k = [a]. \\ Kautz\ network: \qquad & \mathcal{G}_i = \mathcal{K}_i, \quad if \ \ a_k = [a,b]. \end{aligned} \tag{9}$$

The generalised sequence can be computed using the following state-space model (this example is 3rd order):

$$\mathcal{G}_{k+1} = \underbrace{\begin{bmatrix} b & 0 & 0 & \dots \\ b & c & 0 & \dots \\ -ab & (1-ac) & a & \dots \\ ab^2 & -b(1-ac) & (1-ac) & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}}_{A_G} \mathcal{G}_k, \tag{10}$$

$$\begin{aligned} \mathcal{G}(0) = \gamma[1,1,1,-a,ab,-abc,a^2bc,\dots]^T, \\ \gamma = \sqrt{(1-a^2)(1-b^2)(1-c^2)}. \end{aligned}$$

The prediction using d.o.f. and generalised functions [20] are given by

$$C = \begin{pmatrix} c_k \\ c_{k+1} \\ \vdots \end{pmatrix} = \begin{pmatrix} \mathcal{G}(0)^T \\ \mathcal{G}(1)^T \\ \vdots \end{pmatrix} \rho = H_G \rho, \tag{11}$$

where $\rho$ is the $n_G$ dimension decision variable when one uses the first $n_G$ column of $H_G$. The only difference between Laguerre OMPC and Kautz or generalised OMPC is that of $H_G$ matrix. For further details readers are referred to [18] and [20].

*Algorithm 2.1:* GOMPC

$$\rho^* = arg\ \min_{\rho}\ J_G = \rho^T[\sum_{i=0}^{\infty} A_G^i \mathcal{G}(0) S \mathcal{G}(0)^T (A_G^i)^T]\rho$$
$$s.t.\ Mx_k + NH_G\ \rho\ \leq\ b. \qquad (12)$$

Define $c_k^* = H_G \rho_k^*$ and implement $u_k = -Kx_k + e_1^T c_k^*$. Where $e_1^T = \begin{bmatrix} I, 0, \dots, 0 \end{bmatrix}$.

*Remark 2.1:* It is straightforward to show, with conventional arguments, that all algorithms (i.e. LOMPC, KOMPC, GOMPC) using terminal constraints within MPC problem formulation provides recursive feasibility and Lyapunov stability.

### *C. Robust MPC (RMPC): Dual mode MPC for LPV case*

The robust MPC algorithm, here denoted RMPC, developed in [8] is very similar to the nominal optimal MPC algorithm in (7), but it is applicable to the LPV system in (1). RMPC is designed to minimise the nominal predicted performance cost subject to robust constraint satisfaction by the whole class of possible predictions.

Define an augmented system model which incorporates the 'd.o.f.' as follows:

$$X_{k+1} = \Psi_i X_k,\ \Psi_i = \begin{bmatrix} \Phi_i & B_i E \\ 0 & I_L \end{bmatrix},\ X_k = \begin{bmatrix} x_k \\ c_k \end{bmatrix}. \qquad (13)$$

with $\Phi_i = A_i - B_i K$, $E = \begin{bmatrix} I, 0, \dots, 0 \end{bmatrix}$ and $I_L$ as shift matrix. The associated constraints (2) are represented as:

$$\alpha X \leq \delta,\ \alpha = \begin{bmatrix} L_x & 0 \\ -L_u K & L_u E \end{bmatrix},\ \delta = l. \qquad (14)$$

RMPC algorithm differs from OMPC only in the inequalities that need to be satisfied. It has been shown [8] that RMPC has recursive feasibility and guaranteed stability, under the mild condition that the uncertainty is small enough to allow such an invariant set to exist for all the different dynamics $\Psi_i$. The robust invariant set MAS for LPV system subject to linear constraints is introduced in [7]. The MAS is calculated using an invariant condition which avoids the combinatorial explosion of the number of constraints. The MAS for robust MPC $\chi_{r_0}$ in [8] is defined in compact form as:

$$\chi_{r_0} = \{x_k : M_r x_k \leq br\}. \qquad (15)$$

The associated MCAS in [8] is defined in compact form as:

$$\chi_r = \{x_k : \exists C, M_r x_k + N_r C \leq br\}. \qquad (16)$$

The key success of RMPC is the definition of inequalities, i.e. $M_r, N_r$ and $b_r$; the precise details of how to compute these matrices are omitted and are available in [8]. The performance cost can be (one can make improvements to this but such discussions are beyond the remit of this paper) based on a nominal performance cost but it in particular is phrased in terms of perturbations $c_k$ to the nominal control law.

For ease of reference, we summarise RMPC below and the reader is referred to [8] for further details.

*Algorithm 2.2:* RMPC
At each sampling instant, solve the following optimisation problem:

$$\min_{C} J = C^T S C\ \ s.t.\ \ M_r x_k + N_r C \leq b_r,$$

where only the first block element of $C$ is implemented in the control law of (5).

## III. Using Alternative parameterisation within RMPC

A fundamental weakness of conventional MPC algorithms is poor feasibility with a small number of d.o.f. to move the initial states into the MAS. This weakness may be overcome by increasing the number of d.o.f., but this compromises the computational burden. However, the alternative parameterisations proposed in [17], [21], [18], [20] for the nominal case should significant feasibility improvements, for the same number of d.o.f.. Therefore, this section seeks to extend the use of alternative parameterisation to the robust case and thus explore whether similar feasibility benefits are possible or likely. This section will show how such parameterisation can be used to form robust invariant sets and thus deployed in appropriate robust MPC algorithm. Examples in the next section are used to demonstrate the impact on feasibility.

### *A. Alternative parameterisation within RMPC*

The alternative parameterisation (i.e. Laguerre, Kautz and generalised functions) may be used to improve the feasible region of RMPC. Robust generalised MPC (RGMPC) is a robust MPC algorithm where the d.o.f. or input perturbations $c_k$ are parameterised using generalised functions. As in the conventional case, the prediction cost can be represented in terms of the perturbations $\rho$ about the nominal control law, hence:

$$J_G = \rho^T \left[\textstyle\sum_{i=0}^{\infty} \mathcal{G}(i) S \mathcal{G}(i)^T\right] \rho, \qquad (17)$$

with $c_{k+i} = \mathcal{G}(i)^T \rho$ and $\mathcal{G}(i) = A_G \mathcal{G}(i-1)$.

From (14), we drive the autonomous formulation using generalised parameterisation as:

$$X_{k+1} = \Psi_i X_k,\ \Psi_i = \begin{bmatrix} \phi_i & B_i \mathcal{G}_0^T \\ 0 & A_G^T \end{bmatrix},\ X_k = \begin{bmatrix} x_k \\ \rho_k \end{bmatrix}. \qquad (18)$$

These dynamics should full fill the constraints (2),

$$\alpha X \leq \delta;\ \alpha = \begin{bmatrix} L_x & 0 \\ -L_u K & L_u \mathcal{G}_0^T \end{bmatrix},\ \delta = l. \qquad (19)$$

Robust constraint handling is represented by an MCAS or $\chi_r$ which is calculated offline with the the methodology of [7], [8], but deploying alternative functions, that is equations (18,19) within the update model; this is illustrated in the following algorithm.

*Algorithm 3.1:* RGMPC Given a LPV system (18) subject to linear constraints (19).

1) Set the initial values for $A_S$ and $b_S$ to

$$A_S := \alpha; \qquad b_S := l. \tag{20}$$

2) Initialise the index $i := 1$.
3) Repeat until $i$ is not strictly larger than number of rows in $A_S$.
    a) Select row $i$ from (20), check whether adding any of the constraints $\alpha_i \Psi_i X \leq l_i$, $j = 1, \ldots, m$ to $A_S, b_S$ would decrease the size of $\chi_r$, by solving the following linear programming (LP) for $j = 1, \ldots, m$

$$\begin{aligned} c_j = \max_X \quad & \alpha_i \Psi_i X - l_i \\ s.t. \quad & A_S X \leq b_S. \end{aligned} \tag{21}$$

If $c_j > 0$, then add the constraint to $A_S, b_S$ as follows:

$$A_S := \begin{bmatrix} A_S \\ \alpha_i \Psi_i \end{bmatrix}; \quad b_S := \begin{bmatrix} b_S \\ l_i \end{bmatrix}. \tag{22}$$

    b) Increment i.
4) End.

After calculating the inequalities for invariant set the following RGMPC algorithm can be defined.

*Algorithm 3.2:* RGMPC
At each sampling instant, solve the following optimisation problem:

$$\min_\rho J_G \quad s.t. \quad A_S X \leq b_S.$$

Define $c_k^* = H_G \rho_k^*$ and implement $u_k = -K x_k + e_1^T c_k^*$.

*Remark 3.1:* Algorithm 3.1 will terminate in finite steps and only adds constraints and never removes constraints. It is clear that the resulting $\chi_r = \{X : A_S X \leq b_S\}$ will satisfies a robust positive invariant set. Algorithm convergence and invariance of the resulting set is proved similarly as in [7]. After terminating, it is recommended to remove any redundant constraints.

*Remark 3.2:* MAS or $\chi_{r_0}$ is calculated using above algorithm with $[x, c] = [x, 0]$ or using algorithm defined in [7].

### B. Order selection of the generalised parameterisation dynamic

In generalised parameterisations with higher order prediction dynamics have more flexibility to improve feasible region with a limited number of d.o.f.. So there is a clear choice of selecting the order of the parameterisation dynamics.

The prediction dynamics for the $3rd$ order parameterisation from (10) (which can easily be extended to $nth$ order) is quite generic with distinct eigenvalues. From the autonomous formulation using generalised parameterisation in (18), to fulfill the algebraic relations the dimension of the $A_G$ in (10) must be the same as $n_c$. Moreover the key observation from the augmented model in (18) is that $dim(A_G) = n_c$ is an upper bound on maximum parameterisation dynamics order.

## IV. Numerical Examples

The section will illustrate the efficacy of the alternative parameterisation within robust MPC algorithm by several numerical examples given next. The aim is to compare two aspects: (i) the size of the feasible regions; (ii) the number of inequalities required to describe the robust MCAS. For the purposes of visualisation, figures are restricted to second order systems for which it is possible to plot regions of attraction. The comparisons are based on 4 systems with $x \in \mathbb{R}^2, x \in \mathbb{R}^3$ and $x \in \mathbb{R}^4$. The alternative parameterisations are based on Laguerre, Kautz and generalised functions with $3rd$ order dynamics.

### A. Example 1

$$\begin{aligned} A_1 &= \begin{bmatrix} 1 & 0.2 \\ 0 & 1 \end{bmatrix}, \ B_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \\ A_2 &= \begin{bmatrix} 1 & 0.1 \\ 0 & 1 \end{bmatrix}, \ B_2 = \begin{bmatrix} 0 \\ 1.5 \end{bmatrix}, \\ C &= \begin{bmatrix} 1 & 0 \end{bmatrix}, \ Q = C^T C, R = 0.5, \ n_c = 2, \ a = 0.5, \\ (a, b) &= (0.5, 0.58) \ and \ (a, b, c) = (0.65, 0.67, 0.64). \end{aligned} \tag{23}$$

### B. Example 2

$$\begin{aligned} A_1 &= \begin{bmatrix} 0.6 & -0.4 \\ 1 & 1.3 \end{bmatrix}, \ B_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \\ A_2 &= \begin{bmatrix} 0.7 & -0.5 \\ 1 & 1.4 \end{bmatrix}, \ B_2 = \begin{bmatrix} 1.5 \\ 0 \end{bmatrix}, \\ C &= \begin{bmatrix} 1 & 0 \end{bmatrix}, \ Q = C^T C, R = 0.5, \ n_c = 2, \ a = 0.2, \\ (a, b) &= (0.3, 0.35) \ and \ (a, b, c) = (0.41, 0.42, 0.43). \end{aligned} \tag{24}$$

### C. Example 3

$$\begin{aligned} A_1 &= \begin{bmatrix} 1 & 0.1 & 0.1 \\ 0 & 1 & 0.1 \\ 0 & 0 & 1 \end{bmatrix}, \ B_1 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \\ A_2 &= \begin{bmatrix} 1 & 0.2 & 0.2 \\ 0 & 1 & 0.2 \\ 0 & 0 & 1 \end{bmatrix}, \ B_2 = \begin{bmatrix} 0 \\ 0 \\ 1.5 \end{bmatrix}, \\ C &= \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}, \ Q = C^T C, R = 0.5, \ n_c = 2, \ a = 0.4, \\ (a, b) &= (0.55, 0.58) \ and \ (a, b, c) = (0.4, 0.6, 0.58). \end{aligned} \tag{25}$$

### D. Example 4

$$A_1 = \begin{bmatrix} 1 & 0.1 & 0.1 & 0.1 \\ 0 & 1 & 0.1 & 0.1 \\ 0 & 0 & 1 & 0.1 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \; B_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix},$$

$$A_2 = \begin{bmatrix} 1 & 0.2 & 0.2 & 0.2 \\ 0 & 1 & 0.2 & 0.2 \\ 0 & 0 & 1 & 0.2 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \; B_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1.5 \end{bmatrix}, \qquad (26)$$

$$C = \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}, \; Q = C^TC, R = 0.5, \; n_c = 2, \; a = 0.4, \; (a, b) = (0.5, 0.56) \; and \; (a, b, c) = (0.55, 0.57, 0.6).$$

with constraints

$$-1 \leq u_k \leq 1; \quad -10 \leq x_{i_k} \leq 10; \quad i = 1, \dots, 4. \qquad (27)$$

### E. Feasible region comparisons

The region of attraction for examples 1 and 2 are plotted in figure 1 and 2 respectively. It is clear from both figures that the use of alternative (Laguerre, Kautz and generalised (3rd order)) parameterisation techniques within robust MPC algorithms improve the volume of the feasible region or the volume of the maximum control admissible set (MCAS). Table I shows the volume comparisons for examples 1-4 using RMPC, RLMPC, RKMPC and RGMPC (3rd order) algorithms. Alternative parameterisation improves feasible region and thus, based solely on volume considerations and as expected, alternative parameterisation based algorithms to be preferred in robust scenario.

### F. Number of constraints

For completeness, it is important to compare the number of inequalities required to describe the robust MCAS as the complexity of these set descriptions has an impact on the online computational burden, the more inequalities the higher the computational burden in solving the associated QP optimisation (this paper does not discuss issues linked to the exploitation of structure and efficient QP optimisers). The number of inequalities to define $\chi_r$ is compared with same number of d.o.f. in Table II. The number of inequalities with parameterised based algorithms are slightly more in comparison with RMPC algorithms. It is clear that higher order parameterisation improves feasibility needs more d.o.f. which will compromise the complexity benefits.

### G. Summary

The result shown in Figure 1 and 2, it is clear that RGMPC ($3rd$ order) has a large feasible region as compare with other algorithms, but having $n_c = 3$. This is due to the structure of $3rd$ order prediction dynamics which compromises the computational burden. Moreover for Table I and II, an interesting observation is that RGMPC ($3rd$ order) improves feasible region with higher number of inequalities.

RKMPC improves feasible region with slightly higher inequalities as compare with both RLMPC and RMPC for the same number of d.o.f. (i.e. $n_c = 2$). It is also clear from both Table I and II that GKMPC is a better choice with $n_c = 2$.

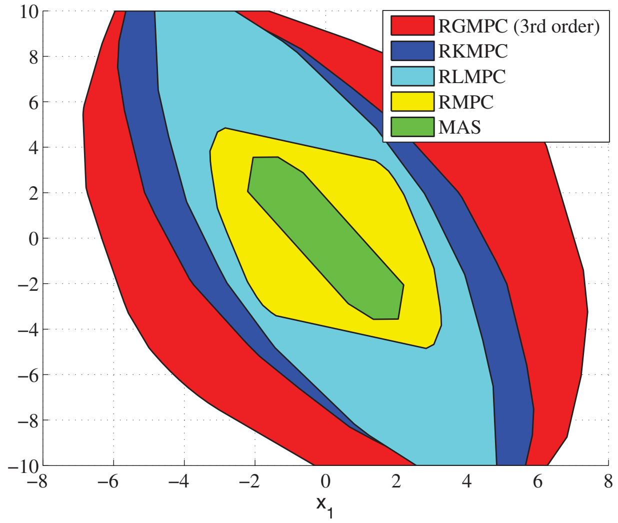


Fig. 1. Feasible regions for model (23) with $n_c = 3$

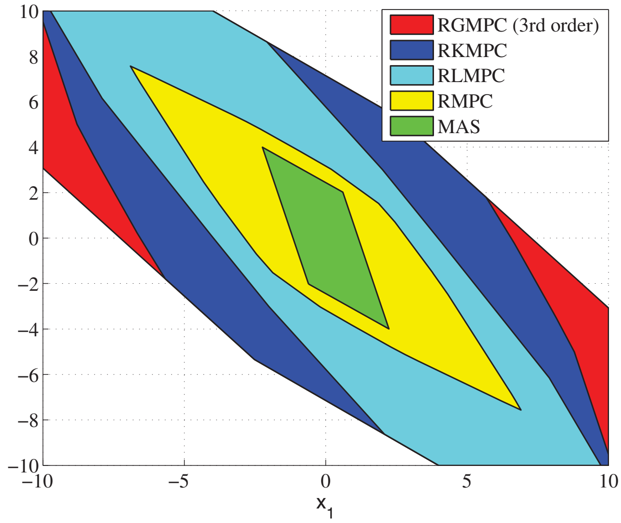


Fig. 2. Feasible regions for model (24) with $n_c = 3$

These results are based on arbitrary choices for the parameters in RGMPC, RKMPC and RLMPC. Further improvements both in feasible region and number of inequalities are possible by tailoring these parameters to the context.

## V. CONCLUSIONS AND FUTURE WORKS

The main contribution of this paper was to extend robust MPC algorithm to make use of alternative parameterisations of the d.o.f. and to consider the impact of doing so. Different alternative parameterisation functions including Laguerre, Kautz and higher order functions can be embedded within the robust MPC approach; the main requirement for this is to

TABLE I
COMPARISON OF FEASIBLE VOLUME

| $n_c = 2$ | | | | |
|---|---|---|---|---|
| Algorithm | Example 1 | Example 2 | Example 3 | Example 4 |
| RMPC | 29 | 38.1 | 1.5 $\times 10^3$ | 1.7 $\times 10^4$ |
| RLMPC | 66.9 | 47 | 4.1 $\times 10^3$ | 5.7 $\times 10^4$ |
| RKMPC | 71.3 | 52.6 | 4.4$\times 10^3$ | 6.1 $\times 10^4$ |
| $n_c = 3$ | | | | |
| Algorithm | Example 1 | Example 2 | Example 3 | Example 4 |
| RMPC | 43 | 56.5 | 2.5 $\times 10^3$ | 2.8 $\times 10^4$ |
| RLMPC | 108.4 | 150.6 | 5.4 $\times 10^3$ | 6.9 $\times 10^4$ |
| RKMPC | 141.7 | 215.6 | 6.1 $\times 10^3$ | 9.4 $\times 10^4$ |
| RGMPC<br>(3rd order) | 225.4 | 237 | 6.2 $\times 10^3$ | 9.5 $\times 10^4$ |

TABLE II
COMPARISON OF INEQUALITIES

| $n_c = 2$ | | | | |
|---|---|---|---|---|
| Algorithm | Example 1 | Example 2 | Example 3 | Example 4 |
| RMPC | 28 | 20 | 36 | 55 |
| RLMPC | 50 | 36 | 57 | 113 |
| RKMPC | 46 | 54 | 65 | 127 |
| $n_c = 3$ | | | | |
| Algorithm | Example 1 | Example 2 | Example 3 | Example 4 |
| RMPC | 48 | 40 | 57 | 93 |
| RLMPC | 74 | 38 | 74 | 109 |
| RKMPC | 105 | 48 | 88 | 153 |
| RGMPC<br>(3rd order) | 141 | 60 | 67 | 163 |

show how a robust invariant set can be computed with different parameterisations of the d.o.f.. Examples based on alternative parameterisation demonstrate that, for a fixed number of d.o.f., in many cases such parameterisations may improve the feasible region without a large change to the number of inequalities required to describe the robust invariant set.

Consequently this approach is worth further investigation. Moreover, there is a need to further investigate in parallel issues such as: which alternative parameterisation is best for particular problem and what choice of parameter(s) within parameterisation will lead to an efficient online optimisation QP structure? From Table II, it is observed that parameter selection effects the number of inequalities. The parameter selection will be based on both number of inequalities and feasible region. Another interesting avenue is to consider the computational efficiency for multi-parameteric quadratic programming (mp-QP) solution to RMPC. Finally, this paper consider parameter uncertainty only and thus extensions to consider disturbances are also required.

## REFERENCES

[1] E. Camacho and C. Bordons, *Model predictive control*, Springer, 2003.

[2] D. Q. Mayne, J. B. Rawlings, C. Rao and P. Scokaret, ”Constrained model predictive control”, *Automatica*, vol. 36, 2000, pp 789-814.

[3] J. A. Rossiter, *Model-based predictive control, A practical approach*, CRC Press, London; 2003.

[4] E. Gilbert and K. Tan, ”Linear sysetms with state and control constraints: The theory and application of maximal output admissible set”, *IEEE Trans. on Automatic Control*, vol. 36, 1991, pp 1008-1020.

[5] G. De Nicolao and L. Magni and R. Scattolini, ”On the robustness of receding horizon control with terminal constraints”, *IEEE Trans. on Automatic Control*, vol. 41, 1996, pp 451–453.

[6] L. Magni and R. Sepulchre, ”Stability margins of nonlinear receding horizon control via inverse optimity”, *Systems & Control Letters*, vol. 32, 1997, pp 241–245.

[7] B. Pluymers and J.A. Rossiter and J. Suykens and B. De Moor, ”The efficient computation of polyhedral invarient sets for linear sysetms with polytopic uncertainty”, *Proc. American Control Conference*, Portland, USA, 2005.

[8] B. Pluymers and J.A. Rossiter and J. Suykens and B. De Moor, ”A simple algorithm for robust MPC”, *Proc. 16th IFAC World Congress*, Prague, Czech Republic, 2005.

[9] B. Pluymers and J. A. K. Suykens and B. de Moor, ”Min-max feedback MPC using a time-varying terminal constraint set and comments on ”Efficient robust constrained model predictive control with a time-varying terminal constraint set”, *Systems & Control Letters*, vol. 54, 2005, pp 1143–1148.

[10] P. O. M. Scokaert and D. Q. Mayne, ”Min-max feedback model predictive control for constrained linear systems”, *IEEE Trans. on Automatic Control*, vol. 43, 1998, pp 1136–1142.

[11] E. C. Kerrigan and J. M. Maciejowski, ”Feedback min-max model predictive control using a single linear program: Robust stability and the explicit solution”, *International Journal of Robust and Nonlinear Control*, vol. 14, 2004, pp 395–413.

[12] M. V. Kothare and V. Balakrishnan and M. Morari, ”Robust constrained model predictive control using linear matrix inequalities”, *Automatica*, vol. 32, 1996, pp 1361–1379.

[13] B. Kouvaritakis and J. A. Rossiter and J. Schuurmans, ”Efficient robust predictive control”, *IEEE Trans. on Automatic Control*, vol. 45, 2000, pp 1545–1549.

[14] M. Cannon and B. Kouvaritakis, ”Optimizing prediction dynamics for robust MPC”, *IEEE Trans. on Automatic Control*, vol. 50, 2005, pp 1892–1897.

[15] L. Imsland and J.A. Rossiter and B. Pluymers and J. Suykens, ”Robust triple mode MPC”, *Internal Journal of Control*, vol. 81, 2008, pp 679–689.

[16] J. A. Rossiter and B. Pluymers, ”The potential of interpolation for simplifying predictive control and application to LPV systems”, *Proc. International Workshop on Assessment and Future Directions in NMPC*, Freudenstadt-Lauterbad, Germany, 2005.

[17] J.A. Rossiter and L. Wang and G. Valencia-Palomo, ”Efficient algorithms for trading off feasibility and performance in predictive control”, *Internal Journal of Control*, vol. 83, 2010, pp 789–797.

[18] B. Khan and J. A. Rossiter and G. Valencia-Palomo, ”Exploiting Kautz functions to improve feasibility in MPC”, *Proc. 18th IFAC World Congress*, Milan, Itlay, 2011.

[19] B. Khan and J. A. Rossiter, ”Triple Mode MPC or Laguerre MPC: a comparison”, *Proc. 18th IFAC World Congress*, Milan, Itlay, 2011.

[20] B. Khan and J. A. Rossiter, ”Generalised parameterisation for MPC”, *Proc. IASTED ISC 2011*, Cambridge, UK, 2011.

[21] J. A. Rossiter and B. Kouvaritakis and M. Cannon, ”Computational efficient algorithms for constraint handling with guaranteed stability and near optimality”, *Internal Journal of Control*, vol. 74, 2001, pp 1678–1689.

[22] P.O.M. Scokaert and J.B. Rawlings, ”Constrained linear quadratic regulation”, *IEEE Trans. on Automatic Control*, vol. 43, 1998, pp 1163–1168.

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# Dynamic Optimization of Polymer Flooding with Free Terminal Time based on Maximum Principle

Shurong Li, Yang Lei, Xiaodong Zhang, Qiang Zhang, Shaowen Peng
College of Information and Control Engineering
China University of Petroleum (East China)
Qingdao, China
yutian_hdpu2003@163.com

*Abstract*—**Polymer flooding is an important technology for enhanced oil recovery (EOR). In this paper, an optimal control model of distributed parameter systems (DPS) for polymer flooding is established, which involves the performance index as maximum of the profit, the governing equations as the seepage equations of polymer flooding, and some inequality constraints as polymer concentration and injection amount limitation. The injection polymer concentration and the terminal time of polymer flooding are chosen as control variables. For this distributed parameter optimal control problem (OCP) with free terminal time, a solution method based on maximum principle is proposed. Firstly, the free terminal time OCP of polymer flooding is transformed into a fixed final time problem by introducing a normalized time variable. Then through application of the maximum principle, adjoint equations and gradients of the objective functional are obtained to optimize the injection polymer concentration and the terminal time simultaneously. Finally, the numerical results of an example illustrate the effectiveness of the proposed method.**

***Keywords-optimal control; maximum principle; distributed parameter system; polymer flooding; free terminal time.***

## I. INTRODUCTION

The optimal control method has been researched in EOR techniques in recent years. Ramirez and Fathi firstly applied it to optimize the injection process of surfactant flooding [1], [2]. Then the optimal control method was used to other enhanced oil recovery techniques [3], such as steam flooding, caustic flooding and gas injection etc. The optimal gas-cycling decision problem of a condensate reservoir has been studied by Ye [4]. The dynamic optimization of water flooding with smart wells has been studied before by Brouwer [5], [6], Sarma [7], [8] and Zhang [9].

Polymer flooding is one of the most effective EOR techniques [10]. Because of the high costs associated with polymer flooding projects, optimal control method must be developed to reduce producing costs while increasing the profit of oil recovered. In this paper, the OCP of a polymer flooding process with free terminal time is considered. The performance index of the OCP is expressed by maximizing the economic benefit. The governing equations are a set of partial differential equations (PDEs), which are a pressure equation, a water saturation equation and a polymer concentration equation respectively. The constraint conditions include the polymer concentration constraint and other inequality constraints. The control variables are chosen as the injection concentrations and the free terminal time. Then the determination of polymer injection strategies turns to solve this OCP of DPS. A normalized time variable is introduced to transform the free terminal time OCP of polymer flooding into a fixed final time problem. Then the necessary conditions for optimality are obtained by Pontryagin's maximum principle. A gradient numerical method is presented for solving the transformed OCP. Finally, an example of polymer flooding project involving a heterogeneous reservoir case is investigated and the results show the efficiency of the proposed method.

The rest of this article is organized as follows: In section II the optimal control model of polymer flooding with free terminal time is built. In section III the original OCP with free terminal time is transformed into a fixed terminal time problem and the necessary conditions for optimality are obtained. In section IV a gradient numerical method is proposed for solving the transformed OCP. In section V an example of polymer flooding accompanied with the optimal results is given. And in section VI some conclusions are derived.

## II. MATHEMATICAL FORMULATION OF OPTIMAL CONTROL

### A. Performance Index

Let $\Omega \in R^2$ denote the domain of reservoir with boundary $\partial\Omega$, $n$ be the unit outward normal on $\partial\Omega$, and $(x,y)\in\Omega$ be the coordinate of a point in the reservoir. We suppose that there exist $N_w$ injection wells and $N_o$ production wells in the oilfield. The injection wells are located at $L_w = \{(x_{wi}, y_{wi}) \mid i = 1,2,\ldots,N_w\}$ and the production wells are located at $L_o = \{(x_{oj}, y_{oj}) \mid j = 1,2,\ldots,N_o\}$, respectively. For polymer flooding, we might wish to increase the profit and reduce the producing cost. Given a free terminal time $t_f$, the performance index is given mathematically by

$$\max\ J = \int_0^{t_f} \iint_\Omega \left[\xi_o(1-f_w)q_{out} - \xi_p q_{in} c_{pin}\right] d\sigma dt, \qquad (1)$$

This work is partially supported by the Natural Science Foundation of China Grant #60974039, the Natural Science Foundation of Shandong Province of China Grant #ZR2011FM002 and the Fundamental Research Funds for the Central Universities Grant #27R1105018A.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

where $\xi_p$ is the cost coefficient of polymer ($10^4\$/m^3$), $\xi_o$ is the price coefficient of oil ($10^4\$/m^3$), $f_w$ is the fractional flow of water, $q_{in}$ is the velocity of polymer injection ($m/day$), $q_{out}$ is the velocity of fluid production ($m/day$) and $c_{pin}$ is the injection concentration of polymer ($g/L$).

### *B. Governing Equations*

Let $p(x,y,t)$, $S_w(x,y,t)$ and $c_p(x,y,t)$ denote the pressure, water saturation and polymer concentration of the reservoir, respectively, at a point $(x,y)\in\Omega$ and a time $t\in[0,t_f]$, then $p(x,y,t)$, $S_w(x,y,t)$ and $c_p(x,y,t)$ satisfy the following partial differential equations (PDEs):

- Pressure equation

$$\frac{\partial}{\partial x}\left(k_p r_o\frac{\partial p}{\partial x}\right)+\frac{\partial}{\partial y}\left(k_p r_o\frac{\partial p}{\partial y}\right)-(1-f_w)q_{out}=h\frac{\partial a_o}{\partial t}, \quad (2)$$

- Water saturation equation

$$\frac{\partial}{\partial x}\left(k_p r_w\frac{\partial p}{\partial x}\right)+\frac{\partial}{\partial y}\left(k_p r_w\frac{\partial p}{\partial y}\right)+q_{in}-f_w q_{out}=h\frac{\partial a_w}{\partial t}, \quad (3)$$

- Polymer concentration equation

$$\begin{aligned}&\frac{\partial}{\partial x}\left(k_d r_d\frac{\partial c_p}{\partial x}\right)+\frac{\partial}{\partial x}\left(k_p r_c\frac{\partial p}{\partial x}\right)+\frac{\partial}{\partial y}\left(k_d r_d\frac{\partial c_p}{\partial y}\right)+\\&\frac{\partial}{\partial y}\left(k_p r_c\frac{\partial p}{\partial y}\right)+q_{in}c_{pin}-f_w q_{out}c_p=h\frac{\partial a_c}{\partial t}.\end{aligned} \quad (4)$$

The boundary conditions and initial conditions are

$$\left.\frac{\partial p}{\partial n}\right|_{\partial\Omega}=0,\ \left.\frac{\partial S_w}{\partial n}\right|_{\partial\Omega}=0,\ \left.\frac{\partial c_p}{\partial n}\right|_{\partial\Omega}=0, \quad (5)$$

$$\begin{aligned}&p(x,y,0)=p^0(x,y),\ S_w(x,y,0)=S_w^0(x,y),\\&c_p(x,y,0)=c_p^0(x,y),\end{aligned} \quad (6)$$

The corresponding parameters in (2)–(4) are defined as

$$k_p=Kh\,,\ k_d=Dh\,, \quad (7)$$

$$r_o=\frac{k_{ro}}{B_o\mu_o}\,,\ r_w=\frac{k_{rw}}{B_wR_k\mu_w}\,,\ r_c=\frac{k_{rw}c_p}{B_wR_k\mu_p}\,,\ r_d=\frac{\phi_p S_w}{B_w}, \quad (8)$$

$$a_o=\frac{\phi(1-S_w)}{B_o},\ a_w=\frac{\phi S_w}{B_w},\ a_c=\frac{\phi_p S_w c_p}{B_w}+\rho_r(1-\phi)C_{rp}, \quad (9)$$

where $K(x,y)$ is the absolute permeability ($\mu m^2$), $h$ is the thickness of the reservoir bed ($m$), $D$ is the diffusion coefficient of polymer ($m^2/s$), $\rho_r$ ($kg/m^3$) is the rock density, and $\mu_o$ ($mPa\cdot s$) is the oil viscosity. Other parameters definition can refer to [11] for details.

### *C. Constraints*

The performance index (1) should be subject to the polymer concentration constraint

$$0\le c_{pin}\le c_{\max}, \quad (10)$$

the injection amount constraint

$$\int_0^{t_f}\iint_\Omega q_{in}c_{pin}d\sigma dt\le m_{p\max}, \quad (11)$$

and the terminal state constraint

$$f_w|_{t=t_f}=98\%, \quad (12)$$

where $c_{\max}$ is the maximum injection concentration and $m_{p\max}$ is the maximum polymer amount.

## III. Necessary Conditions of Optimal Control

### *A. Problem Transformation*

For the orginal OCP of polymer flooding with free terminal time, a normalized time variable is introduced,

$$\tau=t/t_f, \quad (13)$$

Since $t\in[0,\ t_f]$, we have $\tau\in[0,\ 1]$. By using the definite integral by substitution, the performance index (1) is expressed as

$$\max\ J=\int_0^1\iint_\Omega t_f\left[\xi_o(1-f_w)q_{out}-\xi_p q_{in}c_{pin}\right]d\sigma d\tau. \quad (14)$$

The system state vector is denoted by

$$\boldsymbol{u}(x,y,t)=[p,S_w,c_p]^T. \quad (15)$$

The control for the process is the polymer concentration of injected fluid

$$v(x,y,t)=c_{pin},\ (x,y)\in L_w. \quad (16)$$

Then the governing equations (2)–(4) which can be expressed by

$$\frac{\partial\boldsymbol{a}}{\partial t}=\tilde{\boldsymbol{f}}(\boldsymbol{u},\boldsymbol{u}_x,\boldsymbol{u}_y,\boldsymbol{u}_{xx},\boldsymbol{u}_{yy},v,t), \quad (17)$$

are normalized as

$$\frac{\partial\boldsymbol{a}}{\partial\tau}=t_f\tilde{\boldsymbol{f}}(\boldsymbol{u},\boldsymbol{u}_x,\boldsymbol{u}_y,\boldsymbol{u}_{xx},\boldsymbol{u}_{yy},v,\tau), \quad (18)$$

where $\boldsymbol{u}_l=\partial\boldsymbol{u}/\partial l$, $l=x,y$.

If $t_f$ is treated as a new optimization variable and $\boldsymbol{v}=[v,\ t_f]^T$ is denoted as control vector, the orginal OCP of polymer flooding is transformed into the following fixed terminal time problem,

$$\max\ J=\int_0^1\iint_\Omega F(\boldsymbol{u},\boldsymbol{v},\tau)d\sigma d\tau, \quad (19)$$

$$s.t.\ \ \boldsymbol{f}(\dot{\boldsymbol{u}},\boldsymbol{u},\boldsymbol{u}_x,\boldsymbol{u}_y,\boldsymbol{u}_{xx},\boldsymbol{u}_{yy},\boldsymbol{v},\tau)=0 \tag{20}$$

$$\boldsymbol{g}(\boldsymbol{u},\boldsymbol{u}_x,\boldsymbol{u}_y,\boldsymbol{u}_{xx},\boldsymbol{u}_{yy},\tau)=0\ , \tag{21}$$

$$\boldsymbol{u}(x,y,0)=\boldsymbol{u}^0(x,y), \tag{22}$$

$$\int_0^1\iint_\Omega c_1(\boldsymbol{v})d\sigma d\tau \le 0, \tag{23}$$

$$c_2(\boldsymbol{u}|_{\tau=1})=0, \tag{24}$$

$$0\le \boldsymbol{v}\le \boldsymbol{v}_{\max}. \tag{25}$$

where $\dot{\boldsymbol{u}}=\partial\boldsymbol{u}/\partial\tau$. With this transformation, at $t=t_f$, $\tau_f=1$, and in the dimensionless time domain the terminal time is fixed.

### *B. Maximum Principle of DPS*

A convenient way to cope with such an OCP of DPS (19)–(25) is through the use of distributed adjoint variables. We define the Hamiltonian as

$$H=F+\boldsymbol{\lambda}^T\boldsymbol{f}, \tag{26}$$

where $\boldsymbol{\lambda}(x,y,\tau)$ is the adjoint vector. Then the argument functional is given by,

$$\begin{aligned} J_A &= J+\int_0^1\iint_\Omega \boldsymbol{\lambda}^T\boldsymbol{f}(\dot{\boldsymbol{u}},\boldsymbol{u},\boldsymbol{u}_x,\boldsymbol{u}_y,\boldsymbol{u}_{xx},\boldsymbol{u}_{yy},\boldsymbol{v},\tau)d\sigma d\tau \\ &= \int_0^1\iint_\Omega H(\dot{\boldsymbol{u}},\boldsymbol{u},\boldsymbol{u}_x,\boldsymbol{u}_y,\boldsymbol{u}_{xx},\boldsymbol{u}_{yy},\boldsymbol{v},\tau)d\sigma d\tau. \end{aligned} \tag{27}$$

The increment of $J_A$, denoted by $\Delta J_A$, is formed by introducing variations $\delta\boldsymbol{u}$, $\delta\boldsymbol{u}_x$, $\delta\boldsymbol{u}_y$, $\delta\boldsymbol{u}_{xx}$, $\delta\boldsymbol{u}_{yy}$, $\delta\dot{\boldsymbol{u}}$, and $\delta v$ giving

$$\begin{aligned}\Delta J_A = {} & J_A(\boldsymbol{u}+\delta\boldsymbol{u},\boldsymbol{u}_x+\delta\boldsymbol{u}_x,\boldsymbol{u}_y+\delta\boldsymbol{u}_y,\boldsymbol{u}_{xx}+\delta\boldsymbol{u}_{xx},\boldsymbol{u}_{yy}+\delta\boldsymbol{u}_{yy}, \\ & \dot{\boldsymbol{u}}+\delta\dot{\boldsymbol{u}},\boldsymbol{v}+\delta\boldsymbol{v})-J_A(\boldsymbol{u},\boldsymbol{u}_x,\boldsymbol{u}_y,\boldsymbol{u}_{xx},\boldsymbol{u}_{yy},\dot{\boldsymbol{u}},\boldsymbol{v}).\end{aligned} \tag{28}$$

Expanding (28) in a Taylor series and retaining only the linear terms gives the variation of the functional, $\delta J_A$,

$$\begin{aligned}\delta J_A = \int_0^1\iint_\Omega\Bigg[ & \left(\frac{\partial H}{\partial\boldsymbol{u}}\right)^T\delta\boldsymbol{u}+\left(\frac{\partial H}{\partial\boldsymbol{u}_x}\right)^T\delta\boldsymbol{u}_x+\left(\frac{\partial H}{\partial\boldsymbol{u}_{xx}}\right)^T\delta\boldsymbol{u}_{xx}+ \\ & \left(\frac{\partial H}{\partial\boldsymbol{u}_y}\right)^T\delta\boldsymbol{u}_y+\left(\frac{\partial H}{\partial\boldsymbol{u}_{yy}}\right)^T\delta\boldsymbol{u}_{yy}+\left(\frac{\partial H}{\partial\dot{\boldsymbol{u}}}\right)^T\delta\dot{\boldsymbol{u}}+ \\ & \left(\frac{\partial H}{\partial\boldsymbol{v}}\right)^T\delta\boldsymbol{v}\Bigg]d\sigma d\tau.\end{aligned} \tag{29}$$

Since the variations $\delta\boldsymbol{u}$, $\delta\boldsymbol{u}_l$, $\delta\boldsymbol{u}_{ll}$ ($l=x,y$) and $\delta\dot{\boldsymbol{u}}$ are not independent can be expressed in terms of the variations $\delta\boldsymbol{u}$ by integrating the following three terms by parts

$$\begin{aligned}\iint_\Omega\left[\left(\frac{\partial H}{\partial\boldsymbol{u}_l}\right)^T\delta\boldsymbol{u}_l\right]d\sigma = {} & \iint_\Omega\frac{\partial}{\partial l}\left[\left(\frac{\partial H}{\partial\boldsymbol{u}_l}\right)^T\delta\boldsymbol{u}\right]d\sigma- \\ & \iint_\Omega\left[\frac{\partial}{\partial l}\left(\frac{\partial H}{\partial\boldsymbol{u}_l}\right)^T\delta\boldsymbol{u}\right]d\sigma,\end{aligned} \tag{30}$$

$$\begin{aligned}\iint_\Omega\left[\left(\frac{\partial H}{\partial\boldsymbol{u}_{ll}}\right)^T\delta\boldsymbol{u}_{ll}\right]d\sigma = {} & \iint_\Omega\left[\frac{\partial^2}{\partial l^2}\left(\frac{\partial H}{\partial\boldsymbol{u}_{ll}}\right)\right]^T\delta\boldsymbol{u}d\sigma+ \\ & \iint_\Omega\frac{\partial}{\partial l}\left[\left(\frac{\partial H}{\partial\boldsymbol{u}_{ll}}\right)^T\delta\boldsymbol{u}_l-\frac{\partial}{\partial l}\left(\frac{\partial H}{\partial\boldsymbol{u}_{ll}}\right)^T\delta\boldsymbol{u}\right]d\sigma,\end{aligned} \tag{31}$$

$$\int_0^1\left(\frac{\partial H}{\partial\dot{\boldsymbol{u}}}\right)^T\delta\dot{\boldsymbol{u}}=\left[\left(\frac{\partial H}{\partial\dot{\boldsymbol{u}}}\right)^T\delta\boldsymbol{u}\right]\Bigg|_0^1-\int_0^1\frac{\partial}{\partial\tau}\left(\frac{\partial H}{\partial\dot{\boldsymbol{u}}}\right)^T\delta\boldsymbol{u}d\tau. \tag{32}$$

By using the above expressions (30)–(32), the first variation $\delta J_A$ is written as

$$\begin{aligned}\delta J_A = {} & \int_0^1\iint_\Omega\left(\frac{\partial H}{\partial\boldsymbol{u}}-\sum_{l=x,y}\frac{\partial}{\partial l}\frac{\partial H}{\partial\boldsymbol{u}_l}+\sum_{l=x,y}\frac{\partial^2}{\partial l^2}\frac{\partial H}{\partial\boldsymbol{u}_{ll}}+\right. \\ & \left.-\frac{\partial}{\partial\tau}\frac{\partial H}{\partial\dot{\boldsymbol{u}}}\right)^T\delta\boldsymbol{u}d\sigma d\tau+\int_0^1\iint_\Omega\sum_{l=x,y}\frac{\partial}{\partial l}\left\{\left(\frac{\partial H}{\partial\boldsymbol{u}_{ll}}\right)^T\delta\boldsymbol{u}_l+\right. \\ & \left.\left[\frac{\partial H}{\partial\boldsymbol{u}_l}-\frac{\partial}{\partial l}\left(\frac{\partial H}{\partial\boldsymbol{u}_{ll}}\right)^T\right]\delta\boldsymbol{u}\right\}d\sigma d\tau+ \\ & \iint_\Omega\left[\left(\frac{\partial H}{\partial\dot{\boldsymbol{u}}}\right)^T\delta\boldsymbol{u}\right]\Bigg|_0^1 d\sigma+\int_0^1\iint_\Omega\left(\frac{\partial H}{\partial\boldsymbol{v}}\right)\delta v d\sigma\, d\tau.\end{aligned} \tag{33}$$

Applying Pontryagin's Maximum Principle, the necessary conditions for an extremum of $J_A$ are given by

- Adjoint Equations

$$\frac{\partial H}{\partial\boldsymbol{u}}-\sum_{l=x,y}\left(\frac{\partial}{\partial l}\frac{\partial H}{\partial\boldsymbol{u}_l}+\frac{\partial^2}{\partial l^2}\frac{\partial H}{\partial\boldsymbol{u}_{ll}}\right)-\frac{\partial}{\partial\tau}\frac{\partial H}{\partial\dot{\boldsymbol{u}}}=0. \tag{34}$$

- Transversality Boundary Conditions

$$\iint_\Omega\sum_{l=x,y}\frac{\partial}{\partial l}\left\{\left(\frac{\partial H}{\partial\boldsymbol{u}_{ll}}\right)^T\delta\boldsymbol{u}_l+\left[\frac{\partial H}{\partial\boldsymbol{u}_l}-\frac{\partial}{\partial l}\left(\frac{\partial H}{\partial\boldsymbol{u}_{ll}}\right)^T\right]\delta\boldsymbol{u}\right\}d\sigma=0. \tag{35}$$

- Transversality Terminal Conditions

$$\frac{\partial H}{\partial\dot{\boldsymbol{u}}}=0\ ,\ \text{ at }\ \tau=1. \tag{36}$$

- Optimal Control

With the first three necessary conditions being satisfied, the first variation becomes

$$\delta J_A = \int_0^1 \iint_\Omega \left( \frac{\partial H}{\partial \boldsymbol{v}} \right) \delta \boldsymbol{v} d\sigma d\tau. \tag{37}$$

If the variation $\boldsymbol{\delta v}$ is not constrained, then the necessary condition for an extremum is $\partial H / \partial \boldsymbol{v} = 0$ .

If the variation $\boldsymbol{\delta v}$ is constrained, which means that the control is at a constraint boundary, then the necessary condition for maximizing the performance functional is

$$\max_{\boldsymbol{v}} H. \tag{38}$$

### C. Necessary Conditions of OCP for Polymer Flooding

Let $\lambda(x,y,\tau) = (\lambda_1, \lambda_2, \lambda_3)^T$ denote the adjoint vector of OCP for polymer flooding. Applying the theory developed and substituting the governing equations (2)–(4) into (34), the adjoint equations reduce for the polymer flooding under consideration as given in,

$$\begin{aligned}&\sum_{l=x,y} \left\{ \frac{\partial}{\partial l} \left( k_p r_o \frac{\partial \lambda_1}{\partial l} \right) + \frac{\partial}{\partial l} \left( k_p r_w \frac{\partial \lambda_2}{\partial l} \right) + \frac{\partial}{\partial l} \left( k_p r_c \frac{\partial \lambda_3}{\partial l} \right) - \right. \\ &\left. \left[ k_p \frac{\partial r_o}{\partial p} \frac{\partial p}{\partial l} \frac{\partial \lambda_1}{\partial l} + k_d \frac{\partial r_w}{\partial p} \frac{\partial p}{\partial l} \frac{\partial \lambda_2}{\partial l} + \left( k_p \frac{\partial r_c}{\partial p} \frac{\partial p}{\partial l} + \right. \right. \right. \\ &\left. \left. \left. k_d \frac{\partial r_d}{\partial p} \frac{\partial c_p}{\partial l} \right) \frac{\partial \lambda_3}{\partial l} \right] \right\} - q_{out} \left( \xi_o \frac{\partial f_w}{\partial p} - \frac{\partial f_w}{\partial p} \lambda_1 + \frac{\partial f_w}{\partial p} \lambda_2 + \right. \\ &\left. c_p \frac{\partial f_w}{\partial p} \lambda_3 \right) + \frac{\partial a_o}{\partial p} \frac{\partial \lambda_1}{\partial \tau} + \frac{\partial a_w}{\partial p} \frac{\partial \lambda_2}{\partial \tau} + \frac{\partial a_c}{\partial p} \frac{\partial \lambda_3}{\partial \tau} = 0,\end{aligned} \tag{39}$$

$$\begin{aligned}&\sum_{l=x,y} \left[ -k_p \frac{\partial p}{\partial l} \left( \frac{\partial r_o}{\partial S_w} \frac{\partial \lambda_1}{\partial l} + \frac{\partial r_w}{\partial S_w} \frac{\partial \lambda_2}{\partial l} + \frac{\partial r_c}{\partial S_w} \frac{\partial \lambda_3}{\partial l} \right) - \right. \\ &\left. k_d \frac{\partial r_d}{\partial S_w} \frac{\partial c_p}{\partial l} \frac{\partial \lambda_3}{\partial l} \right] - q_{out} \left( \xi_o \frac{\partial f_w}{\partial S_w} - \frac{\partial f_w}{\partial S_w} \lambda_1 + \frac{\partial f_w}{\partial S_w} \lambda_2 + \right. \\ &\left. c_p \frac{\partial f_w}{\partial S_w} \lambda_3 \right) + \frac{\partial a_o}{\partial S_w} \frac{\partial \lambda_1}{\partial \tau} + \frac{\partial a_w}{\partial S_w} \frac{\partial \lambda_2}{\partial \tau} + \frac{\partial a_c}{\partial S_w} \frac{\partial \lambda_3}{\partial \tau} = 0,\end{aligned} \tag{40}$$

$$\begin{aligned}&\sum_{l=x,y} \left[ \frac{\partial}{\partial l} \left( k_d r_d \frac{\partial \lambda_3}{\partial l} \right) - k_p \frac{\partial p}{\partial l} \left( \frac{\partial r_w}{\partial c_p} \frac{\partial \lambda_2}{\partial l} + \frac{\partial r_c}{\partial c_p} \frac{\partial \lambda_3}{\partial l} \right) \right] - \\ &q_{out} \left[ \xi_o \frac{\partial f_w}{\partial c_p} - \frac{\partial f_w}{\partial c_p} \lambda_1 + \frac{\partial f_w}{\partial c_p} \lambda_2 + \left( c_p \frac{\partial f_w}{\partial c_p} + f_w \right) \lambda_3 \right] + \\ &\frac{\partial a_c}{\partial c_p} \frac{\partial \lambda_3}{\partial \tau} = 0,\end{aligned} \tag{41}$$

The boundary conditions of adjoint equations for the OCP of polymer flooding are expressed as

$$\left( r_o \frac{\partial \lambda_1}{\partial l} + r_w \frac{\partial \lambda_2}{\partial l} \right) \Bigg|_{\partial \Omega} = 0, \ \frac{\partial \lambda_3}{\partial l} \Bigg|_{\partial \Omega} = 0, \ l = x, y. \tag{42}$$

The terminal conditions of adjoint equations can be simplified to

$$\lambda_1(x,y,\tau_f) = 0, \ \lambda_2(x,y,\tau_f) = 0, \ \lambda_3(x,y,\tau_f) = 0. \tag{43}$$

## IV. Numerical Solution

We propose an iterative numerical technique for determining the optimal injection strategies of polymer flooding. The computational procedure is based on adjusting estimates of control $\boldsymbol{v}$ to improve the value of the objective functional. If the control $\boldsymbol{v}$ is not optimal, then a correction $\boldsymbol{\delta v}$ is determined so that the functional is made lager, that is, $\delta J_A > 0$ . If $\boldsymbol{\delta v}$ is selected as

$$\boldsymbol{\delta v} = w \cdot \frac{\partial H}{\partial \boldsymbol{v}}, \tag{44}$$

where $w$ is an arbitrary positive weighting factor. Then the functional variation becomes

$$\delta J_A = \int_0^1 \iint_\Omega w \left( \frac{\partial H}{\partial \boldsymbol{v}} \right)^T \left( \frac{\partial H}{\partial \boldsymbol{v}} \right) d\sigma d\tau \geq 0. \tag{45}$$

Thus, choosing $\boldsymbol{\delta v}$ as the gradient direction ensures a local improvement in the objective functional, $J_A$ . Substituting the governing equations into (44), we obtain the gradient of performance index with respect to the injection polymer concentration $v$

$$\nabla J(v) = w t_f q_{in} (\lambda_3 - \xi_p), \ (x,y) \in L_w, \tag{46}$$

and the gradient of performance index with respect to the terminal time $t_f$

$$\nabla J(t_f) = w \int_0^1 \iint_\Omega \left[ \xi_o (1 - f_w) q_{out} - \xi_p q_{in} c_{pin} + \boldsymbol{\lambda}^T \tilde{\boldsymbol{f}} \right] d\sigma d\tau, \tag{47}$$

The computational algorithm of control iteration based on gradient direction is as follows:

(1) Initialization: Make an initial guess for the control $t_f$ and $v(x,y,\tau)$ , $(x,y) \in L_w$ , $\tau \in [0, 1]$ .

(2) Resolution of Governing Equations: Using stored current value of control, integrate the governing equations forward in time with known initial governing conditions. The profit functional is evaluated, and the coefficients involved in the adjoint equations which are function of the state solution are computed and stored.

(3) Resolution of Adjoint Equations: Using the stored coefficients, integrate the adjoint equations numerically backward in time with known final time adjoint conditions by Equation (43). Compute and store $\nabla J(\boldsymbol{v})$ as defined by Equations (46) and (47).

(4) Computation of New Control: Using the evaluated $\nabla J(\boldsymbol{v})$ , an improved function is computed as

$$\boldsymbol{v}^{new} = \boldsymbol{v}^{old} + \nabla J(\boldsymbol{v}). \tag{48}$$

A single variable search strategy can be used to find the value of the positive weighting factor $w$ which maximizes the improvement in the performance functional using Equation (46) and (47).

(5) Termination: Go to Step (2) until reach the following stop criteria

$$\left| J^{new} - J^{old} \right| < \varepsilon, \tag{49}$$

where $\varepsilon$ is a small positive number.

It should be noted that the penalty function method is used to deal with the injection amount constraint and the terminal state constraint (23) and (24). The details of this method can refer to [12].

## V. Example

The two-phase flow of oil and water in a heterogeneous reservoir is considered. The reservoir covers an area of 421.02 × 443.8 $m^2$ and has a thickness of 5 $m$ and is discretized into 90 (9×10×1) grid blocks by finite difference method. There are four injection wells and a production well in reservoir as shown in Figure 1. Polymer is injected when the fractional flow of water for the production well comes to 97% after water flooding. In the performance index calculation, we use the price of oil $\xi_o = 0.0503\ (10^4\$/m^3)$ $[80\ (\$/\text{bbl})]$, and the cost of polymer $\xi_p = 2.5\times10^{-4}\ \ (10^4\$/kg)$. The fluid velocity of production well $q_{out}$ is 0.4624 $m/day$ and the fluid velocity of injection wells $q_{in}$ are all 0.1156 $m/day$. For the constraint (10), the maximum polymer concentration is $c_{\max} = 2.2\ (g/L)$. The parameters of the reservoir description and the fluid data are shown in Table I. Other parameters can refer to [11].

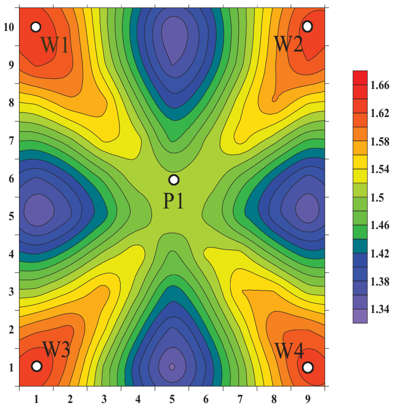


Figure 1. Permeability distribution and well position

The initial injection strategy obtained from engineering method is 1.7 $(g/L)$. The time domain of polymer injection is 0–1500 $days$. When the water fractional flow of production well reaches 98%, the terminal time $t_f = 5498\ (days)$. The performance index is $J = \$1.572\times10^7$ with oil production 32022 $m^3$ and polymer injection 153000 $kg$. For comparison, the results obtained by engineering method are considered as the initial control strategies of the proposed iterative gradient method. The maximum injection polymer amount is $m_{p\max} = 153000\ (kg)$. A backtracking search strategy [12] is used to find the appropriate weighting term $w$ and the stopping criterion is chosen as $\varepsilon = 1\times10^{-5}$. By using the proposed algorithm, we obtain a cumulative oil of 32750 $m^3$ and a cumulative polymer of 153000.02 $kg$ yielding the profit of $J^* = \$1.609\times10^7$. The results show an increase in performance index of \$ $3.7\times10^5$. The optimized terminal time is $t_f = 5165\ (days)$. Figure 2–5 show the injection strategies of the two methods. Figure 6 and Figure 7 show the curves of water fractional flow and accumulative oil production, respectively. The fractional flow of water obtained by proposed method is lower than that by engineering method. Therefore, with the same cumulative polymer injection, the proposed method gets more oil production and higher recovery ratio.

TABLE I. Reservoir Description and Fluid Data

| Symbol | Data | Symbol | Data |
|---|---|---|---|
| $p^0$ | 12 $MPa$ | $S_w^0$ | 0.35 |
| $c_p^0$ | 0 | $\mu_o$ | 15 $cp$ |
| $\mu_w$ | 1 $cp$ | $\phi$ | 0.31 |
| $D$ | 0.002 | $\rho_r$ | 2000 $kg/m^3$ |
| $h$ | 5 $m$ | $C_{rp}$ | $9.38\times10^{-6}$ |

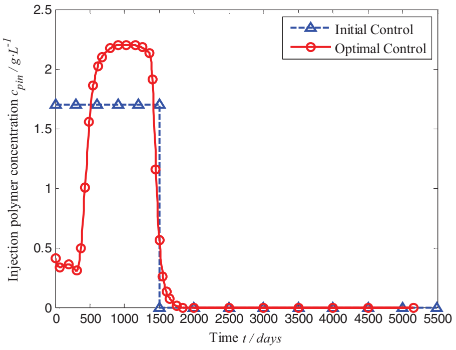


Figure 2. Injection polymer concentration of W1

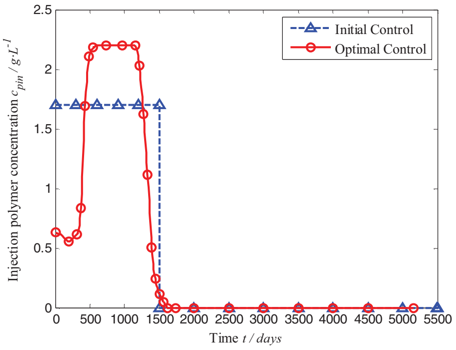


Figure 3. Injection polymer concentration of W2

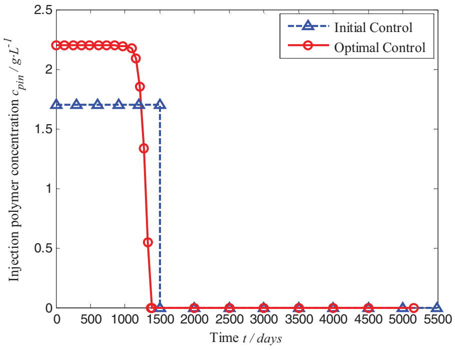


Figure 4. Injection polymer concentration of W3

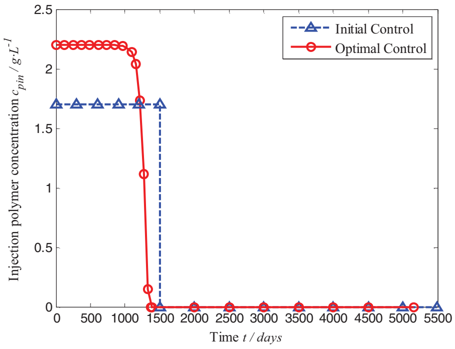


Figure 5. Injection polymer concentration of W4

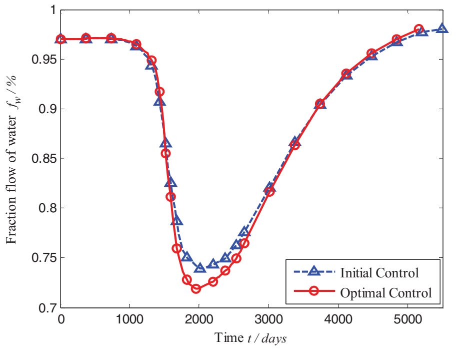


Figure 6. Water fraction flow of the production well P1

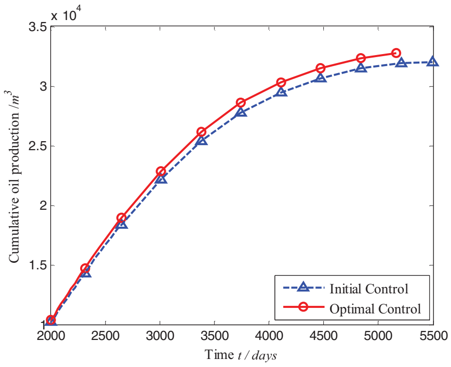


Figure 7. Cumulative oil production

## VI. Conclusion

In this work, a new optimal control model of DPS is established for the dynamic injection strategies making of polymer flooding. The original problem with free terminal time is transformed into a fixed terminal time OCP by introducing a normalized timer variable. Necessary conditions of this OCP are obtained by using Pontryagin's maximum principle. An iterative computational algorithm is proposed for the determination of optimal injection strategies. The optimal control model of polymer flooding and the proposed method are used for a reservoir example and the optimum injection concentration profile is offered. The results show that the profit is enhanced by the proposed method. Meanwhile, more oil production and higher recovery ratio are obtained. The approach used is a powerful tool that can aid significantly in the development of operational strategies for EOR processes.

## References

[1] W. Ramirez, Z. Fathi and J. L. Cagnol, "Optimal injection policies for enhanced oil recovery: part l-theory and computational strategies," Society of Petroleum Engineers Journal, vol. 24, no. 3, pp. 328-332, June 1984.

[2] Z. Fathi, and W. Ramirez, "Optimal injection policies for enhanced oil recovery: part 2-surfactant flooding," Society of Petroleum Engineers Journal, vol. 24, no. 3, pp. 333-341, June 1984.

[3] W. Ramirez, Application of Optimal Control to Enhanced Oil Recovery, New York: Elsevier, 1987

[4] J. Ye, Y. Qi, and Y. Fang, "Application of optimal control theory to making gas-cycling decision of condensate reservoir," Chinese Journal of Computational Physics, vol. 15, no. 1, pp.71-76, January 1998.

[5] D. R. Brouwer and J. D. Jansen, "Dynamic optimization of water flooding with smart sells using optimal control theory," paper SPE 78278 presented at the SPE 13th European Petroleum Conference, Aberdeen, Scotland, 2002, pp. 1-14.

[6] D. R. Brouwer, J. D. Jansen, E. H. Vefring, and C. P. J. W. van Kruijsdijk, "Improved reservoir management through optimal control and continuous model updating," paper SPE 90149 presented at the SPE Annual Technical Conference and Exhibition, Houston, Texas, 2004, pp. 1-11.

[7] P. Sarma, K. Aziz and L. J. Durlofsky, "Implementation of adjoint solution for optimal control of smart wells," paper SPE 92864 presented at the 2005 SPE Reservoir Simulation Symposium, Houston, Texas, 2005, pp. 1-17.

[8] P. Sarma, W. H. Chen, L. J. Durlofsky and K. Aziz, "Production optimization with adjoint models under nonlinear control-state path inequality constraints," paper SPE 99959 presented at the 2006 SPE Intelligent Energy Conference and Exhibition, Amsterdam, The Netherlands, 2006, pp. 1-19.

[9] K. Zhang, J. Yao, L. M. Zhang and Y. J. Li, "Dynamic Real-time Optimization of Reservoir Production," Journal of Computers, vol. 6, no. 3, pp. 610-617, June 2011.

[10] Y. Qing, D. Caili, W. Yefei, T. Engao, Y. Guang, and Z. Fulin, "A study on mass concentration determination and property variations of produced polyacrylamide in polymer flooding," Petroleum Science and Technology, vol. 29, no. 3, pp. 227-235, December 2011.

[11] Y. Lei, S. Li, X. Zhang, Q. Zhang, and L. Guo, "Optimal control of polymer flooding based on mixed-integer iterative dynamic programming," International Journal of Control, vol. 84, no. 11, pp. 1903-1914, November 2011.

[12] J. Nocedal and S. J. Wright, Numerical Optimization, New York: Springer-Verlag, 2000.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Optimal Control for Mixing Enhancement in Boundary Layers at Membrane Walls

Hua Ouyang
School of Chemical Engineering
University of New South Wales
UNS Sydney, NSW 2052, Australia
Email: dr.hua.ouyang@gmail.com

Yuanqing Xia
School of Automation
Beijing Institute of Technology
Beijing 100081, China
Email: xia_yuanqing@bit.edu.cn

*Abstract*—**This paper proposes a scheme for mixing enhancement in the boundary layers of pressure-driven membrane systems. This scheme uses an external electric field to activate the ions in the area adjacent to the membrane surface and generate an electro-osmotic flow. This scheme should reduce fouling and concentration polarization close to the membrane surface and may increase productivity of membrane systems. The objective of the feedback control design for this system needs to determine the voltage (and waveform) applied to the electrodes so that the electric field can effectively increase the mixing in the vicinity of membrane surface, while saving control power. This paper uses a mixing index in terms of the spatial gradients of the perturbation velocity field, which describes the mixing caused by both length stretching and folding. An optimal control problem is defined to maximize mixing in the area adjacent to the membrane and achieve control energy efficiency. In addition, the efficacy of the feedback scheme is validated by Computation Fluid Dynamics (CFD) simulation. The given control law not only solves the optimal problem but also provides the desired waveform for such applications.**

## I. Introduction

Most solid surfaces in contact with water or an aqueous solution will be found to develop some type of electrical charge. The mechanisms that a surface acquires an electrical charge include preferential solution of surface ions, direct ionization of surface groups, substitution of surface ions, specific ion adsorption and so on [1]. The electrical charges gather on the solid surface and form an electric double layer. Physically, the two layers of ions align on the surface and lead to concentration polarization. In the membrane system in seawater filtration or brackish water filtration, the membrane can considered as a solid surface. Concentration polarization will result in fouling formed on the surface of the membrane.

Fouling and concentration polarization reduce the throughput and productivity of membrane systems and significantly increase operating costs. This in turn reduces the profitability of water treatment processes including desalination and recycling. Despite much work and improvement in the design and operation of membrane systems, throughput and productivity continue to be plagued by fouling and concentration polarization. Amongst the mechanisms that have been proposed to reduce fouling and concentration polarization, for example, feed pre-treatment [2], membrane surface modification or cleaning [3], reduction of solute concentration at the membrane surface by mixing enhancement in the flow offers the most promising way. Mixing enhancement in the vicinity of membrane surface can reduce concentration polarization in the region and hence lessen the chance of forming fouling. As a result, the throughput of membrane increases. This paper proposes a new approach to the reduction of solute concentration and fouling at the membrane surface based on producing electro-osmotic flow (EOF) instability.

Electric field is able to activate the ions in a solution to induce an electro-osmotic flow. This paper uses this fact to develop a new approach which enhances the mixing in the area adjacent to the membrane surface, via the use of electric field to generate an EOF in this area. Many results on using electric field to motivate electro-osmotic flows have been reported, for example, [4], [5], [6], [7], [8]. Distinct from most of these results which moves the ions in the whole channel like flow transport, the electric field in this scheme will mostly act on the boundary layer close to the membrane surface, where concentration polarization and fouling occur. This should enable us to use less energy to achieve mixing enhancement and hence it may be more energy efficient than existing electro-kinetic methods.

As mentioned in [9], mixing includes several types: the mixing of a single or similar fluids caused by stretching and folding of fluid; the mixing governed by diffusion and chemical reaction; the mixing caused by breakup and coalescence of fluid. Due to the variety of reasons leading to mixing, there are different mixing indices. Amongst these types, the mixing caused by fluid stretching and folding is the one of interest in the context of this paper. To describe this type of mixing, [9] gives a strict definition of stretching length based on the gradient of relative velocity, which is an important measure of mixing. Specifically, this stretching length uses the stretching tensor to describe mixing. Alexiadis et al [10] uses the vorticity or spin gradient tensor to describe the mixing induced by the vortices (folding) in the circumstance of membrane channel containing circular spacers. This paper adds up these two mixing measures and establishes a new mixing index.

In [11], [12], an objective function involving turbulent kinetic energy and a measure of the spatial gradients of turbulent velocities is used as the cost functional of an optimal flow control problem and maximizing this cost functional leads

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

to mixing enhancement. This cost functional includes a Frobenius norm of the gradient of relative velocity (perturbation velocity). This term reflects the stretching of fluid elements explicitly but the folding measurement is implicit. This term is positively related to mixing but it is not proportional to that of enstrophy which is said to be more related to mixing. In this paper, we further explore the Frobenius norm of the gradient of perturbation velocity and use it as our new mixing index.

The mixing enhancement within the area adjacent to the membrane surface leads to the increase of the throughput of the membrane. This suggests that the mixing within this small vicinity is much more important than that in the bulk solution; i.e., the effect of the electric field will mostly apply to the boundary layer of the system. In this paper, we focus our attention on the mixing enhancement in this rectangular area. This requires us to relate the boundary control action to the mixing in this area. Gauss's divergence theorem states that the outward flux of a vector field through a closed surface is equal to the volume integral of the divergence of the region inside the surface[13]. Based on Gauss's divergence theorem, [11], [12] proposes a heuristic flow control methods which convert a mixing increase problem in a 3D pipe to a boundary control problem. This paper relates the electric field close to the membrane surface to the perturbation velocity and their spatial gradients inside the surface and hence transforms the problem into a boundary control problem.

In addition, the electric field used to activate the fluid flow in the vicinity is generated from a pair of electrodes, which are installed outside the membrane. This restricts us from manipulating the distributive value of electric field when we construct the feeding voltage. Once the electrodes are fixed, the spatial distribution of this electric field is fixed, viz., the voltage will not affect the shape of this distribution. This paper uses the integral of electric field strength on the membrane surface to construct the feedback control law. This makes the results of this paper distinct from those in [12], where distributive flow injection is used.

To improve the energy efficiency, we integrate the perturbation kinetic energy, the new mixing index and the control effort to define a cost functional and formulate an optimal problem for the fluid flow control problem for membrane systems. This requires that the control candidates solve such an optimal problem and maximize the cost functional. In addition, the efficacy of the proposed mixing enhancement scheme and the given control law has been validated by CFD simulations. CFD is a widely used tool for the studies of membrane system. This reliable tool is utilized in order to gain insight into the phenomena taking place inside membrane modules, to assist the design process and improve the performance of modules.

This paper is organized as follows: Section 2 presents a new mixing index and the theory of enhancing mixing in the area adjacent to the membrane surface. This section includes the main results of this paper. Sections 3 uses CFD simulations to illustrate the efficacy of the newly developed mixing enhancement scheme and the control law; Section 4 gives a brief conclusion to the paper. To simplify our presentation, the proof of the main result is given in the Appendix.

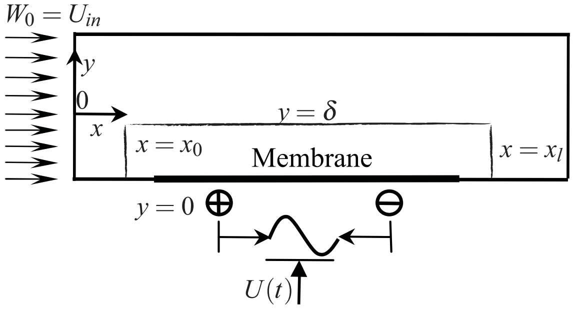


Fig. 1. The rectangular region close to the membrane surface.

## II. Optimal Feedback Control for Mixing Enhancement in the Boundary Layer of Membrane System

In this section, a special scheme is developed, which uses an external electric field to stir up the flow in the area adjacent to the membrane surface and thereby increase the mixing in this area. As mentioned above, different from the previous methods which increase mixing in the whole channel [11], the new method restricts the influence of the external electric field in a region adjacent to the boundary layer. The electrodes are installed outside the membrane and the electric field does not activate the bulk flow in the channel. This reduces energy consumption and brings economic advantages to engineering practice.

As shown in Fig.I, the system we consider in this paper contains a rectangular channel, a piece of membrane installed on the bottom wall and a pair of electrodes. The electrodes are installed to generate the required voltage. An ions solution is fed into the channel from its inlet. The purpose of this research is to develop a control algorithm which generates the voltage signal (voltage and waveform) applied to the electrodes.

This study aims to enhance the mixing of fluid flow in area adjacent to the membrane surface and thereby reducing fouling and concentration polarization on the membrane. Because the purpose of this research is to explore the effect of the electric field on the flow in the area, it is reasonable to assume that the membrane is impermeable in our simulation study. We also assume that the bulk flow in the channel is a laminar flow. In addition, we use the velocity gradient $\frac{\partial w_x}{\partial y}$ on the membrane surface as the measurements of our system to be controlled.

In the channel, fluid flow satisfies the Navier-Stokes equation and the continuity equation

$$\frac{\partial \mathbf{W}}{\partial t} + (\mathbf{W} \cdot \nabla)\mathbf{W} = -\frac{1}{\rho}\nabla P + \frac{\mu}{\rho}\Delta \mathbf{W}, \tag{1}$$

$$\operatorname{div}(\mathbf{W}) = 0. \tag{2}$$

We assume that a velocity field $(\overline{\mathbf{W}}, \overline{P}) = (\overline{W}_x, \overline{W}_y, \overline{W}_z, \overline{P})$ is a steady state solution of the equations (1) and (2) corresponding to fully developed laminar flow in the channel. The solution to the equations (1) and (2) can be obtained analytically. For example, $(6U_{in}(1-\frac{y^2}{h^2}), 0, 12\mu U_{in}\frac{L-x}{h^2})$ is a solution for the system in our simulation. Here, $U_{in}$ is the fluid velocity at the inlet of the channel, $\mu$ is the viscosity of the fluid, $h$ and $L$ are the height and the length of the channel. We can take these velocity components as the time-averaged values of the velocity field. Now, we define the perturbation variables

$$\mathbf{w} = (w_x, w_y, w_z) = \mathbf{W} - \bar{\mathbf{W}}, p = P - \bar{P}. \quad (3)$$

Substituting these variables into the equations (1) and (2), the Navier-Stokes equation and continuity equation become

$$\frac{\partial \mathbf{w}}{\partial t} + \left(\bar{\mathbf{W}} \cdot \nabla\right)\mathbf{w} + \left(\mathbf{w} \cdot \nabla\right)\bar{\mathbf{W}} + \left(\mathbf{w} \cdot \nabla\right)\mathbf{w} = -\frac{1}{\rho}\nabla p + \frac{\mu}{\rho}\Delta \mathbf{w}, \quad (4)$$

$$\operatorname{div}(\mathbf{w}) = 0. \quad (5)$$

in the domain $\Omega = \{(x,y,z) = [x_0,x_l] \times [-h/2, h/2]\}$ where $x_l - x_0$ is the length of the region we consider.

We define a rectangular prism $[x_0,x_l] \times [0,\delta]$, as shown in Fig. I, which contains all the flow being perturbed by the external electric field. Technically, it requires that the perturbation velocity components and the spatial gradients of the perturbation velocity field are all zeros on the surface $y = \delta$; i.e., the flow on and above this surface will not be perturbed by the electric field. Also, on the upstream and downstream sides of the rectangular prism, i.e., at $x = 0$ and $x = x_l$, the perturbation velocity components and the spatial gradients of the perturbation velocity field are also zero.

As the boundary condition on the membrane surface involves actuating the control law, it is worthy discussing the actuation scheme of the control at the first place. According to [5], [14], when an electric field is applied, the charges in the electric double layer induce fluid flow in the area adjacent to the membrane surface. As the boundary layer is very thin, the fluid flow on the membrane surface explains most effect of the electric field. Therefore, it is reasonable to assume that the external electric field only induces fluid flow on the membrane surface, rather than in $y$ direction. The induced flow velocity is called slip velocity $u_s$ and can be expressed as the product of electro-osmotic mobility $\mu_{EO}$ and the local electric field $E$ as $\overrightarrow{u}_s = \mu_{EO}\overrightarrow{E} = \frac{\varepsilon\zeta}{\mu}\overrightarrow{E}$ where $\zeta$ is the zeta potential, $\varepsilon$ is the permittivity, and $\mu$ is the viscosity of the fluid. In the context of this paper, the velocity $\overrightarrow{u}_s$ and the electric field $\overrightarrow{E}$ both take the $x$-axial direction as positive direction and the reverse direction as negative. Then, we write the slip velocity simply $u_s = \mu_{EO}E = \frac{\varepsilon\zeta}{\mu}E$. Therefore, the effect of the electric field on the fluid flow is transformed into a slip velocity on the membrane surface and the control problem becomes a boundary control problem.

Now, we can define the boundary condition for the equation (4). The equation (4) needs to satisfy the following boundary conditions on the rectangular prism $[x_0,x_l] \times [0,\delta]$

$x = x_0$: $w_x = 0$, $w_y = 0$, $\frac{\partial w_x}{\partial x} = 0$, $\frac{\partial w_x}{\partial y} = 0$, $\frac{\partial w_y}{\partial x} = 0$, $\frac{\partial w_y}{\partial y} = 0$;
$x = x_l$: ditto;
$y = \delta$: $w_x = 0$, $w_y = 0$, $\frac{\partial w_x}{\partial x} = 0$, $\frac{\partial w_x}{\partial y} = 0$, $\frac{\partial w_y}{\partial x} = 0$, $\frac{\partial w_y}{\partial y} = 0$;
$y = 0$: $w_x = u_s$, $w_y = 0$.

The boundary conditions on the surface $y = \delta$ are also the conditions for selecting $\delta$.

In this section, we assume that the measurement of the system to be controlled, $\frac{\partial w_x}{\partial y}|_{y=0}$ is known for constructing feedback control signal.

To facilitate the development of the new mixing enhancement approach, we define two concepts: perturbation kinetic energy and mixing index. The perturbation kinetic energy, which is equivalent to the turbulent kinetic energy as defined in [11] when turbulence is the main cause of mixing, is defined as

$$E(\mathbf{w}) = \frac{1}{2}\int_{\Omega} |\mathbf{w}|^2 dV = \frac{1}{2}\int_{x_0}^{x_l}\int_0^{\delta}\int_0^{z_r} \left(w_x^2 + w_y^2 + w_z^2\right) dxdydz. \quad (6)$$

The mixing in this paper is defined as a measure of the spatial gradients of the perturbation velocity field:

$$M(\mathbf{w}) = \int_{\Omega} |\nabla \mathbf{w}|^2 = \int_{\Omega} \mathrm{Tr}\{\nabla \mathbf{w}^T \nabla \mathbf{w}\} dV. \quad (7)$$

Previously, Ottino [9] defined a stretching mixing rates as

$$\frac{dL}{dt} = \int_{\Omega} Tr\left(\Phi^T \Phi\right) dV, \quad (8)$$

where $\Phi = \frac{1}{2}\left(\nabla \mathbf{w} + (\nabla \mathbf{w})^T\right)$.

Defining $\Psi = \frac{1}{2}\left(\nabla \mathbf{w} - (\nabla \mathbf{w})^T\right)$, then, the mixing estimation parameter in [10], which mainly reflects the extent of mixing caused by vortices, can be rewritten in the following form

$$\frac{d\Lambda}{dt} = \int_{\Omega} Tr\left(\Psi^T \Psi\right) dV. \quad (9)$$

It is obvious that (7) is the sum of (8) and (9):

$$M(\mathbf{w}) = \int_{\Omega} |\nabla \mathbf{w}|^2 dV = \int_{\Omega} \mathrm{Tr}\left(\Phi^T \Phi\right) dV + \int_{\Omega} \mathrm{Tr}\left(\Psi^T \Psi\right) dV$$

That is, the mixing index (7) describes the mixing caused by both length stretching and folding induced by perturbations.

The proposed mixing enhancement scheme uses electric field to induce perturbations to the boundary layer. Fluid stretching and vortices account for most of the mixing caused by the perturbations. Therefore, the new mixing index can be used to describe the extent of the mixing caused by the perturbation in the circumstance of this paper. It is worth pointing out that this mixing index describes the mixing enhancement due to the perturbation caused by EOF and it does not reflect mixing inherent in the steady-state system and related to the original velocity gradients and momentum diffusion.

The control actuation of the proposed scheme is implemented through a pair of fixed electrodes and this gives the electric field a special distribution. This makes the proposed methods distinct from the flow control scheme in the literatures, for example, [12]. Based on aforementioned actuation mechanism, the slip velocity

$$u_s = U(t)f(x), \quad (10)$$

where $U(t)$ is the voltage applied to the electrodes, which is generated from our control algorithm, and

$$f(x) = C_1 \left[ \frac{x - l_1}{(x - l_1)^2 + C_2} - \frac{x - l_2}{(x - l_2)^2 + C_2} \right]$$

describes the distribution of the electric field with the parameters $C_1 = \frac{\varepsilon_w \zeta}{2\mu \ln\left(\frac{l_{i,2} - l_{i,1} - r_c}{r_c}\right)}$, and $C_2 = (h_m + r_c + \Delta y)^2$. Here, $r_c$ is the radius of electrodes, $h_m$ is the thickness of membrane and $\Delta y$ is the distance between the electrode and the membrane outside. The other constants $\varepsilon_w$, $\zeta$, $\mu$ are the permittivity, zeta potential, viscosity, respectively. Define $F = \int_0^{x_l} f^2(x)dx$, then

$$\begin{aligned} F = & \int_0^{x_l} C_1^2 \left[ \frac{x - l_1}{(x - l_1)^2 + C_2} - \frac{x - l_2}{(x - l_2)^2 + C_2} \right]^2 dx \\ = & C_1^2 \int_0^{x_l} \left[ \frac{x - l_1}{(x - l_1)^2 + C_2} \right]^2 dx - 2C_1^2 \int_0^{x_l} \frac{x - l_1}{(x - l_1)^2 + C_2} \times \\ & \frac{x - l_2}{(x - l_2)^2 + C_2} dx + C_1^2 \int_0^{x_l} \left[ \frac{x - l_2}{(x - l_2)^2 + C_2} \right]^2 dx. \end{aligned} \tag{11}$$

Now, we calculate the term on the right side of the equation (11) one by one. First we calculate the first term on the right side of (11). From the fact that $\left[\frac{x-l_1}{(x-l_1)^2+C_2}\right]^2 = \left[\frac{Ax+B}{(x-l_1)^2+C_2}\right]' + \left[\frac{Cx+D}{(x-l_1)^2+C_2}\right]$ where $A = -\frac{1}{2}$, $B = \frac{1}{2}l_1$, $C = 0$, $D = \frac{1}{2}$ and '$'$' means derivative, it follows that $\int_0^{x_l} C_1^2 \left[\frac{x-l_1}{(x-l_1)^2+C_2}\right]^2 = \left[\frac{-1/2x+1/2}{(x-l_1)^2+C_2}\right]_0^{x_l} + \frac{1}{2\sqrt{C_2}} \arctan \frac{x}{2\sqrt{C_2}} \Big|_{-l_1}^{x_l - l_1} = -\frac{x_l - 1}{2(x_l - l_1)^2 + 2C_2} - \frac{1}{2(l_1^2 + C_2)} + \frac{1}{2\sqrt{C_2}} \arctan \frac{x_l - l_1}{2\sqrt{C_2}} - \frac{1}{2\sqrt{C_2}} \arctan \frac{-l_1}{2\sqrt{C_2}}$. In the same way, we have

$$\begin{aligned} & \int_0^{x_l} \left[ \frac{x - l_2}{(x - l_2)^2 + C_2} \right]^2 dx = \left[ \frac{-1/2x + 1/2}{(x - l_2)^2 + C_2} \right]_0^{x_l} \\ & + \frac{1}{2\sqrt{C_2}} \arctan \frac{x}{2\sqrt{C_2}} \Big|_{-l_2}^{x_l - l_2} \\ & = -\frac{x_l - 1}{2(x_l - l_2)^2 + 2C_2} - \frac{1}{2(l_2^2 + C_2)} + \frac{1}{2\sqrt{C_2}} \arctan \frac{x_l - l_2}{2\sqrt{C_2}} \\ & - \frac{1}{2\sqrt{C_2}} \arctan \frac{-l_2}{2\sqrt{C_2}}. \end{aligned}$$

The second term on the right side of (11) $\frac{x-l_1}{(x-l_1)^2+C_2} \times \frac{x-l_2}{(x-l_2)^2+C_2} = \frac{Ax+B}{(x-l_1)^2+C_2} + \frac{Cx+D}{(x-l_2)^2+C_2}$ where $A = 0$ and $B$, $C$, $D$ are the solution of the following linear equations

$$\begin{bmatrix} 2(l_2 - l_1) & 1 & 1 \\ l_1^2 - l_2^2 & -2l_2 & -2l_1 \\ 0 & l_2^2 + C_2 & l_1^2 + C_2 \end{bmatrix} \begin{bmatrix} C \\ B \\ D \end{bmatrix} = \begin{bmatrix} 2 \\ -2(l_1 + l_2) \\ 2l_1 l_2 \end{bmatrix}.$$

Then, we can calculate $\int_0^{x_l} \frac{x-l_1}{(x-l_1)^2+C_2} \times \frac{x-l_2}{(x-l_2)^2+C_2} dx = \frac{B}{\sqrt{C_2}} \arctan \frac{x}{2\sqrt{C_2}} \Big|_{-l_1}^{x_l - l_1} + C\frac{1}{2} \ln[(x - l_1)^2 + C_2]_0^{x_l} + (D - Cl_2)\frac{1}{C_2} \arctan \frac{x}{2\sqrt{C_2}} \Big|_{-l_2}^{x_l - l_2}$ and thus $F$ can be integrated analytically. From the above calculation, we can see that if the electrodes are installed, the distribution of the strength of the external electric field has been determined and $F$ is a constant; i.e., the voltage is the only design variable of the electric field.

Now, we define the main problem to be solved in this section:

**Problem A:** The optimal feedback control problem is defined as finding an appropriate $U(t)$ for the control law $u_s(x,t) = U(t)f(x)$, to maximize the following cost functional

$$\begin{aligned} J(u) = \lim_{t \to \infty} \Bigg[ & E(\mathbf{w}(t)) + \int_0^t \Bigg( \frac{\mu}{\rho} M(\mathbf{w}) + \Gamma(\mathbf{w}(t)) \\ & - \alpha \int_{\bar{\Omega}} u_s^2(\tau) dA - \frac{\mu^2}{4\rho^2 F \alpha} \left( \int_{\bar{\Omega}} f(x) \frac{\partial w_x}{\partial y}(\tau) dA \right)^2 \Bigg) d\tau \Bigg]. \end{aligned} \tag{12}$$

where $\bar{\Omega}$ is the surface $y = 0$, and $\alpha > 0$ is a constant related to the amplitude of $U(t)$, which is used to adjust the applied voltage.

In the cost functional (12), the first term describes the perturbation kinetic energy of the flow; the second term is the mixing index; the third term describes the stretching caused by laminar flow and its definition is given in the Appendix The rest two terms will be explained in the following.

The following theorem gives a solution to Problem A:

*Theorem 1:* Given a constant $\alpha > 0$, the control law

$$U(t) = -\frac{\mu}{2\rho \alpha F} \int_{\bar{\Omega}} f(x) \frac{\partial w_x}{\partial y} dA \tag{13}$$

solves Problem A. Also, the slip velocity

$$u_s(x,t) = -\frac{\mu}{2\rho \alpha F} f(x) \left( \int_{\bar{\Omega}} f(x) \frac{\partial w_x}{\partial y} \right) dA \tag{14}$$

is the boundary condition on the lower wall of the channel in the system (4) and (5).

The proof of this theorem is given in the Appendix.

Here, $U(t)$ is a continuous signal. The amplitude of this voltage signal is $U_A = \| \frac{\mu\lambda}{2\rho} \left( \int_{\bar{\Omega}} f(x) \frac{\partial w_x}{\partial y} \right) / F \|$. Here, the formula of $U(t)$ is not an explicit function of time $t$ but the system dynamics behind $\frac{\partial w_x}{\partial y}$ implies $U(t)$ is a function of $t$. The term $\frac{\partial w_x}{\partial y}$ itself is a function of time. In the area close to the membrane surface, the absolute perturbation velocity $|w_x|$ decreases in $y$ direction and hence the sign $\frac{\partial w_x}{\partial y}$ is opposite to that of $w_x$. Since the output is fed back to the control input, the output penalty works in conjunction with the input penalty to minimize control effort.

## III. Simulations and Mixing measurements

In this section, the fluid dynamics of the EOF in a membrane system are simulated using ANSYS CFX to validate the efficacy of the control feedback approach; i.e., to test the effect of mixing enhancement in the vicinity of the membrane.

In the simulation, we consider the $2D$ case and use a channel with $L = 0.11$m and height $h = 0.004$m. The electrodes are cylindrical and $h_m = 0.00025$m, $r_c = 0.005$m. The distance between the two electrodes is 0.015m.

Because the purpose of the simulation is to validate the mixing enhancement in the area adjacent to the membrane surface when the electric field is applied, we assume that the membrane is impermeable. In our simulation, we also used the following parameters for the system: $h_m = 0.00025$m, $\mu = 0.001\text{kg}\cdot\text{m}^{-1}\cdot\text{s}^{-1}$, $\rho = 1000\text{kg}\cdot\text{m}^{-3}$, $\zeta = 0.02\text{kg}\cdot\text{m}^2\cdot\text{s}^{-3}\cdot\text{A}^{-1}$, $r_c = 0.005$m, $\varepsilon_w = 7.0832\times10^{-10}\text{m}^{-3}\cdot\text{kg}^{-1}\cdot\text{s}^4\cdot\text{A}^2$. The feeding flow velocity to the channel is a typical velocity in membrane systems used in water treatment, $\bar{W}_x = 0.14\text{m}\cdot\text{s}^{-1}$. In our simulation, the constant is selected as $\alpha = 0.008$, the Reynolds number of the flow is 280, and the step time is $10^{-5}$s.

To give the system an initial perturbation, we first apply an oscillating voltage to the system, and then run the CFD simulation using the feedback control law given in this paper. The feedback is calculated using (13). As shown in Fig. 2, the simulation results show that the feedback oscillates around zero over time. The mixing extent caused by the electric field is measured by $M(\mathbf{w})$, which is the integral of spatial gradients of the perturbation velocity field over $\Omega$. As the electric field is the only perturbation in the channel, $\Omega$ is selected as to contain all the perturbations in the channel. Fig. 3 shows the mixing effect of the mixing enhancement scheme, in comparison with case without electric field ($M(\mathbf{w}) = 0$). The mixing index has a scale of $10^{-6}$) in Fig. 3. This is because the system itself has a scale of $10^{-6}$. As shown in Fig. 3, the mixing index also has oscillating features and this shows that the mixing is caused by the input voltage. Therefore, this illustrates the efficacy of the proposed mixing enhancement scheme and control law.

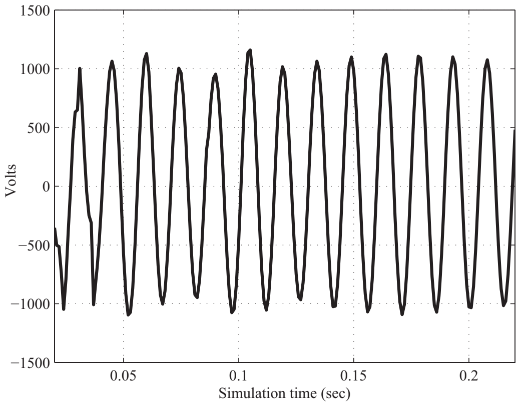


Fig. 2. The feedback control signal of the closed-loop system.

## IV. Conclusion

This paper has proposed a method for mixing enhancement in the vicinity of the membrane surface and increasing the productivity of the membrane system. A new mixing index has been defined and incorporated into the cost functional of an optimal control problem. This paper uses the integral of an electric field distribution function to handle the actuation problem due to fixed electric field distribution and distributive slip velocity. An optimal control problem has been defined in this paper and the control law given in this paper solves this optimal problem and maximizes the cost functional. A CFD simulation has been used to demonstrate the effect of the control law on mixing in the vicinity area adjacent to the membrane and has illustrated the efficacy of the proposed method. It also illustrates that the control law gives the desired waveform for such applications.

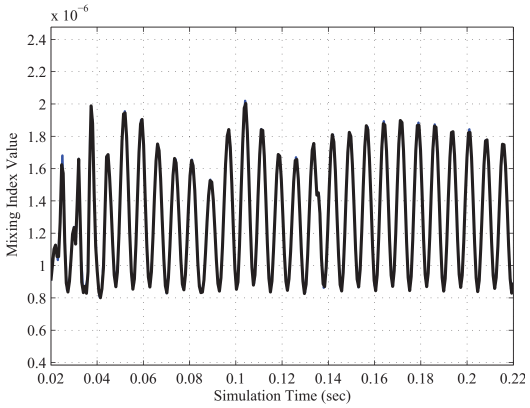


Fig. 3. The mixing index value when the control law (13) is applied to the system.

## Appendix

### A. Proof of Theorem 1:

The proof includes two parts: calculating the derivative of turbulent kinetic energy and verifying that the control maximizes the cost functional (12). We first consider the time derivative of the perturbation kinetic energy

$$\begin{aligned}\beta = \frac{d}{dt}E(\mathbf{w}) &= \int_\Omega \frac{\partial \mathbf{w}}{\partial t}\cdot\mathbf{w}dV\\ &= -\int_\Omega \left(\left(\bar{\mathbf{W}}\cdot\nabla\right)\mathbf{w} + \left(\mathbf{w}\cdot\nabla\right)\bar{\mathbf{W}}\right)\cdot\mathbf{w}dV - \int_\Omega \left(\mathbf{w}\cdot\nabla\right)\mathbf{w}\cdot\mathbf{w}dV\\ &\quad - \int_\Omega \frac{1}{\rho}\nabla p\cdot\mathbf{w}dV + \int_\Omega \frac{\mu}{\rho}\Delta\mathbf{w}\cdot\mathbf{w}dV. \qquad (15)\end{aligned}$$

Now, we consider the terms on the right side one by one.

$$\begin{aligned}-\int_\Omega \left(\mathbf{w}\cdot\nabla\right)\mathbf{w}\cdot\mathbf{w}dV &= -\frac{1}{2}\int_\Omega \nabla(\mathbf{w}\cdot\mathbf{w})\mathbf{w}dV\\ &= -\frac{1}{2}\int_\Omega \operatorname{div}\left((\mathbf{w}\cdot\mathbf{w})\mathbf{w}\right) = -\frac{1}{2}\int_{\partial\Omega}\left[(\mathbf{w}\cdot\mathbf{w})\mathbf{w}\right]\cdot\mathbf{n}dA.\end{aligned}$$

On the surfaces $x = x_0, x = x_l$ and $y = \delta$, $\mathbf{w} = [w_x; w_y] = 0$. This results that $[(\mathbf{w}\cdot\mathbf{w})\mathbf{w}] = 0$. As $\mathbf{w}$ is perpendicular to $\mathbf{n}$ on the surface $y = 0$, the right side of the above equality is equal to zero. Therefore, $-\int_\Omega (\mathbf{w}\cdot\nabla)\mathbf{w}\cdot\mathbf{w}dV = 0$.

By the divergence theorem of Gauss, $-\frac{1}{\rho}\int_\Omega \nabla p\cdot\mathbf{w}dV = -\frac{1}{\rho}\int_\Omega \operatorname{div}(p\mathbf{w})\,dV = -\frac{1}{\rho}\int_{\partial\Omega} p\mathbf{w}\cdot\mathbf{n}dA$. As $\mathbf{w} = 0$ on the surfaces $x = x_0, x = x_l$ and $y = \delta$ and the fact that $\mathbf{w}\cdot\mathbf{n} = 0$ on the surface $y = 0$, we have $-\frac{1}{\rho}\int_\Omega \nabla p\cdot\mathbf{w}dV = 0$.

Consider the fourth term on the right side of (15). Using the Einstein summation notation, $\int_\Omega \frac{\mu}{\rho}\Delta\mathbf{w}\cdot\mathbf{w}dV = -2\nabla w_i\cdot\nabla w_i - w_i(\Delta w_i)]dV = \frac{\mu}{\rho}\int_\Omega \frac{1}{2}\Delta(w_i w_i)dV - \frac{\mu}{\rho}\int_\Omega |\nabla w|^2 dV$.

The term $\frac{1}{2}\frac{\mu}{\rho}\int_\Omega \Delta(w_i w_i)dV = \frac{1}{2}\frac{\mu}{\rho}\int_{\partial\Omega}(\nabla|\mathbf{w}|^2)\cdot\mathbf{n}dA$. On $\partial\Omega$, as $\mathbf{w} = 0$ on the surfaces $x = x_0, x = x_l$ and $y = \delta$, we only need to consider the surface $y = 0$ where $\mathbf{n} = -j$. From the fact that $w_y|_{y=0} = 0$, it follows that

$$\frac{1}{2}\frac{\mu}{\rho}\int_{\Omega}\Delta(w_i w_i)dV$$
$$= -\frac{1}{2}\frac{\mu}{\rho}\int_0^{z_r}\int_{x_0}^{x_l} 2\left(w_x\frac{\partial w_x}{\partial y} + w_y\frac{\partial w_y}{\partial y}\right)_{y=0} dxdz$$
$$= -\frac{\mu}{\rho}\int_0^{z_r}\int_{x_0}^{x_l}\left(w_x\frac{\partial w_x}{\partial y}\right)_{y=0} dxdz.$$

Let $\Gamma = \int_{\Omega}((\bar{W}\cdot\nabla)w + (w\cdot\nabla)\bar{W})\cdot wdV$. Consider that $\bar{W}_y = \frac{\partial \bar{W}_y}{\partial y} = \frac{\partial \bar{W}_y}{\partial x} = 0$ and our problem is two dimensional, we have

$$\Gamma(\mathbf{w}) = \int_{\Omega}\left(\bar{W}_x\frac{\partial w_x}{\partial x}w_x + \bar{W}_x\frac{\partial w_y}{\partial x}w_y + w_y\frac{\partial \bar{W}_x}{\partial y}w_x\right)dV.$$

Then, we can conclude that

$$\frac{dE(\mathbf{w}(t))}{dt} = -\Gamma(\mathbf{w}) - \frac{\mu}{\rho}\int_0^{z_r}\int_0^{l} w_x\frac{\partial w_x}{\partial y}|_{y=0}dxdz - \frac{\mu}{\rho}M(\mathbf{w}).$$

Now, we substitute this result into the cost functional (12) and prove that the control law (13) maximizes this cost functional.

$$\frac{1}{2}\frac{\mu}{\rho}\int_{\Omega}\Delta(w_i w_i)dV$$
$$= -\frac{1}{2}\frac{\mu}{\rho}\int_0^{z_r}\int_{x_0}^{x_l} 2\left(w_x\frac{\partial w_x}{\partial y} + w_y\frac{\partial w_y}{\partial y}\right)_{y=0} dxdz$$
$$= -\frac{\mu}{\rho}\int_0^{z_r}\int_{x_0}^{x_l}\left(w_x\frac{\partial w_x}{\partial y}\right)_{y=0} dxdz.$$

Let $\Gamma = \int_{\Omega}((\bar{W}\cdot\nabla)w + (w\cdot\nabla)\bar{W})\cdot wdV$. Consider that $\bar{W}_y = \frac{\partial \bar{W}_y}{\partial y} = \frac{\partial \bar{W}_y}{\partial x} = 0$ and our problem is two dimensional, we have

$$\Gamma(\mathbf{w}) = \int_{\Omega}\left(\bar{W}_x\frac{\partial w_x}{\partial x}w_x + \bar{W}_x\frac{\partial w_y}{\partial x}w_y + w_y\frac{\partial \bar{W}_x}{\partial y}w_x\right)dV.$$

Then, we can conclude that

$$\frac{dE(\mathbf{w}(t))}{dt} = -\Gamma(\mathbf{w}) - \frac{\mu}{\rho}\int_0^{z_r}\int_0^{l} w_x\frac{\partial w_x}{\partial y}|_{y=0}dxdz - \frac{\mu}{\rho}M(\mathbf{w}).$$

Now, we substitute this result into the cost functional (12) and prove that the control law (13) maximizes this cost functional.

$$J(u_s) = \lim_{t\to\infty}\left[E(\mathbf{w}(t)) + \int_0^t\left(-\frac{dE(\mathbf{w}(t))}{d\tau} - \Gamma(\mathbf{w}(t))\right.\right.$$
$$-\frac{\mu}{\rho}\int_{\bar{\Omega}}u_s\frac{\partial w_x}{\partial y}|_{y=0}dA + \Gamma(\mathbf{w}(t)) - \alpha\int_{\bar{\Omega}}u_s^2(\tau)dA$$
$$\left.\left.-\frac{\mu^2}{4\rho^2F\alpha}\left(\int_{\bar{\Omega}}f(x)\frac{\partial w_x}{\partial y}(\tau)|_{y=0}dA\right)\right)^2 d\tau\right]$$
$$= E(\mathbf{w}(0)) + \lim_{t\to\infty}\left[\int_0^t\int_{\bar{\Omega}}\left(-\frac{\mu}{\rho}u_s(\tau)\frac{\partial w_x}{\partial y}|_{y=0}dA\right.\right.$$
$$-\alpha\int_{\bar{\Omega}}u_s^2(\tau)dA$$
$$\left.\left.-\frac{\mu^2}{4\rho^2F\alpha}\left(\int_{\bar{\Omega}}f(x)\frac{\partial w_x}{\partial y}(\tau)|_{y=0}\right)dA\right)^2 d\tau\right].$$

Furthermore, we have

$$J(u_s) = E(\mathbf{w}(0)) + \lim_{t\to\infty}\left[\int_0^t\left(\frac{\mu}{\rho}\int_{\bar{\Omega}}-U(t)f(x)\frac{\partial w_x}{\partial y}|_{y=0}dA\right.\right.$$
$$\left.\left.-\alpha U(t)^2F - \frac{\mu^2}{4\rho^2F\alpha}\left(\int_{\bar{\Omega}}f(x)\frac{\partial w_x}{\partial y}(\tau)|_{y=0}dA\right)^2\right)d\tau\right]$$
$$= E(\mathbf{w}(0)) + \lim_{t\to\infty}\alpha F\int_0^t\left[-\frac{U(t)}{\alpha F}\left(\frac{\mu}{\rho}\int_{\bar{\Omega}}f(x)\frac{\partial w_x}{\partial y}|_{y=0}\right)dA\right.$$
$$\left.-U(t)^2 - \frac{\mu^2}{4\rho^2\alpha^2F^2}\left(\int_{\bar{\Omega}}f(x)\frac{\partial w_x}{\partial y}(\tau)|_{y=0}dA\right)^2\right]d\tau$$
$$= E(\mathbf{w}(0))$$
$$-\lim_{t\to\infty}F\int_0^t\left(U(t) + \frac{\mu}{\rho}\frac{1}{2\alpha F}\int_{\bar{\Omega}}f(x)\frac{\partial w_x}{\partial y}|_{y=0}dA\right)^2 d\tau. \quad (16)$$

When $U(t) = -\frac{\mu}{2\rho\alpha F}\int_{\bar{\Omega}}f(x)\frac{\partial w_x}{\partial y}|_{y=0}dA$, the integral in (16) is zero. Then, the maximum of (12) is achieved. Therefore, (14) holds. This completes the proof. □

## References

[1] D. Myers, *Surfaces, interfaces, and colloids*. Wiley Online Library, 1999.

[2] M. Mänttäri, A. Pihlajamäki, E. Kaipainen, and M. Nyström, "Effect of temperature and membrane pre-treatment by pressure on the filtration properties of nanofiltration membranes," *Desalination*, vol. 145, no. 1-3, pp. 81–86, 2002.

[3] H. Ma, C. Bowman, and R. Davis, "Membrane fouling reduction by backpulsing and surface modification," *Journal of Membrane Science*, vol. 173, no. 2, pp. 191–200, 2000.

[4] S. Qian and H. Bau, "Theoretical investigation of electro-osmotic flows and chaotic stirring in rectangular cavities," *Applied mathematical modelling*, vol. 29, no. 8, pp. 726–753, 2005.

[5] M. Oddy, J. Santiago, and J. Mikkelsen, "Electrokinetic instability micromixing," *Analytical Chemistry*, vol. 73, no. 24, pp. 5822–5832, 2001.

[6] S. Shin, I. Kang, and Y. Cho, "Mixing enhancement by using electrokinetic instability under time-periodic electric field," *Journal of Micromechanics and Microengineering*, vol. 15, p. 455, 2005.

[7] A. Rawool and S. Mitra, "Numerical simulation of electroosmotic effect in serpentine channels," *Microfluidics and Nanofluidics*, vol. 2, no. 3, pp. 261–269, 2006.

[8] A. Stroock, M. Weck, D. Chiu, W. Huck, P. Kenis, R. Ismagilov, and G. Whitesides, "Patterning electro-osmotic flow with patterned surface charge," *Physical review letters*, vol. 84, no. 15, pp. 3314–3317, 2000.

[9] J. Ottino, *The kinematics of mixing: stretching, chaos, and transport*. Cambridge Univ Pr, 1989, vol. 3.

[10] A. Alexiadis, D. Wiley, D. Fletcher, and J. Bao, "Laminar flow transitions in a 2d channel with circular spacers," *Industrial & engineering chemistry research*, vol. 46, no. 16, pp. 5387–5396, 2007.

[11] O. Aamo and M. Krstić, *Flow control by feedback: stabilization and mixing*. Springer Verlag, 2003.

[12] A. Balogh, O. Aamo, and M. Krstic, "Optimal mixing enhancement in 3-d pipe flow," *Control Systems Technology, IEEE Transactions on*, vol. 13, no. 1, pp. 27–41, 2005.

[13] M. R. Spiegel, *Schaum's outline of theory and problems of vector analysis and an introduction to tensor analysis*. Schaum's Outline Series, 1959.

[14] W. Wu, P. Selvaganapathy, and C. Ching, "Transport of particles and microorganisms in microfluidic channels using rectified ac electro-osmotic flow," *Biomicrofluidics*, vol. 5, p. 013407, 2011.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# *Adaptive Feed-Forward and Feedback Control for Oxygen Ratio in Fuel Cell Stacks*

Omar Ragb, D L YU and JB Gomm
Control system research group, School of Engineering
Liverpool John Moore University
Liverpool, UK
om.khazali_yo@yahoo.com

*Abstract*—**Automatic control of fuel cell stacks (FCS) using non-adaptive and adaptive radial basis function (RBF) neural network methods are investigated in this paper. The neural network RBF inverse model is used to estimate the compressor voltage for fuel cell stack control at different current demands, reduction in the compressor gain (30% and 20%) and manifold leak (15%) in order to prevent the oxygen starvation. A PID controller is used in the feedback to adjust the difference between the requested and the actual oxygen ratio by compensating the neural network inverse model output. This method is designed and conducted in three stages, starting with the collection of data from the available fuel cell stack model and finished with the non adaptive and adaptive RBF neural network control. RBF neural networks with the K-means and P-nearest Neighbour's training algorithms are used for the investigation. Furthermore, the RBF inverse model is made adaptive to cope with the significant parameter uncertainty, disturbances and environment changes. Simulation results show the effectiveness of the adaptive control strategy.**

***Keywords-Fuel cell stacks; Non-adaptive; Adaptive; Radial Basis Function Neural Network; Feed-forward; Feedback oxygen starvation.***

## I. Introduction

Burning current natural sources causes many environment problems today. A lot of harmful gases, such as $CO_2$, rise in the environment as a result of burning fossil fuels and destructs the ozone layer, which leads to climatic change and what is known as the greenhouse effect. To recover this problem, the world has been looking for energy sources that are clean and safe on the environment. Fuel cells are a kind of clean and safe energy source on the environment. Polymer electrolyte membrane (PEM) fuel cells emerge as one of the most clean and promising alternatives to reduce fossil fuel dependency [1]. In the last years many researchers have presented some methods to control the fuel cell stacks, in order to prevent the oxygen starvation and improve the fuel cell control, which are now reviewed. Sedighizadeh M. [2] discussed the application of wavelet networks in the implementation of adaptive controllers for PEMFC's. Jiang Z. et al. [3] presented an adaptive control strategy for active power sharing in the hybrid power source. This control strategy is able to adjust the output current set-point of the fuel cell according to the state of the charge (or voltage) of the battery. An adaptive MPPT controller using the extrermum-seeking algorithm [4] is used to automatically keep the fuel cell working at maximum power point (MPP) all the time. Fiacchini M. et al. [5] is presented an adaptive control scheme for the safe operation of a fuel cell system. In particular, the aim of control action is to avoid that the oxygen ratio reaches dangerous values. In this paper adaptive and non-adaptive control methods are implemented to achieve better control for the fuel cell breathing. Furthermore, in this paper, we first explain the fuel cell working principles followed by description of the dynamic model of fuel cell stacks. We then formulate the RBF adaptive and non adaptive model. Finally we demonstrate simulation results for the fuel cell control with adaptive and non adaptive controllers.

## II. Fuel cell Dynamics

### *A. Fuel Cell Working and prenceples*

Fuel cells consume a hydrogen fuel (on the anode side) and oxygen (on the cathode side) and produce electric energy with water and some heat through a chemical reaction [1], to satisfy different power requirements (Fig. 1). Generally, the reactants flow in and reaction products flow out while the electrolyte remains in the cell. Fuel cells differ from batteries in that they do not need recharging, they operate quietly and efficiently, and when hydrogen is used as fuel they generate only electric power and drinking water. So, they are called zero emission engines. William Grove discovered the basic operating principle of fuel cells by reversing water in 1839 [6]. In particular, proton exchange membrane fuel cells (PEM-FCs), also known as polymer electrolyte membrane fuel cells, are considered to be more developed than other fuel cells technologies, because they have high power density, solid electrolyte, operate at low temperature, long cell and stack life and low corrosion [6]. The PEM-FC takes its name from the special plastic membrane used as the electrolyte. This membrane electrode assembly (MEA), not thicker than a few hundred microns, is the heart of a PEM-FC and, when supplied with fuel and air, generates electric power at cell voltages around 0.7 Volt and power densities of up to about 1 W/cm electrode area. Fig. 2 shows a schematic of a PEM-FCS and MEA. The MEA is typically located between a pair of current collector plates (platinum-impregnated porous electrodes) with machined flow fields for distributing fuel and oxidant to the

Omar AL-Mukhtar University, Libya

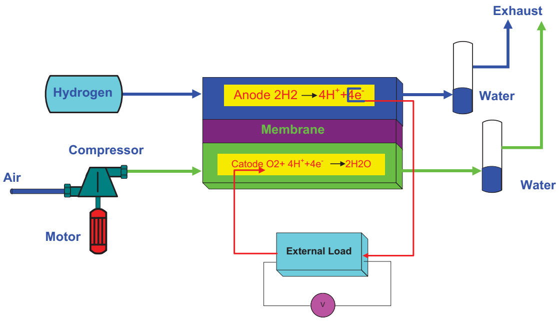


Figure 1. PME- FC reaction and structure

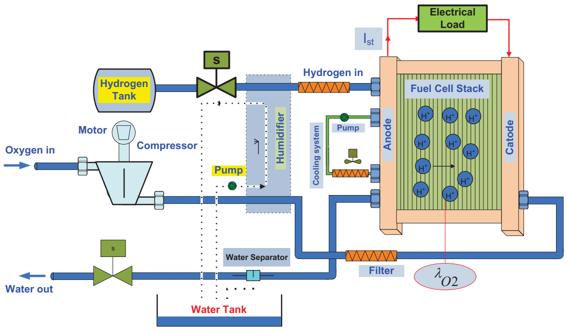


Figure 2. PEM fuel cell structure

anode and cathode, respectively. A water jacket for cooling is often placed at the back of each reactant flow field followed by a metallic current collector plate. The cell can also contain a humidification section for the reactant gases, which are kept close to their saturation level in order to prevent dehydration of the membrane electrolyte. Many FCs are connected electrically in series (Fig. 2) to form an FC stack (FCS).

### B. *Fuel Cell Stack Model*

The fuel cell stack (FCS) model simulated in this paper consists of four interacting sub-models (Fig. 3) which are the stack voltage, the anode flow, the cathode flow, and the membrane hydration models [6]. The voltage model contains an equation to calculate stack voltage based on fuel cell temperature, pressure, reactant gas partial pressures and membrane humidity, in summary, the fuel cell voltage *E* is given by

$$E = 1.2229 - 0.85\times10^{-3}(T_{fc} - 29815) + 43085\times10^{-5}T_{fc}(p_{H2}) + \frac{1}{2}p_{O2} \quad (1)$$

where, $T_{fc}$ is the fuel cell temperature in Kelvin, $p_{H2}$ and $p_{O2}$ are the partial pressures of hydrogen and oxygen respectively, details in [1,6]. In this model the stack temperature is assumed to be constant at 80°C. The model which is used in our investigations is given in [6]. The FCS Simulink model is created in Matlab 6.5.

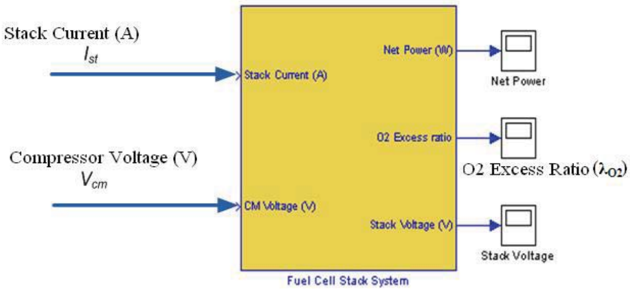


Figure 3. Simulink model of integrated PEM fuel cell

## III. FEED-FORWARD CONTROL DESIGN BASED ON NN

### A. *NN Inverse Model*

The radial basis function neural network (RBFNN) has an ability to model any non-linear function. However, this kind of neural network can need many nodes to achieve the required approximating properties [7]. The first step in the fuel cell modelling is the generation of a suitable training data set. The accuracy of the neural network modelling performance will be influenced by the training data. In the fuel cell stack data collection, the training data must be representative fuel cell behavior in order to analyze the performance of RBF fuel cell models in practical operating conditions. This means that input and output signals should sufficiently cover the region in which the system is going to be controlled [8]. As shown in Fig. 3, the fuel cell stack used for this research has to inputs compressor voltage $v_{cm}$ and the load current $I_{st}$, has three outputs (stack voltage *SV*, net power *NP* and oxygen ratio $y=O_2$). A set of random amplitude signals (RAS) were designed (0~3000 samples) for the fuel cell current load demand ($I_{st}$) and the compressor voltage ($v_{cm}$) to obtain a representative set of input data. The RASs of the current load demand and fuel cell compressor voltage were bounded between 100 and 300 Ampere for the current and between 100 and 235 volts for the compressor voltage see table I. Appling these two random input signals on the fuel cell model produces three outputs which are oxygen ratio (*y*), stack voltage (*SV*) and net power (*NP*) see table II and Fig. 4.

TABLE I. RAS INPUTS SIGNAL FOR $\hat{v}_{cm}$ MODELLING

| Parameters | Minimum | Maximum |
|---|---|---|
| $V_{cm}$ | 100 Volts | 235 Volts |
| $I_{st}$ | 100 | 300 |

TABLE II. OUTPUTS SIGNAL FOR $\hat{v}_{cm}$ MODELLING

| Parameters | Minimum | Maximum |
|---|---|---|
| $y=O_2$ | 0.7051 | 5.24 |
| SV | 112.64 Volts | 282.79 Volts |
| PN | 3850 | 6630 |

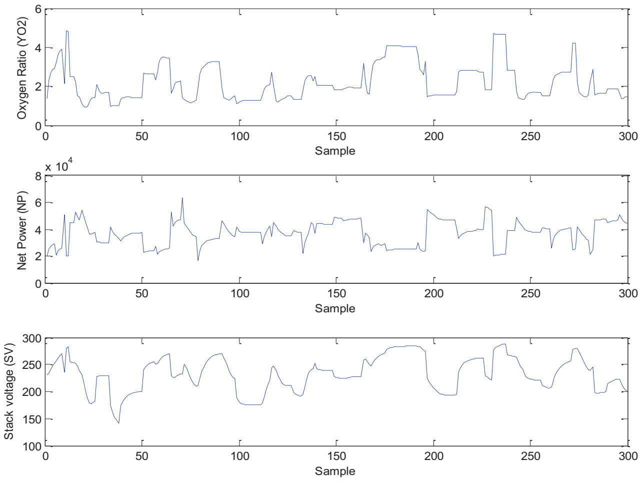


Figure 4. Fuel cell stack parameters output after applied RAS of $I_{st}$ & $v_{cm}$

For RBF neural network training, the K-means algorithm is used to choose the centers, $\rho$-nearast neighbor algorithm decides the widths and the recursive training algorithm [9] calculates the weights for the output layer. Here, a RBFNN based inverse model is used to predict the compressor voltage $\hat{v}_{cm}(k)$ which is the manipulated variable in the next sample time.The RBFNN block diagram is illustrated in Fig. 5, where the RBFNN input at sample $k$ is a vector $x(k)$ it also given by the following equation:

$$x(k) = [v_{cm}(k-1)\ y(k-1)\ I_{st}(k)\ I_{st}(k-1)\ SV(k-1)$$

$$PN(k-1)\ PN(k-2)\ PN(k-3)]^T \tag{2}$$

The neural network output is:

$$\hat{v}_{cm}(k) = g[v_{cm}(k-1)\, y(k-1)\, I_{st}(k)\, I_{st}(k-1)$$

$$SV(k-1)\, N(k-1)\, PN(k-2)\, PN(k-3)] \tag{3}$$

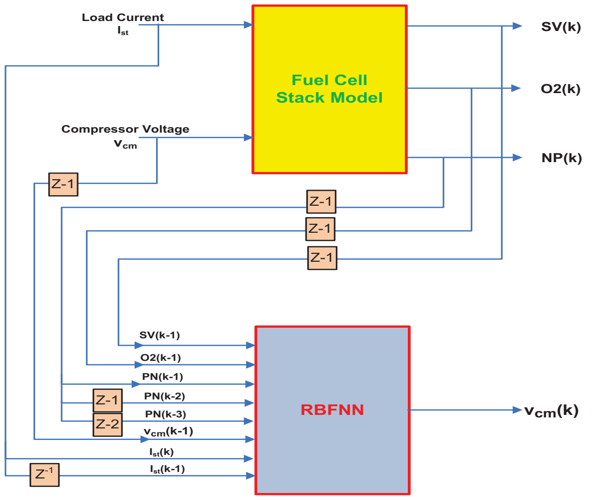


Figure 5. The RBFNN block diagram with Input variables

where g(.) is the nonlinear neural network function and $\hat{v}_{cm}(k)$ is the estimated compressor voltage. In order to train this neural network model, RASs were applied to the two fuel cell stack inputs $I_{st}$ and $v_{cm}$ and data for the fuel cell were collected for $O_2$, $SV$ and $PN$ at each sample time. The raw data were scaled using the following equation before training:

$$xscale(k) = \frac{x(k) - \min\{x(i)\}}{\max\{x(i)\} - \min\{x(i)\}} \qquad i \in [1, N]$$

The training data set with 2000 samples are used to train the RBFNN model. Then, the test set with 1000 is applied to the trained model and the model output prediction results are displayed in Fig. 6. The mean absolute error (MAE) is used to evaluate the modelling and control performance in this research, which is given by the following equation:

$$MAE = \frac{1}{N}\sum_{k=1}^{N}\left|\hat{v}_{cm}(k) - v_{cm}(k)\right| = \frac{1}{N}\sum_{k=1}^{N}\left|e(k)\right| \tag{5}$$

where $\hat{v}_{cm}(k)$ is the prediction by the inverse neural network model and $v_{cm}$ is the compressor voltage. The $\hat{v}_{cm}(k)$ in Fig. 6 is the normalized value and the MAE is 0.0142. The output of the neural network is nearly equal to the actual compressor voltage input. This is because $y(k-1)$, which can be calculated online at sample time $k$, was also used to predict the value of compressor voltage $\hat{v}_{cm}(k)$. Also this inverse RBFNN model can predict the required $\hat{v}_{cm}(k)$ for one sample step. The $\hat{v}_{cm}(k)$ can be calculated according to (3).

### B. *Non-Adaptive FF and FB Control Scheme*

The RBFNN-based non-adaptive feed-forward with feedback control system structure in our implementation is shown in Fig. 7. After trained the RBFNN inverse model and we got satisfy results. So, all the Recursive Least Square parameters ($w(0)$,µ and $\rho(0)$ will be saved in order to use them in non-adaptive and adaptive feed-forward controllers, then we will use this model in the feed-forward path to predict the scaled compressor voltage $\hat{v}_{cm}(k)$ and is given by (6).

$$\hat{v}_{cm}(k) = \phi^T \times w \tag{6}$$

So, to get the non scaled compressor voltage the (7) should be applied:

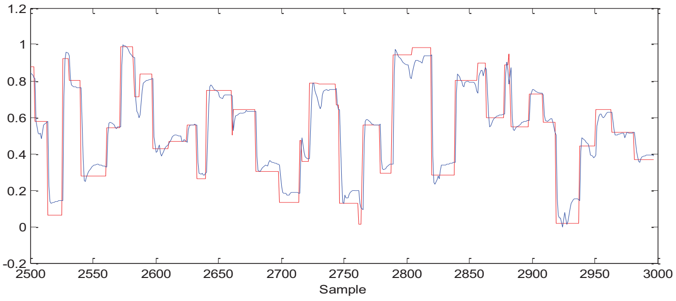


Figure 6. $v_{cm}$ Validation data for RBFNN model MAE (Mean Absolute Errors=0.0142)

$$v_{cm}(k) = \min v_{cm} + \hat{v}_{cm}(k) + \max v_{cm} - \min v_{cm}) \qquad (7)$$

On the other hand, to enhance the performance in steady state, the PID controller is added to form the feedback controller. In this case the activating compressor voltage is the sum of two controller outputs variables, one is from the RBF based feed-forward neural network controller, the other from the feedback PID controller. The current demand changing during the control is shown in Fig. 8. The following digital PID controller equation is used in [9]:

$$v_{cm\,2}(k) = y(k-1) + K_p\left[\left(1+\frac{T}{T_i}+\frac{T_d}{T}\right)e(k) - \left(1+\frac{2T_d}{T}\right)e(k-1) + \frac{Td}{T}e(k-2)\right] \quad (8)$$

After fine tuning, the PID controller that is used here with RBF based neural network controller for oxygen ratio regulation is

$$\hat{v}_{cm2}(k) = y(k-1) + 585e(k) - 45e(k-1)2.2\times10^{-4}e(k-2) \quad (9)$$

Here the sampling time is chosen to be 0.1sec. The measured oxygen ratio with time delay is the feedback signal of system.

### C. *Adaptive FF and FB Control Scheme*

The different of the strategy of adaptive RBFNN based FF control of oxygen ratio is the widths and the weights will be updated at each sample time. The RBF-NN based adaptive feed-forward and feedback control system structure is illustrated in Fig. 9. After training the RBFNN inverse model and we got satisfy good match. The PID controller is added to form the feedback control as discussed in the non-adaptive control.

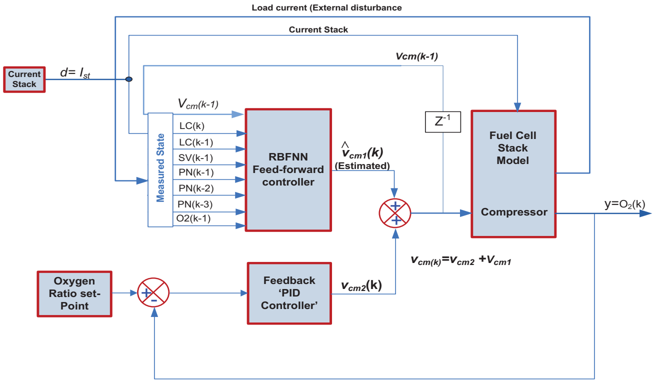


Figure 7. RBFNN FF and PID controller on the FCs

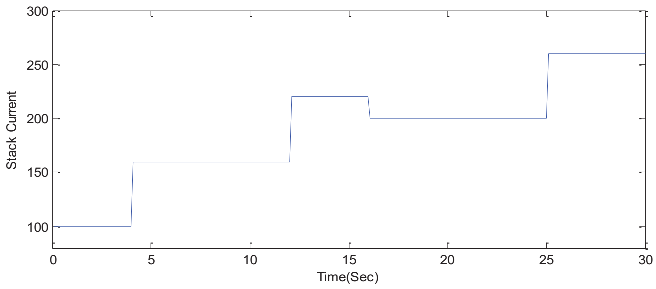


Figure 8. Fuel cell current demand changing during the control

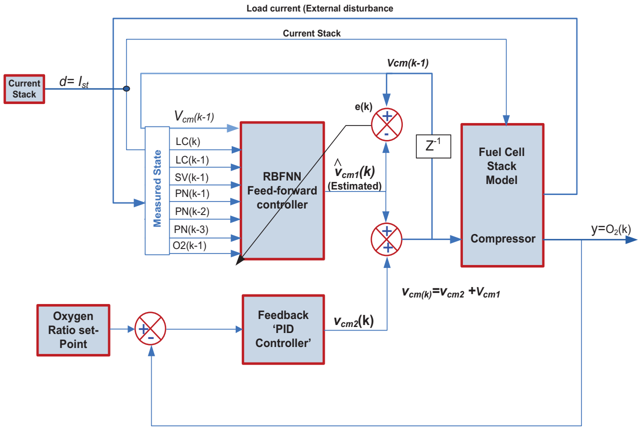


Figure 9. RBFNN FF and PID Controller on the Fuel Cell Stack

## IV. SIMULATION RESULTS

The output responses simulation of the oxygen ratio for non-adaptive and adaptive RBFNN feed-forward control for fuel cell stack are illustrated in fig. 10 and 11, the significant a difference can be seen between the performances of the two controllers. The tracking MAE of oxygen for non-adaptive and adaptive are 0.0045 and 0.0036 respectively.

### A. *Control Performance with 30% and 20% Reduction in the Compressor Gain*

The compressor is a machine to press the air inside the cathode but in some cases there are some problems associated with the compressor, which leads to reduce efficiency of the compressor.

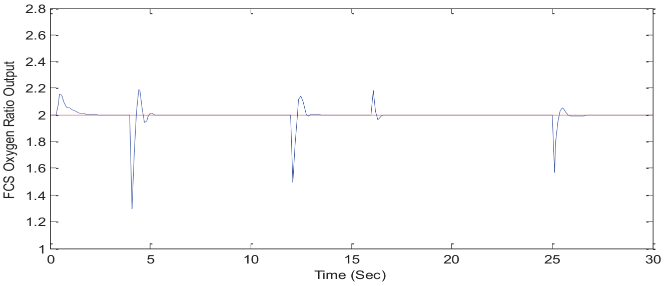


Figure 10. Simulation Result of Non-Adaptive RBF-Based FF & FB Control on FCS Oxygen Ratio

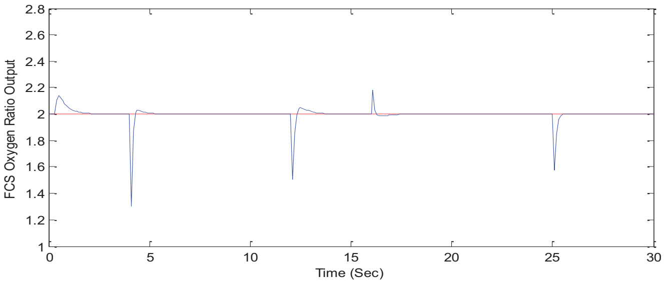


Figure 11. Simulation Result of Adaptive RBF-Based FF & FB Control on FCS Oxygen Ratio

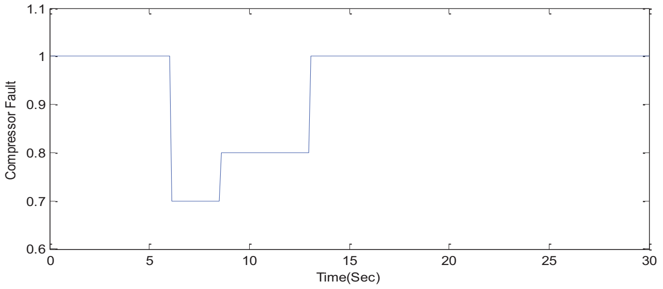


Figure 12. FCS Compressor Efficiency Fault Simulation

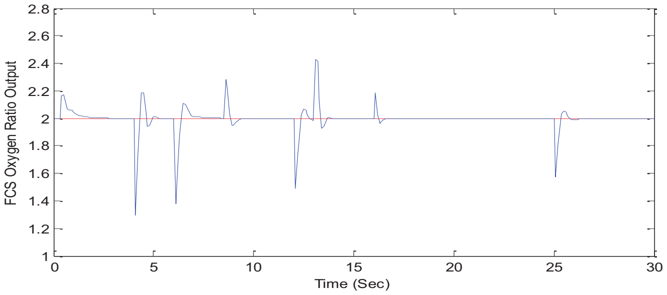


Figure 13. Simulation Result of Non-Adaptive FF and FB Controller on $O_2$ with 30~ 20% Reduction in Compressor Gain

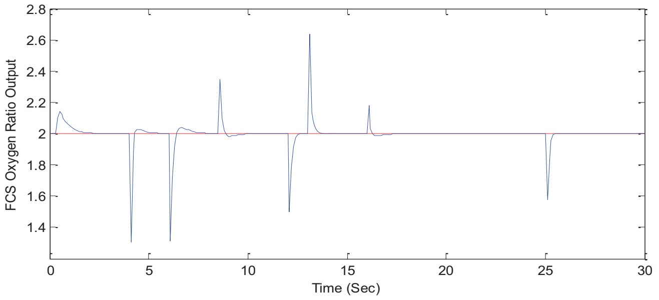


Figure 14. Simulation of Adaptive FF and FB Control on $O_2$ Ration With 30~20% Reduction In Compressor Gain

So, it needs regular maintenance because it contains gears and movement parts with oil circulation. When there is a problem in any part will affect on the efficiency of the compressor. The adaptive and non-adaptive RBFNN feed-forward controllers are evaluated with the current signal demand (Fig. 8) and the compressor gain reduction (Fig. 12). After reducing the compressor gain by 30% and 20% respectively, which will decrease the value of $O_2$ flow into the FCS port, the non-adaptive RBFNN model has a little capability to deal with this problem because it cannot retune itself according to this error, as a result it can't overcome this change and its MAE is 0.0083 (Fig. 13). However the adaptive RBF has the capability to retune itself to cope this situation and hence, achieves an improved control performance with a mean absolute error of 0.0075 (Fig. 14).

### B. *Control Performance with Compressor Gain Reduction and Manifold Leak*

Manifold leak is called component fault and to represent the air leakage fault, the manifold pressure equation in [6] is modified to (10):

$$\frac{dm_{sm}}{dt} = W_{cp} - W_{sm} - \Delta L \tag{10}$$

Where, $W_{cp}$ is the inlet mass flow (compressor flow), $W_{sm}$ is the supply manifold outlet mass flow and the added term $\Delta L$ is used to simulate the leakage from the supply air manifold, which decrease the air outflow from the manifold. $\Delta L = 0$ represents no air leak in the intake manifold. The air leakage level is simulated as 15% of total air intake in manifold as shown in Fig.15, and was simulated by changing the Simulink model of the FCS. From the oxygen ratio output shown in Fig. 16, the performance of non adaptive RBF FFC is acceptable and the MAE is 0.0126%. However, the performance of the adaptive RBF FFC as shown in Fig. 17 is better than the non adaptive. The adaptive RBF FFC can handle the manifold leak and therefore, achieves an improved control performance with mean MAE equal to 0.0090.

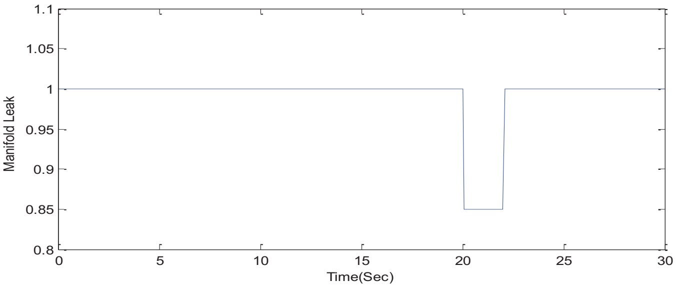


Figure 15. Manifold leak error simulate

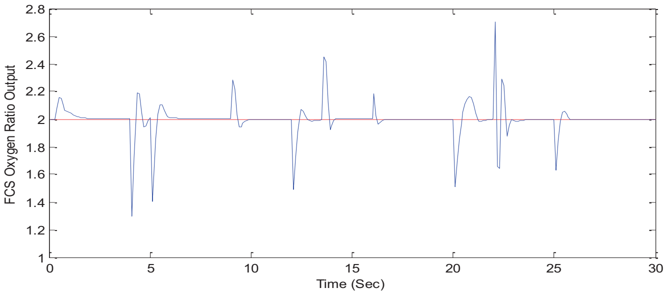


Figure 16. Oxygen ratio control result of the non adaptive RBF FFC with compressor gain reduction and leak

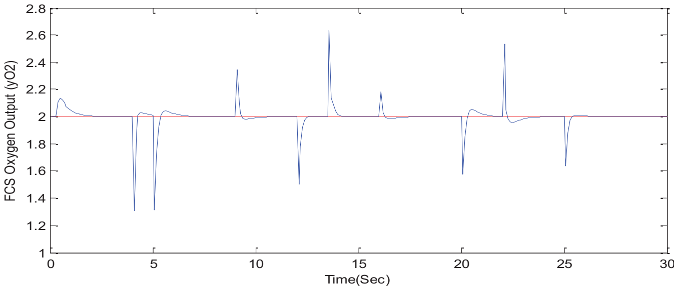


Figure 17. Oxygen ratio control result of the adaptive RBF FFC with compressor gain reduction and leak

TABLE III. EVALUATED

| Type of control method | Mean Absolute Error MAE |
|---|---|
| Non-adaptive RBFNN Control+PID controller | 0.0045 |
| Adaptive RBFNN +PID controller | 0.0036 |
| Non-adaptive RBFNN Control+PID controller with compressor gain reduction 30% and 20% | 0.0087 |
| Adaptive RBFNN Control+PID controller with compressor gain reduction 30% and 20% | 0.0075 |
| Non adaptive RBFNN Control+PID controller with compressor gain reduction 30% and 20% and the manifold leak 15% | 0.0126 |
| Adaptive RBFNN Control+PID controller with compressor gain reduction 30% and 20% and the manifold leak 15% | 0.0090 |

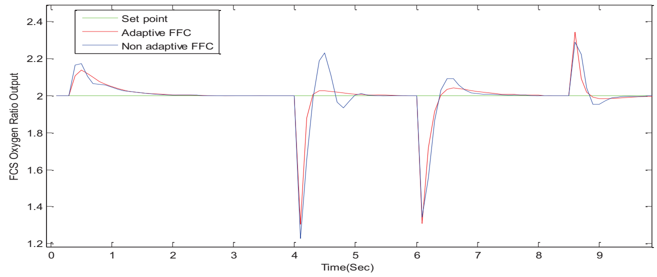


Figure 18. Comparing Adaptive and Non-Adaptive Control Performance when there is gain reduction

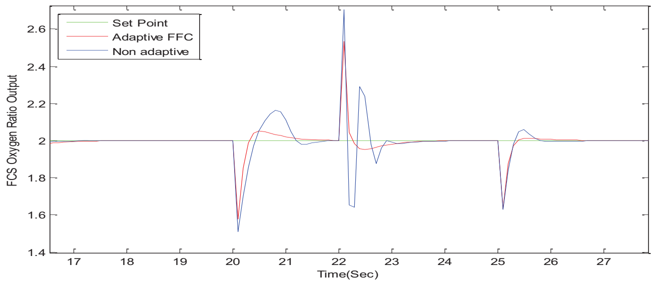


Figure 19. Comparing Adaptive and Non-Adaptive Control Performance when there are gain reduction and manifold leak

## V. DISCUSSION

From the Fig. 10 and 11 and the table III, when there is no reduction in the compressor gain, the RBFNN adaptive can give good performance than non-adaptive FFC with MAE equal to 0.0036 and 0.0045 respectively. However, reducing the compressor gain by 30% and 20% respectively, which will decrease the value of $O_2$ flow into the FCS port, the non-adaptive RBF controller has little capacity to deal with this situation because it was trained off-line and is fixed. As result it was unable to cope with an environment change see Fig. 13. However, the adaptive RBFNN can adjust the compressor voltage according to the real-time condition of the FCS and compensate for the influence of this system error after several sample times see Fig. 14. Fig. 18 show the comparison on oxygen ratio rates between Fig. 13 and 14. After that, the performance of the non adaptive and adaptive RBF FFC is evaluated with manifold leak Fig. 15 additional to current disturbance Fig. 8 and compressor gain reduction Fig. 12, the simulation results are as given in Fig. 16 and 17. The adaptive performance also, has more capacity to overcome the environmental change than the non adaptive because it was trained on-line according to the error see table III and Fig. 19.

## VI. CONCLUSION AND FUTURE WORK

This paper presents an adaptive and non-adaptive RBF control strategy to estimate the compressor voltage for fuel cell stack control to prevent fuel cell oxygen starvation and compressor surge during rapid load demands, gain reduction (30% and 20%) and manifold leak 15%. APID controller is used in the feedback to adjust the difference between the requested and the actual oxygen ratio by compensating the neural network inverse model output. When there is no reduction in the compressor gain the control strategy can adjust the compressor voltage of the fuel cell according to the current demand. From the other hand, when there is reduction (30% and 20%) in the compressor gain and 15% manifold leak. The non-adaptive RBF controller less ability than adaptive RBF controller because the adaptive was trained off-line, as a result it was unable to cope with an environmental changes. Furthermore, the adaptive RBFNN is able to adjust the compressor voltage to adapt to the real-time condition of the FCS and compensate for the influence of this system error after several sample times. The simulation results show the effectiveness of the adaptive control strategy.

## REFERENCES

[1] J. T. Pukrushpan, A. G. Stefanopoulou, and H. Peng, “Control of fuel cell breathing,” IEEE ControlSystems Magazine, vol. 24, no. 2, pp. 30-46, April 2004.

[2] M. Sedighizadeh, and A. Rezazadeh, “A Neuro adaptive control Strategy for movable power source of proton exchange membrane fuel cell using wavelets,” World Academy of Science, Engineering and Technology, vol. 36, pp. 285 -289, 2007.

[3] Z. Jiang, L. Gao, and A. Roger, “Adaptive control strategy for active power sharing in hybrid fuel cell/battery power sources,” IEEE Transactions on Energy Conversion, vol. 22, No. 2, pp. 507-515, 2007.

[4] Z. Zhong, H. Hai-bo, X. Zhu, C. Guang-yi and R. Yuan, “Adaptive maximum power point tracking control of fuel cell power plants,” Elsever Journal of Power Sources, vol. 176, pp. 259-269, 2008.

[5] M. Fiacchini, T. Alamo, C. Albea, and F. Camacho, “Adaptive model predictive control of the hybrid dynamics of a fuel cell system, control application,” IEEE International Conference on, Singapore. DOI: 10.1109/CCA., 2007. 4389435.

[6] J. T. Pukrushpan, A. G. Stefanopoulou, and H. Peng, “Control of fuel cell power systems, springer,” vol. 13, No. 1, pp. 3-14, 2004.

[7] Y. J. Zhai, and D. L. Yu, “Radial-basis-function-based feedforward–feedback control for air–fuel ratio of spark ignition engines,” School of Engineering, LJMU, Liverpool, UK, vol. 222 Part D: J. Automobile Engineering, pp. 416-428, 2007.

[8] Y.J. Zhai, D.L.Yu, Reza T., Y. Al-Hamidi, “Fast predictive control for air-fuel ratio of SI engines using a nonlinear internal model,” international Journal of Engineering, Science and Technology, vol. 3, No. 6, 2011, pp. 1-17, 2011.

[9] Kristine Z. T., L. Bell, and L. Harry, T. Van, “A recursive least squares implementation for lcmp beamforming under quadratic constraint,” IEEE Transactions on Signal Processing, vol. 49. No. 6, 2001.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# A Relaxation-based Approach for the Orthogonal Procrustes Problem with Data Uncertainties

Shakil Ahmed
Department of Electrical
and Electronic Engineering,
Imperial College London, UK
shakil.ahmed08@imperial.ac.uk

Imad M. Jaimoukha
Department of Electrical
and Electronic Engineering,
Imperial College London, UK
i.jaimouka@imperial.ac.uk

*Abstract*—**The orthogonal Procrustes problem (OPP) deals with matrix approximations. The solution of this problem gives an orthogonal matrix to best transform one data matrix to another, in a Frobenius norm sense. In this work, we use semidefinite relaxation (SDR) to find the solutions of different OPP formulations. For the standard problem formulation, this approach yields an exact solution, i.e. no relaxation gap. We also address uncertainties in the data matrices and formulate a min-max robust problem. The robust problem, being non-convex, turns out to be a difficult optimization problem; however, it is relatively straight forward to approximate it into a convex optimization problem using SDR. Our preliminary results on robust problem show that the solution of the relaxed uncertain problem does not guarantee zero relaxation gap, and as a result, we cannot always find a solution, which satisfies the orthogonality constraint. In such cases we use orthogonalization, which gives the nearest orthogonal matrix from the SDR based solution. All these relaxed formulations, can be easily converted into a semidefinite program (SDP), for which polynomial time efficient algorithms exists. For the nominal problems, the presented approach may not be computationally efficient than other existing methods. In this work, our main contribution is to demonstrate that the SDR approach provides a unified framework to solve not only the standard OPP but can also solve the problems with uncertainties in the data matrices, which other existing approaches cannot handle.**

## I. INTRODUCTION

Orthogonal Procrustes problem (OPP) is a well known mathematical problem [1], [2]. It deals with finding a geometrical transformation that involves rotations or reflections with orthogonality constraint. In simple words, given two arbitrary real matrices $A$ and $B$ of the same dimension, OPP finds an orthogonal matrix $X$, which can best transform one matrix to the other, such that the Frobenius norm of the error $AX-B$ is minimized. There are many formulations of this problem, which can address rotations, reflections and translations having applications in various areas, such as image processing, computer vision, statistics, satellites and aerospace. In this work, we are mainly interested in the formulations, which can address rotations and reflections.

In image processing and machine learning, OPP is used in pose estimation, which involves estimation of a camera or some object position and orientation, either relative to a model reference frame, or at a previous time using a camera or a range sensor [3]. This application involves both rotation and translation. In satellites and aerospace applications, OPP is used for rigid body attitude determination and involve only rotations. In these applications, information of some vector quantities, such as the earth magnetic field, sun and star direction, object position, is obtained from both a sensor and a mathematical model to estimate the rigid body attitude. In statistics, OPP is used for principal component analysis, which is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. For other variants of the Procrustes problems and their applications, please see [2].

Many solutions of the OPP exist in literature, suiting different applications. One mostly used solution of the orthogonal Procrustes problem is based on the singular values decomposition (SVD) [1], [4]. Different variants of the SVD based solutions are used in many applications. Few customized algorithms were also developed to meet some application specific requirements. For example, different solutions proposed for the Wahba problem [5], which is a subclass of OPP for rotations, and used in satellite attitude determination. Although, some proposed algorithms are based on the SVD approach [6], most of the algorithms used in practical applications are based on customized solutions for high computational speed, such as QUEST [7], ESOQ, ESOQ2 [8], [9].

Semidefinite relaxation (SDR) is considered a powerful and computationally efficient approximation technique for difficult optimization problems [10], such as non-convex and robust problems with quadratic cost and constraints. In this work, we demonstrate that this relaxation approach provides a unified framework to solve different formulations of the Procrustes problem and some of its more sophisticated extensions. We present SDR based convex formulations for the standard OPP and its variant for rotations, which needs to deal with the additional nonlinear constraint of $det(X)=1$. The relaxed formulations, however, result in no gap, giving an exact solution for the nominal problems. We also show that the SDR framework is much more general and can handle many extensions of this problem, which the other existing approaches cannot deal with. In this regard, our main contribution is to use the relaxation approach for the OPP with uncertain data matrices, which is a comparatively less addressed topic because of its

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

mathematical complexity [11], [12]. We formulate and solve a relaxed version of the robust min-max problem. For the relaxed robust problem, the gap is not always zero. For such cases, we present a method to extract the nearest possible solution from the optimal, satisfying the constraints of the original problem.

The paper is organized as follows. In Section II, an orthogonal Procrustes problem is presented along with an existing solution based on the singular value decomposition. The application of the SDR is discussed in Section III, both for the standard problem and its extension for rotations. Section IV discusses the robust formulation and its solution. Section V presents numerical simulations to support the presented results and to show the effectiveness of the robust formulation.

*Notation:* For a matrix, $A \succ 0$ ($A \succeq 0$) means that $A$ is positive definite (semidefinite). The Frobenius norm of a matrix A is $\|A\|_F = \sqrt{\mathrm{tr}(A^TA)}$. $I_n$ denotes the identity matrix of size $n$, while $0_{n\times m}$ represents a matrix of $n$ rows and $m$ columns with all zero entries. The set of singular value (eigenvalue) of a $A$ is represented by $\sigma(A)$ ($\lambda(A)$). Operator $\mathrm{diag}(\lambda_1,\lambda_2,\dots,\lambda_n)$ represents a matrix of size $n\times n$, having only diagonal elements $\lambda_1,\lambda_2,\dots,\lambda_n$. Operator $\mathrm{trace}(A)$ is sum of all the diagonal entries of a matrix $A$.

## II. THE ORTHOGONAL PROCRUSTES PROBLEM

The orthogonal Procrustes problem (OPP) is mathematically defined as

$$\begin{aligned} \min_{X} \quad & \|AX-B\|_F^2 \\ \text{subject to} \quad & XX^T = I, \end{aligned} \tag{1}$$

where $A,B \in \mathbb{R}^{m\times n}$ are given data matrices, $m \geq n$ and $X \in \mathbb{R}^{n\times n}$ is the unknown orthogonal matrix, which belongs to an orthogonal group of order $n$, i.e. $\mathbb{X} =: \{X : XX^T = I, \det(X) = \pm 1\}$.

One important subclass of the orthogonal Procrustes problem includes an additional nonlinear constraint $\det(X) = +1$ (see, for example, Wahba problem [5]). This problem deals specifically with rotations and has a wide range of applications. In these applications, we are only interested in $X \in \mathbb{R}^{3\times 3}$ i.e. the solution now belongs to $SO(3)$, a special orthogonal group of order 3, defined as $\mathbb{X}_+ =: \{X : X \in \mathbb{R}^{3\times 3}, XX^T = I, \det(X) = +1\}$.

One can find many solutions of this problem in the literature. Most solutions are generally application specific, satisfying some special requirements, such as computational efficiency, numerical stability. We will present here one existing approach, which is considered numerically the most robust, and is based on the singular value decomposition (SVD) [1], [6].

### A. An SVD based solution

To find an SVD based solution, first we write the objective function of (1) as

$$\begin{aligned} \|AX-B\|_F^2 &= \mathrm{tr}\left((AX-B)(AX-B)^T\right) \\ &= \mathrm{tr}\left(AXX^TA^T + BB^T - AXB^T - BX^TA^T\right). \end{aligned} \tag{2}$$

To simplify the expression, we use the constraint $XX^T = I$ in (2) and write $AXX^TA^T = AA^T$. Further, neglecting the constant terms, i.e. $AA^T + BB^T$, we can write following maximization problem, which is equivalent to (1) in its argument.

$$\begin{aligned} \max_{X} \quad & \mathrm{tr}(BX^TA^T) \\ \text{subject to} \quad & XX^T = I. \end{aligned} \tag{3}$$

To solve this problem, we use the permutation property of the trace operator [16] and write the cost function as

$$\begin{aligned} \mathrm{tr}(BX^TA^T) &= \mathrm{tr}(X^TA^TB), \\ &= \mathrm{tr}(X^TU\Sigma V^T) \\ &= \mathrm{tr}(V^TX^TU\Sigma) \\ &\leq \mathrm{tr}(\Sigma) = \sum_i \sigma_i. \end{aligned} \tag{4}$$

Here $U\Sigma V^T$ is the singular value decomposition of the term $A^TB$, where $U,V$ are unitary matrices. The inequality in (4) becomes an equality when $V^TX^TU = I$, i.e. $X = UV^T$, which is the required solution.

For problems with additional constraint $\det(X) = +1$ (rotations), to maximize (4), we use few properties of the determinant operator [16] and write as

$$\begin{aligned} \det(V^TX^TU) &= \det(U^TVX) \\ &= \det(U^TV)\det(X) \\ &= \det(U^TV) = \pm 1. \end{aligned} \tag{5}$$

If $\det(U^TV) = 1$, the maximum is attained for $V^TX^TU = I$, while if $\det(U^TV) = -1$, the maximum is attained for $V^TX^TU = \mathrm{diag}(1,1,\dots,1,-1)$.

A unified solution for both problems can be given as

$$X = U\,\mathrm{diag}(1,1,\dots,\det(U^TV))V^T. \tag{6}$$

## III. A RELAXATION APPROACH FOR OPP

We addressed the orthogonal Procrustes problem (1) using semi-definite relaxation approach [12], [13]. We will present a relaxed formulation of the standard OPP (1) as well as the OPP for rotations.

### A. Relaxation of the standard OPP

To derive a semi-definite relaxation of the standard OPP, we use (2) and simplify the expression using the constraint $XX^T = I$. By introducing a linear objective, we can write the following optimization problem, which is equivalent to (1),

$$\begin{aligned} \min_{X,M} \quad & \mathrm{tr}(M) \\ \text{subject to} \quad & M - AA^T - BB^T + AXB^T + BX^TA^T \succeq 0, \\ & XX^T = I. \end{aligned} \tag{7}$$

As the orthogonality constraint $XX^T = I$ is not convex, we relax it to a convex quadratic inequality $XX^T \preceq I$. Using this relaxation, we write an approximate problem, which has a linear cost and linear matrix inequality constraints, as

$$\begin{aligned} \min_{X,M} \quad & \mathrm{tr}(M) \\ \text{subject to} \quad & \mathscr{M} \succeq 0, \\ & \mathscr{X} \succeq 0, \end{aligned} \tag{8}$$

where $\mathscr{M} = \begin{bmatrix} M + AXB^T + BX^TA^T & A & B \\ A^T & I & 0 \\ B^T & 0 & I \end{bmatrix}$ and $\mathscr{X} = \begin{bmatrix} I & X \\ X^T & I \end{bmatrix}$ are Schur complement [16] for the first and the relaxed second constraint in (7).

We have the following result regarding the gap between the original problem and its relaxation.

*Theorem 1:* There is no gap between problem (1) and its relaxation (8) and the SDR solution of (8) is the optimal solution of (1).

*Proof:* The proof is evident from (4). To elaborate this point, assume that the gap between (1) and (8) is zero, i.e. $\text{tr}(M) = AA^T + BB^T - 2BX^TA^T$. Now the minimum value of $\text{tr}(M)$ will be obtained when $BX^TA^T$ is maximum. From (4), we know that the maximum value of $BX^TA^T$ is achieved when $X = UV^T$, where $UV^T$ is the singular value decomposition of $A^TB$ and the obtained $X$ satisfies the orthogonality constraint. Hence, the solution of the SDR is obtained by pushing $XX^T$ towards $I$, resulting in no gap between (8) and (1). ■

### B. Relaxation of the OPP for rotations

The OPP formulation for rotations needs to handle the additional nonlinear constraint $\det(X) = 1$, which cannot be directly handled in the SDR framework. The OPP for rotations is of much interest from practical point of view for many applications. In such applications, we are only interested in the data matrices, where $n = 3$, e.g. applications such as finding camera orientation or rigid body attitude. In such cases, the rows of $A$ and $B$ represent information of same quantities obtained from different sources.

One approach to handle rotation problem in the SDP framework is proposed by [15]. We refer to [17, Proposition 4.1] for more details. We present the main point in a simplified form.

Consider a symmetric matrix $Z \in \mathbb{R}^{4\times 4}$, then an exact SDP representation of the convex hull of $SO(3)$ is given by the following expression, where the orthogonal matrix $X \in SO(3)$ is represented in terms of the elements of the matrix $Z$ as

$$X(Z) = \begin{bmatrix} z_{11}+z_{22}-z_{33}-z_{44} & 2z_{23}-2z_{14} & 2z_{24}+2z_{13} \\ 2z_{23}+2z_{14} & z_{11}-z_{22}+z_{33}-z_{44} & 2z_{34}-2z_{12} \\ 2z_{24}-2z_{13} & 2z_{34}+2z_{12} & z_{11}-z_{22}-z_{33}+z_{44} \end{bmatrix}, \tag{9}$$

if and only if the matrix $Z$ satisfies the following constraints

$$\begin{aligned} Z &\succeq 0, \\ \text{trace}(Z) &= 1. \end{aligned} \tag{10}$$

In the above expression $z_{ij} = Z(i,j)$. To understand this fact, consider that positive semidefinite $4\times 4$ matrices with both trace and rank 1 are given as

$$Z = \frac{1}{a^2+b^2+c^2+d^2}\begin{bmatrix} a^2 & ab & ac & ad \\ ab & b^2 & bc & bd \\ ac & bc & c^2 & cd \\ ad & bd & cd & d^2 \end{bmatrix}. \tag{11}$$

The image of this rank 1 matrix under the linear map (9) is precisely the group $SO(3)$ [17]. This parameterization is known as the *Cayley transform.*

Now embedding the trace constraint, i.e. $z_{11} + z_{22} + z_{33} + z_{44} = 1$ within the definition of $Z$, we can write an exact SDP representation for the rotation problem as

$$\begin{aligned} \min_{Z,M} \quad & \text{tr}(M) \\ \text{subject to} \quad & \mathscr{M} \succeq 0, \\ & Z \succeq 0, \end{aligned} \tag{12}$$

where

$$\mathscr{M} = \begin{bmatrix} M + AX(Z)B^T + BX(Z)^TA^T & A & B \\ A^T & I & 0 \\ B^T & 0 & I \end{bmatrix}. \tag{13}$$

The new $X(Z)$ with embedded trace constraint is given as

$$X(Z) = \begin{bmatrix} 2z_{11}+2z_{22}-1 & 2z_{23}-2z_{14} & 2z_{24}+2z_{13} \\ 2z_{23}+2z_{14} & 2z_{11}+2z_{33}-1 & 2z_{34}-2z_{12} \\ 2z_{24}-2z_{13} & 2z_{34}+2z_{12} & -2z_{22}-2z_{33}+1 \end{bmatrix}. \tag{14}$$

The solution of this problem gives $Z$ satisfying the constraints given in (10). This $Z$ is then used to calculate $X$, which is the optimal solution of the rotation problem. Later on, we will make use of this transformation to solve the rotation problem with data uncertainties.

## IV. THE OPP WITH DATA UNCERTAINTIES

Depending upon application, the data matrices $A$ and $B$ are generally obtained from different sources, e.g. some camera, sensors, mathematical models. This input information has always some sort of uncertainty, such as noise, sensor and modelling errors. The level of uncertainty, however depends upon the quality of measurement or mathematical modelling. In some applications the level of uncertainty can be high. It is a well known fact that these uncertainties can severely affect the accuracy of the obtained solution. The error in the solution will be large in the worst case uncertainties. To overcome the issue of sensitivity of the solution to data uncertainties, in this section, we will formulate and solve a robust optimization problem. In the formulated robust min-max problem, we obtain a solution, which could minimize the chosen cost function under the worst case uncertainties. The solution of the robust OPP gives better estimate of the unknown matrix $X$, with large data uncertainties.

### A. Uncertainty Representation in the Data Matrices

We consider the following data uncertainty structure:

$$\begin{bmatrix} \tilde{A} & \tilde{B} \end{bmatrix} = \begin{bmatrix} A & B \end{bmatrix} + E\Delta\begin{bmatrix} F_1 & F_2 \end{bmatrix}, \tag{15}$$

where $\tilde{A}$ and $\tilde{B}$ are uncertain data matrices, $A$ and $B$ represent the nominal data, $E, F_1$ and $F_2$ are known matrices and $\Delta$ is the uncertainty matrix such that $\Delta\Delta^T \preceq I$.

Perturbation model of this form is common in robust estimation, filtering and control [18], [19]. By a suitable selection of $E, F_1, F_2$ and $\Delta$, this model can represent both structured and unstructured uncertainty. For example, a norm-bounded full $\Delta$ will represent unstructured uncertainty, while a norm-bounded diagonal $\Delta$ with a suitable choice of other matrices will represent structured uncertainty. The suitable choice of $E, F_1$ and $F_2$ specify both the components of $A$ and $B$ affected by the uncertainty $\Delta$ and also the amount of uncertainty, e.g. as a percentage of input data.

In this work, we use this general uncertainty structure to formulate the robust problem. The choice of the constant matrices will define it to be structured or unstructured. Further, we will consider a ball uncertainty, i.e. $\Delta\Delta^T \preceq I$.

### *B. The Robust Problem*

For the robust problem formulation, we will follow the min-max approach, i.e. to minimize the objective function under the worst case uncertainties. Using the uncertainty model (15), we define the following robust problem:

$$\begin{aligned} \min_X \max_\Delta \quad & \|\tilde{A}X - \tilde{B}\|_F^2 \\ \text{subject to} \quad & XX^T = I, \\ & \Delta\Delta^T \preceq I, \end{aligned} \tag{16}$$

where $\tilde{A} = A + E\Delta F_1$ and $\tilde{B} = B + E\Delta F_2$.

To solve the min-max problem, we use SDR approach, following the same steps as used while solving the nominal problem (8). The main challenge here is to handle uncertainty in the max problem. By expanding the cost of (16), we get

$$\begin{aligned} J(X,\Delta) =& \|\tilde{A}X - \tilde{B}\|_F^2 \\ =& \operatorname{tr}((A+E\Delta F_1)XX^T(A+E\Delta F_1)^T - \\ & (A+E\Delta F_1)X(B+E\Delta F_2)^T - \\ & (B+E\Delta F_2)X^T(A+E\Delta F_1)^T + \\ & (B+E\Delta F_2)(B+E\Delta F_2)^T). \end{aligned} \tag{17}$$

The constraint $XX^T = I$ is used to further simplify the cost. To transform this problem into a tractable LMI formulation, we first replace the cost $J(X,\Delta)$ in (17) with a linear objective function trace$(M)$, and write the following equivalent problem

$$\begin{aligned} \min_{X,M,\Delta} \quad & \operatorname{tr}(M) \\ \text{subject to} \quad & M - J(X,\Delta) \succeq 0 \;\forall\; \Delta : \Delta\Delta^T \preceq I, \\ & XX^T = I. \end{aligned} \tag{18}$$

We can further simplify the optimization problem by relaxing the first constraint and making it independent of $\Delta$. For this we use following identity, where the left hand side is equal to the right hand side:

$$\begin{aligned} & M - J(X,\Delta) = \lambda E(I - \Delta\Delta^T)E^T + \\ & \begin{bmatrix} I & E\Delta \end{bmatrix} \begin{bmatrix} M - J_n(X) - \lambda EE^T & T_2(X) \\ (T_2(X))^T & \lambda I - J_\Delta(X) \end{bmatrix} \begin{bmatrix} I \\ \Delta^T E^T \end{bmatrix} \end{aligned} \tag{19}$$

where $J_n(X), T_2(X)$ and $J_\Delta(X)$ are defined as

$$\begin{aligned} J_n(X) &= AA^T + BB^T - AXB^T - BX^TA^T, \\ T_2(X) &= BX^TF_1^T + AXF_2^T - AF_1^T - BF_2^T, \\ J_\Delta(X) &= F_1F_1^T + F_2F_2^T - F_1XF_2^T - F_2X^TF_1^T. \end{aligned}$$

Further we define the matrix in the second term on right hand side of (19) as

$$\mathscr{T}(M,X,\lambda) = \begin{bmatrix} M - J_n(X) - \lambda EE^T & T_2(X) \\ (T_2(X))^T & \lambda I - J_\Delta(X) \end{bmatrix}. \tag{20}$$

The right hand side of (19) is either zero or positive because $(I - \Delta\Delta^T) \succeq 0$ and we impose $\lambda \geq 0$ and $\mathscr{T} \succeq 0$, ensuring that $M$ is an upper bound on $J(X,\Delta)$. Finally we write a relaxation of (16) as

$$\begin{aligned} \min_{X,M,\lambda} \quad & \operatorname{tr}(M) \\ \text{subject to} \quad & \mathscr{T}(M,X,\lambda) \succeq 0, \\ & \mathscr{X} \succeq 0, \\ & \lambda \geq 0, \end{aligned} \tag{21}$$

where $\mathscr{X}$ is the same as defined in (8).

*Remark 1:* Unlike Theorem 1, the gap between the robust problem (16) and its semidefinite relaxation (21) is not necessarily zero.

One possible reason of the non-zero gap may be that the relaxed maximization for all $\Delta : \Delta\Delta^T \preceq I$, is achieved at an $X$, which lies inside $XX^T = I$, i.e. the obtained $X$ does not satisfy the orthogonality constraint of the original problem.

### *C. Orthogonalization of X*

When the gap between the robust problem and its semidefinite relaxation is not zero, the $X$ does not satisfy the orthogonality constraint. For such cases, we find the nearest orthogonal $X$ in the Frobeneous norm sense. Such an $X$ can be obtained by solving following optimization problem:

$$\begin{aligned} \min_{X_o} \quad & \|X_o - X\|_F^2 \\ \text{subject to} \quad & X_oX_o^T = I, \end{aligned} \tag{22}$$

where $X_o$ is the required orthogonal matrix. This problem is same as the nominal Procrustes problem (1) with $A = I$ and $B = X$. The required matrix $X_o = U_xV_x^T$, where $U_x\Sigma_xV_x^T = X$ is the singular value decomposition of $X$.

### *D. Effect on the robust performance*

It can be argued that $X_o$ is not the optimal solution of (21), however it may be considered a suitable solution of (16), because this solution not only satisfies the orthogonality constraint and also results in a minimum cost variation. To evaluate this, let $e_1 = \tilde{A}X_o - \tilde{B}$ and $e_2 = \tilde{A}X - \tilde{B}$, then $\|e_1 - e_2\| = \|\tilde{A}(X_o - X)\| \leq \|\tilde{A}\|\|X_o - X\|$. A minimum value of $\|X_o - X\|$ will ensure that orthogonal $X_o$ will result in a minimum cost variation from the SDR solution. However, this new solution may not have the same properties from the robustness perspective. Some analysis of the performance is presented in the simulation section, however, the robustness properties of $X_o$ needs further analysis.

### *E. Robust problem for rotations*

Using the same transformation as discussed in Section III-B, we can write the robust problem (21) for rotations as

$$\begin{aligned} \min_{Z,M,\lambda} \quad & \mathrm{tr}(M) \\ \text{subject to} \quad & \mathscr{T}(M,Z,\lambda) \succeq 0, \\ & Z \succeq 0, \\ & \lambda \geq 0, \end{aligned} \tag{23}$$

where $\mathscr{T}(M,Z,\lambda)$ is the same as (20) with $X$ replaced by $X(Z)$.

However, it is observed that the proposed transformation also does not always work for problems with uncertainties in the data matrices. So for the cases where $X$ is not orthogonal, we still need to use orthogonalization as discussed above, to obtain a solution of the rotation problem.

Finally, based on the numerical simulations performed for the robust problem, we present following remark.

*Remark 2:* It is observed in the numerical simulations that limiting the size of the maximum uncertainty in the robust problem (21) can significantly reduce the number of cases, where the solution $X$ is not orthogonal. This observation also applies for rotation OPP (23), where the number of cases when $\mathrm{rank}(Z) \neq 1$, reduce significantly, reducing the likelihood of occurrence of the non-orthogonal $X$.

## V. Simulation Results

This section presents numerical simulations to evaluate the performance of the presented relaxation approaches and to support different discussions and results.

### *A. Analysis of the SDR for the standard OPP*

Firstly, we compare the solution of (8) with the SVD solution. For this comparison, we generate random $A$ and $B$ matrices of the size $10 \times 10$ using MATLAB command `randn`. Results of the simulation is given in Figure 1. The plot compares and cost $\|AX - B\|_F^2$ using both the SVD and the SDR based solutions. Gap between the relaxed and the original problem is also shown, which is zero, supporting Theorem 1. The last subplot shows $\det(X)$, which is $\pm 1$ indicating that the random problems are either rotations or reflections.

### *B. Analysis of the SDR for rotations OPP*

The performance of the OPP for rotations is shown in Figure 2. First subplot presents the relaxation gap, which is zero, validating the exactness of (12) for rotations. Other subplots show some other parameters of the $Z$ and $X$ matrices. It can be observed that the parameters, such as trace and rank of $Z$, and determinant of $X$ are as desired for all random cases.

### *C. Robust performance evaluation*

In this section, we evaluate the performance of the solutions obtained by solving the approximate formulations of the standard and the robust problems for a set of bounded uncertainties in the input matrices. For this analysis, we considered a structured uncertainty description with a suitable choice of $E, F_1$, and $F_2$ matrices. To obtained data for this test, first we generated a random $A$ matrix using MATLAB's `randn` command, and an orthogonal matrix $X$ using `orth` command, both in $\mathbb{R}^{3\times 3}$. Using this orthogonal matrix, we calculate the matrix $B$. This $A,B$ pair represents an exact data set, i.e. matrix $A$ can be exactly transformed to $B$ using $X$. In this pair, we added uniformly distributed random error within a range of 30% of the size of the elements of the true matrices, to obtain a data set with errors. We then solved both the nominal problem (4) and the robust problem (21). A large number of tests were performed by adding unifromly distrubuted random error in the nominal data within the set uncertainty bounds. The cost value was evaluated for both the nominal and the robust solution. The histogram of the test results is given in Figure 3, where $x-$axis represents the cost value, while $y-$axis represents the frequency of occurrance of the tests. It can be observed that the dispersion of the cost value using the nominal solution is much larger than the robust solution. This benefit is more obvious for the worst case scenarios. However, for nominal cases the robust solution suffers from an offset as compared to the nominal solution.

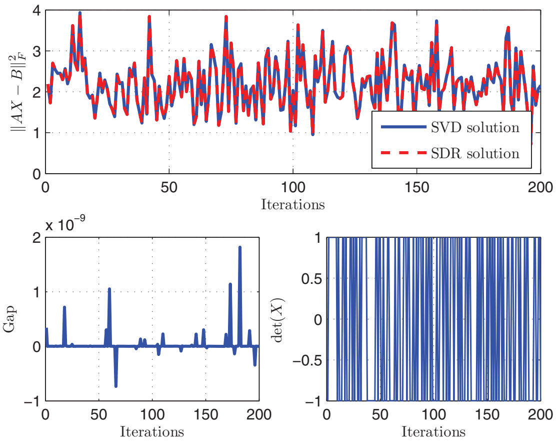


Fig. 1. Analysis of the semidefinite relaxation of the standard OPP (8).

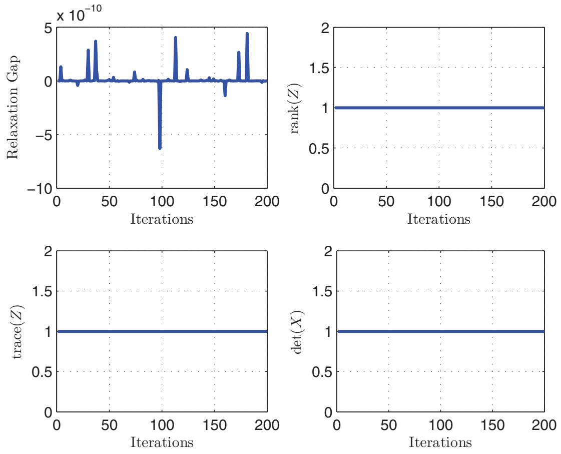


Fig. 2. Analysis of the semidefinite relaxation of OPP for rotations (12).

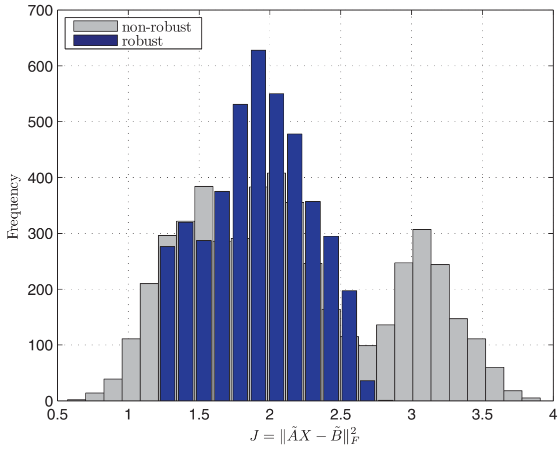


Fig. 3. Evalutation of the robust performance.

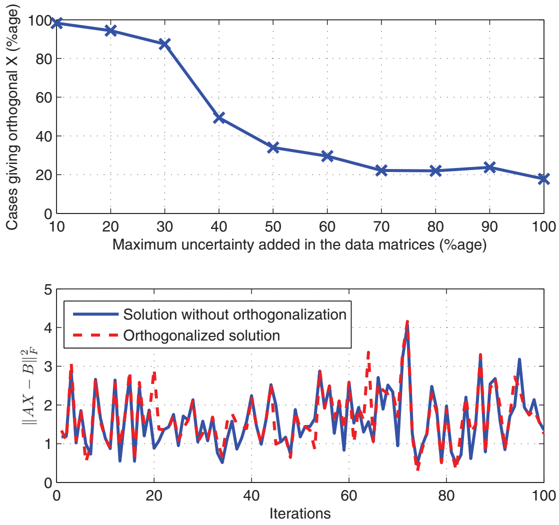


Fig. 4. First subplot shows number of orthogonal solutions (%age) verses uncertainty bounds, and second subplot compares cost using the SDR solution before and after orthogonalization.

### D. Orthogonalization step

Lastly, we analyze the orthogonalization step and its effects on the cost value. For the first point, we performed simulations by varying maximum uncertainty level in the the nominal data and plotted the number of cases (in percentage), which give an orthogonal $X$. Results are shown in the first subplot of Figure 4. The analysis shows that the number of tests having orthogonal $X$ increase significantly, when the size of the uncertainty is small, supporting Remark 2. Further, we compare the cost variation using the optimal solution of the approximate robust problem (21) and its orthogonalized solution for 200 runs of random data with within bounded uncertainty, showing that the cost variation due to the orthogonalization is not much.

## VI. Conclusions and Future Work

We presented an approach based on semidefinite relaxations to solve different formulations of the orthogonal Procrustes problem. It was demonstrated that the relaxation of the nominal problem results in no gap between the actual and the relaxed problems. The SDR framework also allows to handle uncertainties in the data matrices. It was further demonstrated that while considering uncertainties, the gap is not necessarily zero, which results in a solution not satisfying the orthogonality constraint of the original problem. In such cases, orthogonalization was proposed for the obtained solution. However, there are some issues, which need further analysis. In this regard, our future work may consider further study of the relaxation gap with uncertainties. Moreover, some work may be done on casting these problems in a standard SDP framework and analyzing its computational complexity using some efficient SDP algorithm.

## References

[1] Nicholas J. Higham, The Symmetric Procrustes Problem, *BIT*, 28 (1988), pp. 133143).

[2] John C Gower and Garmt B Dijksterh, *Procrustes Problems*, OUP Oxford, 2004, UK.

[3] Robert M. Haralick, Hyonam Joo, Chung-Nan Lee, Xinhua Zhuang, Vinay G. Vaidya and Man Bae Kim, Pose Estimation from Corresponding Point Data, *IEEE Transactions on Systems, Man. and Cybernetics*, Vol. 19, NO. 6, November/December 1989.

[4] Schonemann, P.H., A generalized solution of the orthogonal Procrustes problem, *Psychometrika*, volume 31 (1964, page 110.

[5] Wahba, G., A least square estimate of spacecraft attitude, *SIAM Review*, 7(3), 409, 1965.

[6] Markley, F.L., Attitude determination using vector observations and the singular value decomposition. *Journal of Astronautical Sciences*, 36(3), 245258, 1988.

[7] Shuster, M. D. and Oh, S., Three-Axis Attitude Determination from Vector Observation, *Journal of guidance and Control*, vol. 4, issue 1, pp. 70 77, 1981.

[8] Mortari, D., ESOQ: A closed form solution to the Wahba problem. *Journal of the Astronautical Sciences*, 45(2), 195204, 1997.

[9] Mortari, D., Second estimator of the optimal quaternion, *Journal of Guidance, Control and Dynamics*, 23(5), pp. 885 888, 2000.

[10] Luo Z. Q., Ma W. K., So A. M, Ye Y. and Zhang S., Semidefinite Relaxation of Quadratic Optimization Problems, *IEEE Signal Processing Magazine, Special Issue on Convex Optimization for Signal Processing*, vol. 27, issue 3, pp. 20 - 3, May 2010.

[11] Ahmed, S. and Kerrigan, E. C., Robust Static Attitude Determination using Robust Optimization, *Proc.* 18$^{th}$ *IFAC World Congress*, Milano, Italy, August 2011.

[12] Ahmed, S., Kerrigan, E. C. and Imad M. Jaimoukha, Semi-definite relaxation for an attitude determination problem, *Proc.* 50$^{th}$ *CDC*, Orlando, Florada, USA, December 2011.

[13] Malik, U., Jaimoukha, I. M., Halikias, G. D., and Gungah, S. K., On the gap between the quadratic integer programming problem and its semidefinite relaxation, *Mathematical Programming*, vol. 107, pp. 505 515, 2006.

[14] Halikias, G. D., Jaimoukha, I. M., Malik, U., and Gungah, S. K., New bounds on the unconstrained quadratic integer programming problem, *Mathematical Programming*, vol. 39, pp. 543 554, 2007.

[15] Pablo A. Parrilo, Semidefinite programming and convex algebraic geometry, *Foundations of Computational Mathematics (FoCM 2008)*, plenary talk, June 2008.

[16] Golub, G. H. and Loan, C. F. V., *Matrix Computations*, The Johns Hopkins University Press, USA, 3rd ed., 1996.

[17] Raman Sanyal, Frank Sottile and Bernd Sturmfels, ORBITOPES. *Mathematika*, 57 , pp 275-314 doi:10.1112/S002557931100132X.

[18] Laurent El Ghaoui and Herve Lebret, Robust Solutions to Least Squares Problems with Uncertain Data. *SIAM J. Matrix Anal. Appl.*, Vol. 18, No. 4, pp. 1035-1064, October 1997.

[19] A. H. Sayed, V. H. Nascimento, and F. A. M. Cipparrone, A Regularized Robust Design Criterion for Uncertain Data. *SIAM J. Matrix Anal. Appl.*, Vol. 23, No. 4, pp. 1120-1142, October 2002.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Fuzzy logic control for solar powered hydrogen production, storage and utilisation system

Fan Zhang, Kary Thanapalan, Andrew Procter, Jon Maddy, Alan Guwy
University of Glamorgan
Hydrogen Centre
Baglan Energy Park
Port Talbot, SA12 7AX
United Kingdom
Email: fanvac@gmail.com

*Abstract*—**Climate change concerns, increasing global energy demand, coupled with limited supply of fossil fuels, calls for development of new power source. Solar energy is a very promising renewable energy source to moderate the growth of energy demand. The combination of electrolyser and fuel cell which use hydrogen as an energy carrier extends the utilisation of the solar energy. For this integrated solar powered hydrogen production, storage and utilisation system, one of the problems is to develop an efficient control system to improve the performance of the overall system. This paper presents a power management strategy based on fuzzy logic technology to address such problem. The target of this power management strategy is to meet the power demand, to maximise the hydrogen production and to minimise the usage of battery. Therefore, the overall system's efficiency will be increased and lifetime of the battery pack will be extended. The numerical results based on real solar data for a one year period shown that the proposed fuzzy logic controller behaved as expected, it was able to meet the power demand and to store the hydrogen when possible while maintain the battery's state of charge at desired level.**

***Keywords - Hydrogen production and storage, Renewable energy, Power management, Fuzzy logic control***

## I. Introduction

The global energy demand is increasing rapidly. The concerns of limited supply of conventional fossil fuels and the emission of green house gases call for new solutions to this energy problem. Renewable energy can be part of the solution to provide clean and sustainable energy supply.

However, the inherent intermittency in most of the renewable sources causes problems for power-on-demand requirements. Hydrogen is considered to be a promising candidate as an energy carrier to compensate this problem.

Renewable powered hydrogen production and utilisation systems can be either standalone or grid connected. Over the past few years, the development and application of such systems has increased significantly. A large body of research work addressed various topics for such systems, including modelling, simulation, control and performance evaluation [1]–[7].

A typical system normally consists of one or several renewable power sources, such as photovoltaic (PV) array, wind turbines, micro-hydro, geothermal, etc. An electrolyser is used to convert excess energy to produce hydrogen as an energy carrier. The fuel cell will be used as power source to convert the chemical energy from hydrogen into electrical energy when the power demand is higher than the supply of renewable sources. Batteries or a battery pack are usually used to maintain a constant DC bus voltage and to store or supply short-term energy. The emphasis of such systems is not only to improve the performance of existing hydrogen production, storage and utilization technologies, but also to integrate various units effectively with renewable energy sources through overall power management strategies [8].

Several researchers addressed power management strategy (PMS) design and applications for such system. However, from above examples [1]–[6], [9], [10], it can be seen that for most of the reported PMS, the state of charge (SOC) level of the battery is the main parameter that governs the operation sequence of the electrolyser and the fuel cell. The start-up and shut-down of the electrolyser and fuel cell is relied on the fixed SOC limits. Reference [11] indicated that the two important shortcomings for this type of algorithm are:

1) it did not take into consideration of the system's state except for the batteries' SOC;
2) it did not allow the control of the production or consumption of the hydrogen, which would help manage the energy in the system.

The fuzzy logic (FL) controller developed by [11] determines the appropriate hydrogen production/consumption rate as a function of the system's power inputs and outputs and the batteries' SOC. However, the authors in [11] have not considered the system sizing that may lead to significant drop of hydrogen storage level in 7-day period. Meanwhile, although the battery SOC was maintained within a reasonable region (i.e., between 40% and 60% as the results demonstrated), the oscillation means that considerable amount of energy may be wasted during charging and discharging process of the battery pack.

This study focuses on developing PMS based on FL control methodology for solar powered hydrogen production, storage and utilisation system to improve the overall system's efficiency. The purpose of proposed FL controller is twofold, one is to maximise the hydrogen production which is a function

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

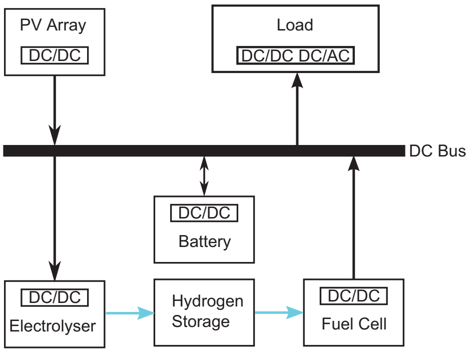


Fig. 1. System configuration of the solar powered hydrogen production, storage and utilisation system

of the battery SOC and the difference of the power flow between available solar power and load demand, the other one is to minimise the usage of batteries. The entire system is modelled with consideration of system sizing and systematic modelling approach. Then the FL controller is designed in order to achieve aforementioned control objectives. Finally, performance of the overall control system is validated over a one-year period using real solar data. The numerical results shown that the designed system has the capability to satisfy the daily load demand as well as produce hydrogen using excessive solar power for future usage, storage, transportation or hydrogen powered car refuelling.

## II. System configuration

The structure of this standalone system is shown in Fig. 1 where the black arrow represents the electricity flow and the blue arrow represents the hydrogen flow.

In this study, a photovoltaic array is considered as the main power source to the system to meet the predefined load demand. Due to the inherent intermittent nature of solar power, a battery pack is included only to smooth the fluctuation of the solar power if necessary. When excess power is available, hydrogen will be produced using an electrolyser. When a power shortage occurs, a fuel cell will be used to convert the chemical energy from hydrogen to electric energy.

The complete model is developed in order to represent the physical system and to design and validate the performance of the proposed controller. The detailed model of DC/DC or DC/AC converter is neglected and they are treated as an ideal device to produce power which is determined by the FL controller.

### A. Load profile

The purpose of this solar hydrogen system is to use solar power to satisfy the load demand of the office block of Hydrogen Centre of University of Glamorgan at Baglan Energy Park, Port Talbot, United Kingdom and utilise excess power to produce hydrogen for later use, storage, transportation or for fuel cell car refuelling, therefore, the load profile analysis is essential for determine the size and the control design of the overall system.

| Appliance | Power (W) | Duty cycle (hours) | Energy consumption (Wh) |
|---|---|---|---|
| Fridge | 23 | 24 | 552 |
| Microwave | 700 | 0.3 | 210 |
| Kettle | 3000 | 0.4 | 600 |
| Lighting | 400 | 4 | 1600 |
| Laptops | 390 | 8 | 3120 |
| Auxiliary devices | 1200 | 24 | 28800 |
| Total | 5713 | – | 34882 |

TABLE I
List of appliances' power, duty cycle and energy consumption

The average energy consumption $E_{load}$ can be estimated by equation

$$E_{load} = \sum_{i=1}^{n} I_{load} V_{load} D_{load} \tag{1}$$

where $I_{load}$, $V_{load}$, $D_{load}$ are the current, voltage and duty cycle of each appliance used in one day, respectively. $n$ is the number of the appliances.

The considered office block is utilised between 09:00 and 17:00 and includes 6 laptops, 1 mini refrigerator, 1 microwave, 1 kettle, lighting and several auxiliary devices such as security system, monitor system, phone system, LAN, boiler etc. The list of appliances together with their power, duty cycle is given in Table I. It is reasonable to assume that this daily load will keep constant throughout the year, for weekdays.

### B. Battery model

The battery stack creates a linkage between all components in the solar powered hydrogen system, it is used as an energy buffer to compensate the power fluctuation.

The battery state of charge $S$ is the only state variable of the battery system model and is given by

$$S = \frac{\left(Q_{max} - \int_0^t I_b dt\right)}{Q_{max}} \tag{2}$$

where $Q_{max}$ is the battery's maximum capacity [12].

The battery current $I_b$ is defined by

$$I_b = -(I_{pv} - I_{load} - I_{el} + I_{fc}) \tag{3}$$

where $I_{pv}$ is the PV array's current, $I_{load}$ is the load current, $I_{el}$ is the electrolyser current and $I_{fc}$ is the fuel cell current.

When charging the battery, the current is negative and positive current indicates discharging mode of the battery. It is found that in order to lengthen the battery lifetime, overcharging and deep discharging should be avoided.

### C. Electrolyser model

The electrolyser operating current is defined by

$$I_{el} = \frac{P_{el}}{V_{el}} \tag{4}$$

where $P_{el}$ is the power of electrolyser.

| | |
|---|---|
| $f_1$ | 269.9112 |
| $f_2$ | 1.1013 |
| $r_1$ | $9.2434 \times 10^{-5}$ |
| $r_2$ | $-2.2120 \times 10^{-7}$ |
| $s$ | 0.1084 |
| $t_1$ | $-1.0636$ |
| $t_2$ | 8.7077 |
| $t_3$ | 267.7084 |
| $A$ | 0.013 |
| $n_c$ | 48 |

TABLE II
PARAMETERS FOR THE ELECTROLYSER

For a given cell temperature, the operating voltage of the electrolyser is expressed as

$$V_{el} = V_{rev} + \frac{r_1 + r_2 T}{A} I_{el} + slog\left(-\frac{t_1 + t_2/T + t_3/T^2}{A} I_{el} + 1\right) \tag{5}$$

where $V_{rev}$ is the reversible cell voltage, $r_1$, $r_2$, $t_1$, $t_2$, $t_3$ are parameters which can be determined experimentally [13].

The hydrogen production rate is proportional to the electrolyser current, hence, the total hydrogen production for an electrolyser which consists of several series connected cells, can be expressed as

$$\dot{n}_{H_2} = \eta_F \frac{n_c I_{el}}{2F} \tag{6}$$

where $\eta_F$ is the Faraday efficiency and can be calculated as $\eta_F = \frac{(I_{el}/A)^2}{f_1 + (I_{el}/A)^2} f_2$, $n_c$ is the number of series connected cells, $F$ is the Faraday constant.

The validated parameters used in the electrolyser model are listed in Table II.

### D. Fuel cell model

The fuel cell operating current is defined as

$$I_{fc} = \frac{P_{fc}}{V_{fc}} \tag{7}$$

where $P_{fc}$ is the fuel cell power.

The fuel cell model is derived and validated in [14]. A brief detail of the fuel cell model is described here, detailed information can be found in [14]. The fuel cell stack voltage can be calculated by multiplying the cell voltage by the number of cells $n_{fc}$ of the stack, hence

$$V_{fc} = n_{fc} \times (E - V_{act} - V_{ohm} - V_{con}) \tag{8}$$

where

$$E = \frac{1}{2F}\left\{\Delta G + \Delta S(T - T_r) + RT\left(ln(P_{H_2}) + \frac{1}{2}ln(P_{O_2})\right)\right\} \tag{9}$$

where $E$ is the open circuit voltage of the fuel cell, $\Delta G$ is the change in the free Gibbs energy, $\Delta S$ is the change of the entropy, $R$ is the universal constant of the gases, $P_{H_2}$ and $P_{O_2}$ are the partial pressures of hydrogen and oxygen respectively. $V_{act}$ represented the activation over potential at the electrodes; $V_{ohm}$ represents the ohmic over potential caused by electrical and ionic conduction loss; $V_{con}$ represented the concentration over potential caused by mass transport limitations of the reactants to the electrodes.

The activation over potential $V_{act}$, including anode and cathode can be calculated as

$$V_{act} = -\left\{\xi_1 + \xi_2 T + \xi_3 Tln\left(\frac{P_{O_2}}{5.1 \times 10^6 e^{\frac{-498}{T}}}\right) + \xi_4 Tln(I_{fc})\right\} \tag{10}$$

$\xi$ represents parametric coefficient for the cell model.

The ohmic voltage drop $V_{ohm}$ is determined by the following expression

$$V_{ohm} = I_{fc}\left(\rho_m t_m A_{fc}^{-1} + c\right) \tag{11}$$

In this model a general expression for resistance is defined to include all the important parameters of the membrane. The resistance to the transfer of protons through the membrane is assumed to be a constant ($c$) and included in the equation as an additional term. $\rho_m$ is the specific resistivity of the membrane for the electron flow. $t_m$ is the thickness of the membrane, $A_fc$ is the cell active area.

The voltage drop due to the mass transport can be determined by

$$V_{con} = -Bln(1 - \theta) \tag{12}$$

and

$$\theta = \left(I_{fc} A_{fc}^{-1}\right)\left(\left(I_{fc} A_{fc}^{-1}\right)_{max}\right)^{-1} \tag{13}$$

where $B$ is a parametric coefficient that depends on the cell and its operation state.

The hydrogen consumption rate of the fuel cell stack is given by

$$\dot{n}_{fc} = \frac{n_{fc} I_{fc}}{2F} \tag{14}$$

## III. FUZZY LOGIC BASED POWER MANAGEMENT STRATEGY

### A. Control objects

The control object is to make full use of the solar power to maximise the hydrogen production and minimise the usage of battery stack, i.e., keep the battery SOC between 50 -60% which will extend its lifespan and increase the overall system's efficiency.

The operating strategies for power management system are listed as follows:

1) The highest priority is to utilise the solar power to satisfy the predefined power demand.
2) If excess solar energy is available, it will be sent to the electrolyser to generate hydrogen for future usage.
3) The electrolyser will keep running as long as the addition power from PV array is available. The electrolyser

maximum power is set to 16kW. The addition power transferred from PV array to the electrolyser will not exceed this limit.

4) If there is still additional power generated by PV array when the load demand is satisfied and the electrolyser is running at its peak limit, extra power will be utilised to charged the battery if the battery SOC is low.
5) Dumping power is only required if the electrolyser is running at its maximum power level and the battery SOC is at desired level.
6) If the power generated by PV array is insufficient to support the electrical demand, the difference is supplied by fuel cell stack which can run up to 4kW.
7) Only when the power demand is higher than the power supply of PV/FC combination, will the battery be used to provide short-term compensation.

### B. Fuzzy logic controller

The required control objects are implemented by a two-input-one-output fuzzy logic controller. Two inputs to the controller are the difference power flow ($dP$) and the battery's SOC. The difference power flow indicates the difference between available solar power and the power demand. The controller will use this information to decide whether excess solar power is available for hydrogen production. The SOC is used to maintain the battery's SOC at desired level in order to prevent overcharge / deep discharge of the battery.

The fuzzy mechanism consists of triangular membership functions for the two inputs and for the output is shown in Fig. 2 and Fig. 3 respectively. The reason for choosing triangular membership function is mainly for the simpler computation of membership value. The dP flow is divided into eight variables. Negative power supply means fuel cell or battery is required to supply the difference and positive power means addition solar power can be used to produce hydrogen or charge the battery. The battery SOC is also described by three variables. It is desired that the SOC will be maintained within the region between 50-60%. Negative power output means fuel cell power is required to compensate the power shortage and positive power output indicates that excess solar power is available for electrolyser to produce hydrogen.

The rules of the FL controller are demonstrated in Table III and the definitions of the linguistic variables are described in Table IV. In Table III, the top row of the table shows the difference of power flow ($dP$) and the left column is the battery's SOC. The cells of the table at the intersection of rows and columns contain the linguistic value for the output corresponding to the value of the first input written at the beginning of the row and to the value of the second input written on the top of the column. The rule output was defuzzified using a centroid computation.

## IV. Simulation and results

The model of the aforementioned solar hydrogen system is created in MATLAB/Simulink and the proposed fuzzy logic controller is implemented using the Fuzzy Logic Toolbox. The

| | | $dP$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | NL | NM | NS | Z | PS | PM | PL | PEL |
| SOC | L | NL | NL | NM | NS | Z | PS | PM | PL |
| | C | NL | NM | NS | Z | PS | PM | PL | PEL |
| | H | NM | NS | Z | PS | PM | PL | PEL | PEL |

TABLE III
FUZZY CONTROL RULE TABLE

| Linguistic | Linguistic meanings |
|---|---|
| PEL | Positive extreme large |
| PL | Positive large |
| PM | Positive medium |
| PS | Positive small |
| Z | Zero |
| NS | Negative small |
| NM | Negative medium |
| NL | Negative large |
| L | Low |
| C | Correct |
| H | High |

TABLE IV
LINGUISTIC VARIABLES IN THE FUZZY INFERENCE SYSTEM

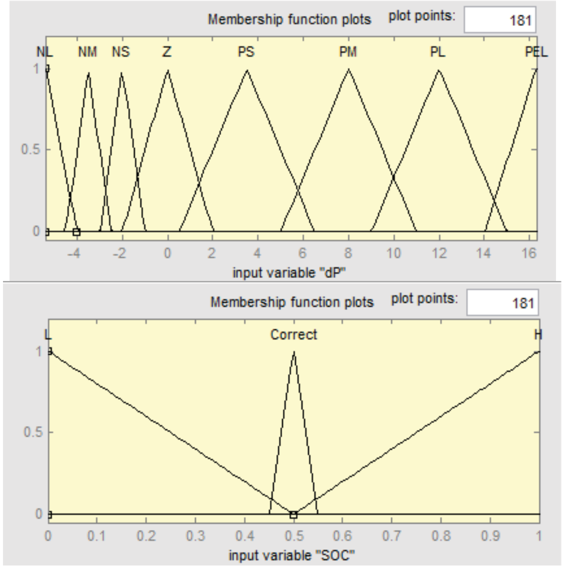


Fig. 2. Membership functions for input variables

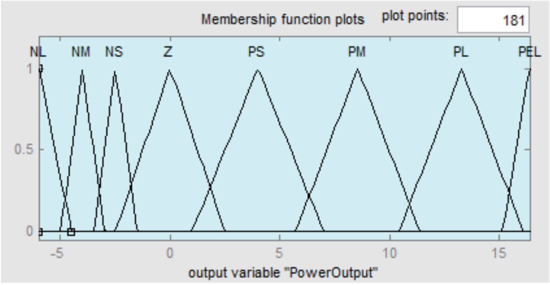


Fig. 3. Membership function for output variable

purpose of the simulation is to observe the performance of proposed system over a period of year including day and night.

The real solar data gathered from the PV panel for entire year of 2010 is utilised in this analysis. The PV array is rated at 18kWp. The annual solar power delivered by the PV array is illustrated in Figure 4. The fluctuation of delivered solar power can be seen from the figure. In summer, the maximum power solar panel can deliver can reach up to 18kW, while during winter and early spring, this value can drop down lower than 0.1kW.

The power generated by the PV array will be used to satisfy the user power demand as stated previously and the difference between available power and the load demand is shown in Figure 5. The negative power indicates that there is insufficient solar power to meet the power demand therefore fuel cell is required to supply the difference. The positive region means that excess solar power is available and can be utilised by electrolyser to produce hydrogen. The figure demonstrated that during winter and spring season, from November to April, due to low solar irradiation and short daylight, the power generated by solar array cannot satisfy the load demand, therefore, the fuel cell has to operate for most of the time to compensate this power shortage. During summer season, the frequency of fuel cell usage is reduced and the electrolyser is running more frequently to convert solar power into hydrogen.

Under the control of the proposed fuzzy logic controller, the hydrogen production rate is shown in Figure 6. The hydrogen consumed by the fuel cell is shown in Figure 7. A more clear demonstration of monthly hydrogen production by electrolyser, hydrogen consumption by fuel cell and the net hydrogen in storage overall this test period are shown in Figure 8. The results fit the seasonal profile demonstrated in the previous figures very well and clearly indicate that by using such system, the load demand can be satisfied and there is an abundance of hydrogen produced which can be used for storage, transportation or fuel cell refuelling.

Figure 9 shows the SOC of battery stack. The initial SOC was set at 50%. From the figure it can be seen that the proposed FL controller can maintain the battery's SOC around 55% during the entire year as desired, which increases the efficiency of overall system and prevents abusive use of the battery and hence contributes to extend the lifespan of the battery pack.

## V. Conclusion

This paper presented a solar powered hydrogen production, storage and utilisation system. The system uses solar power as a main power source to meet the power demand, and utilises hydrogen as an energy vector to store excess solar power via electrolyser. When required, the produced hydrogen will be converted back to electric energy using fuel cell. The proposed FL controller determines the time to produce hydrogen and to convert it back to electricity. The purpose of the FL controller is to maximise the hydrogen production and minimise the usage of the battery stack to increase the system's efficiency and to extend the lifetime of the battery stack.

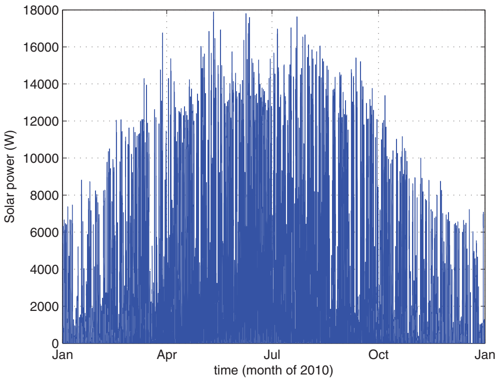


Fig. 4. Annual solar power generation

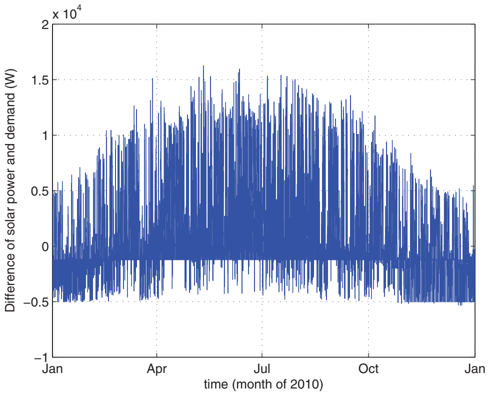


Fig. 5. Annual difference of power

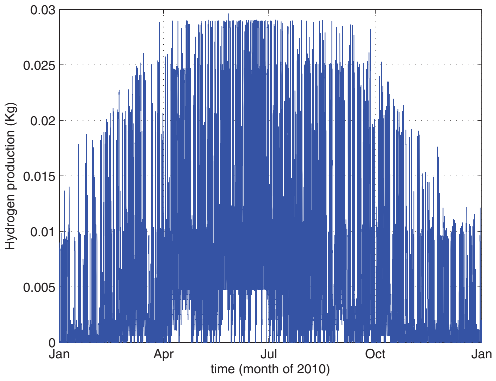


Fig. 6. Electrolyser $H_2$ production rate

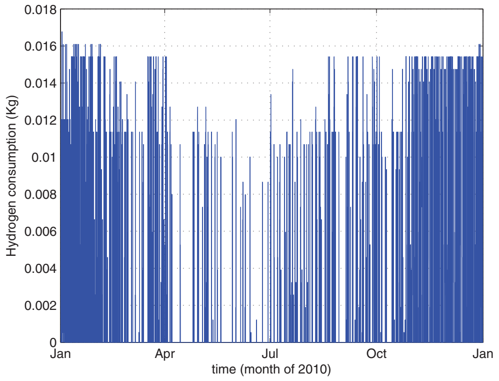


Fig. 7. Fuel cell $H_2$ consumption rate

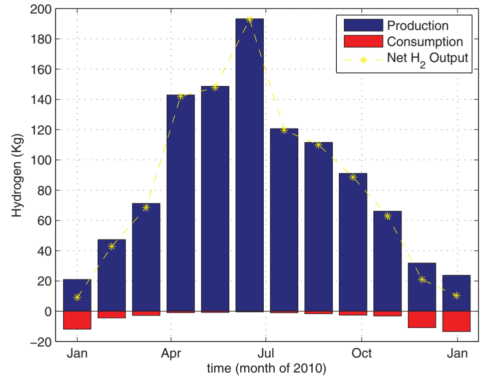


Fig. 8. Monthly variation of hydrogen production, consumption and net hydrogen output

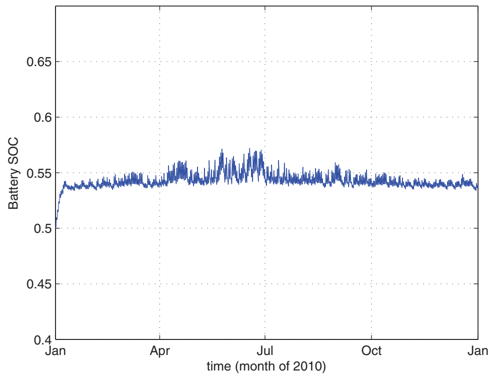


Fig. 9. SOC for entire year

The proposed system is simulated for a complete year with real solar data gathered from University of Glamorgan Hydrogen Centre. The purpose of the simulation is to evaluate the performance of the proposed controller. The numerical results shown that the proposed controller behaved as expected, it was able to satisfy the power demand and to store the hydrogen when possible while maintain the battery's SOC at desired level.

## Acknowledgment

The authors would like to acknowledge the CymruH2Wales project, part of the Low Carbon Research Institute Convergence Programme. This project has been supported by the European Development Fund through the Welsh Government.

## References

[1] R. Carapellucci and L. Giordano, "Modeling and optimization of an energy generation island based on renewable technologies and hydrogen storage systems," *International Journal of Hydrogen Energy*, vol. 37, pp. 2081–2093, 2012.

[2] C. Ziogou, D. Ipsakis, C. Elmasides, F. Stergiopoulos, S. Papadopoulou, P. Seferlis, and S. Voutetakis, "Automation infrastructure and operation control strategy in a stand-alone power system based on renewable energy sources," *Journal of Power Sources*, vol. 196, pp. 9488–9499, 2011.

[3] M. Eroglu, E. Dursun, S. Sevencan, J. Song, S. Yazici, and O. Kilic, "A mobile renewable house using PV/wind/fuel cell hybrid power system," *International Journal of Hydrogen Energy*, vol. 32, pp. 7985–7992, 2011.

[4] F. Valenciaga and C. Evangelista, "Control design for an autonomous wind based hydrogen production system," *International Journal of Hydrogen Energy*, vol. 35, no. 11, pp. 5799–5807, 2010.

[5] S. Pedrazzi, G. Zini, and P. Tartarini, "Complete modeling and software implementation of a virtual solar hydrogen hybrid system," *Energy Conversion and Management*, vol. 51, no. 1, pp. 122–129, 2010.

[6] D. Ipsakis, S. Voutetakis, P. Seferlis, F. Stergiopoulos, and C. Elmasides, "Power management strategies for a stand-alone power system using renewable energy sources and hydrogen storage," *international journal of hydrogen energy*, vol. 34, no. 16, pp. 7081–7095, 2009.

[7] G. Gómez, G. Martínez, J. Gálvez, R. Gila, R. Cuevas, J. Maellas, and E. Bueno, "Optimization of the photovoltaic-hydrogen supply system of a stand-alone remote-telecom application," *International journal of hydrogen energy*, vol. 34, no. 13, pp. 5304–5310, 2009.

[8] S. Deshmukh and R. Boehm, "Review of modeling details related to renewably powered hydrogen systems," *Renewable and Sustainable Energy Reviews*, vol. 12, no. 9, pp. 2301–2330, 2008.

[9] M. Uzunoglu, O. Onar, and M. Alam, "Modeling, control and simulation of a PV/FC/UC based hybrid power generation system for stand-alone applications," *Renewable Energy*, vol. 34, no. 3, pp. 509–520, 2009.

[10] S. Vosen and J. Keller, "Hybrid energy storage systems for stand-alone electric power systems: optimization of system performance and cost through control strategies," *International Journal of Hydrogen Energy*, vol. 24, no. 12, pp. 1139–1156, 1999.

[11] A. Bilodeau and K. Agbossou, "Control analysis of renewable energy system with hydrogen storage for residential applications," *Journal of Power Sources*, vol. 162, no. 2, pp. 757–764, 2006.

[12] K. Thanapalan, J. Williams, G. Premier, and A. Guwy, "Design and implementation of renewable hydrogen fuel cell vehicles," *Renewable Energy and Power Quality Journal*, vol. 9, 2011.

[13] Ř. Ulleberg, "Modeling of advanced alkaline electrolyzers: a system simulation approach," *International Journal of Hydrogen Energy*, vol. 28, no. 1, pp. 21–33, 2003.

[14] K. Thanapalan, J. Williams, G. Liu, and D. Rees, "Modelling of a PEM fuel cell system," in *the Proc. of IFAC World Congress*, vol. 8, 2008.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Predictive Control Strategy for a Supercritical Power Plant and Study of Influences of Coal Mills Control on its Dynamic Responses

Omar Mohamed, Bushra Al-Duri
College of Engineering and Physical Sciences
University of Birmingham
Birmingham, B15 2TT, UK
ORM808@bham.ac.uk; B.Al-Duri@bham.ac.uk

Jihong Wang
School of Engineering
University of Warwick
Coventry CV4 7AL, UK
jihong.wang@warwick.ac.uk

*Abstract*—**the paper is to investigate dynamic responses of supercritical power plants (SCPP) and study the potential strategies for improvement of their responses for Grid Code compliance. An approximate mathematical model that reflects the main features of SCPP is developed. The model unknown parameters are identified using Genetic Algorithms (GA) and the model is validated over a wide operating range. A model based predictive control (MPC) is then proposed to speed up the dynamic responses of the power plant by adjusting the reference of the plant local controls instead of direct control signal applications. Simulation results have shown encouraging improvement in performance of the plant with no interference with its associated local controllers.**

***Keywords - Supercritical Boiler; Mathematical Modeling; Parameter identification; Genetic Algorithms; Model based predictive control.***

## I. INTRODUCTION

It is well known that the Supercritical (SC) power plant is a complex process and has a large thermal inertia. Due to its once-through boiler structure, there are concerns for its dynamic response speed as there is no drum to buffer energy in the system and also there are concerns in Grid Code compliance ([1]). The first attempt towards optimal control of oil-fired SC power plants was reported in 1978 [2] with a state space model for identification and control optimization using a dynamic programming technique. Nonlinear model based predictive control (NMBPC) was reported in [3] using a reduced order physical system model to predict the next step control values. Dynamic matrix control (DMC) was published in [4] designed for SC power plants using linear model identified from step response tests. In [5], a model of an existing SC once-through power plant was reported for simulation study of plant frequency responses. The recurrent neural network modeling and modified predictive optimal control approach for coal fired SC and ultra-supercritical (USC) power plants were reported in [6] [7][8]. The paper is to study the control strategy by taking prompt actions in mill control to speed up the whole process dynamic responses.

The main contributions of the paper are: 1) a nonlinear vertical spindle mill model representing the whole milling process is integrated to a SC power plant model, which is developed by the research group at Warwick and Birmingham Universities in collaboration with the industrial partners. This has improved the whole SC power plant (SCPP) model as the previous SCPP mathematical models generally assume instantaneous response from the fuel source. In the paper, the influences of milling process capability and mill control to the whole power plant dynamic responses are investigated. 2) The paper proposed a Model Predictive Control (MPC) method to provide the updated optimal demand/set point values for the coal flow, feed water flow and the main steam valve position reference. Then those values are fed to the mill, boiler and turbine local controllers. If the amount of desired coal flow is optimally predicted in advance, there will be more stored coal in the mills to give quicker responses. The study has indicated that the proposed MPC strategy for adjusting the reference values of the plant local controls plays an important role in improving the whole plant dynamic response speeds.

## II. SC POWER PLANT DESCRIPTION AND ITS MATHEMATICAL MODEL

Vertical spindle mills are the dominant types used for SC coal fired power plants ([9-10]). The raw coal enters the mill inlet tube and carries the coal to the middle of grinding rotating table. Hot primary air flows into the mill from the bottom to carry the coal output from grinding process to the classifier that is a multi-stage separator located at the top of the mill. The heavier coal particles fall down for further grinding and the pulverized coal is carried pneumatically to the furnace. Inside the boiler, the chemical energy released from combustion is converted to thermal energy. The heat is exchanged between the hot flue gas to the water through heat exchangers. The boiler contains thin tubes as heating surfaces which form the economizers (ECON), waterwall (WW), low temperature superheater (LSH), platen superheater (PSH), final stage superheater (FSH), and reheaters (RH). The water is forced at high pressure (SC pressure) inside the economizer and passes through all those heating sections. Since pressure is above the critical point, the sub-cooled water in the economizers converted to the supercritical steam in the superheaters without evaporation. The SC steam is then expanded through turbines. The high pressure (HP) turbine is energized by the steam supplied at final stage superheater and the reheaters are used to reheat the exhausted low pressure steam from the HP turbine before it returns to the IP turbine. The mechanical power is converted to electrical power by synchronous generator coupled to the turbines. In the work described in the paper, a 600MW SC power plant is selected with the boiler specifications at boiler maximum continuous rating (BMCR) shown in Table I

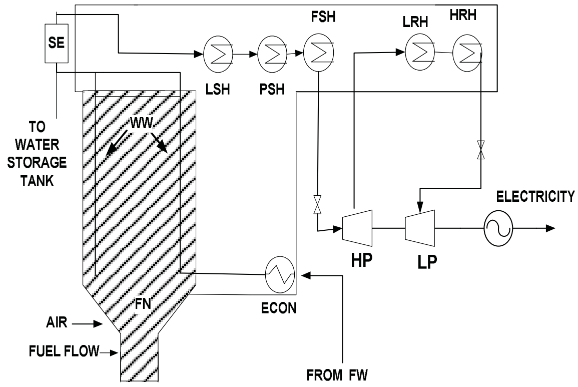


Fig.1 Schematic view of the SCPP under investigation

TABLE I. BOILER SPECIFICATION

| | |
|---|---|
| Flow rate of superheated steam(t/h) | 1780 |
| Steam pressure (MPa) | |
| FSH outlet | 25.4 |
| ECON inlet | 27.6 |
| steam temperature (C°) | |
| FSH outlet | 570 |
| ECON inlet | 288 |
| Fuel (t/h) | Pulverized coal of 276 |

For the purpose of dynamic simulation studies and control system development, a nonlinear mathematical model with 20 differential equations for supercritical boiler-turbine-generator systems, rooted from physical principles, has been developed and integrated with a vertical spindle mill model [9]. Some assumptions are made to simplify the model structures which are:

- Fluid properties are uniform at any cross section, and the fluid flow in the boiler tubes is one-phase flow.
- In the heat exchanger, the pipes for each heat exchanger are lumped together to form one pipe.
- Only one control volume is considered in the waterwall.
- The dynamic behavior of the air and gas pressure is neglected.
- Only the change in internal energy is considered, the deviations or changes of kinetic energy and potential energy of fluid are neglected.

Due to page limitation, only a brief description of the model is given in the paper, the detailed procedures for the model derivation and parameter identification are reported in our work [10][11]. The boiler model is developed by deriving the nonlinear dynamics of pressure and temperature in each heat exchanger from mass and energy balance equations of a certain control volume. Those equations are strongly coupled by the equations of SC steam flow and heat flow in the boiler. The heat flow is directly related to the fuel through constant gains and fuel calorific value. It should be noted that major boiler model parameters are either calculated from steam tables or identified using the data from certain operating unit responses. The former method is more suitable for steady state system model so the latter approach has been adopted with the real power plant measured data. The turbines HP and IP are modeled by the same principles of energy conservation and simply linked to the outlet of RH and FSH outlets of the boiler. The generator nonlinear model [12] has been coupled to the turbine model through torque equilibrium with other algebraic equations. The model has two direct inputs of feedwater flow and fuel flow and one indirect input which is the valve position reference. The model equations have been implemented by MATLAB/SIMULINK so that the output scope can be easily accessed at any point in the model. The computer graphical implementation includes gains, integrators, differentiators, transfer functions in s domain, summing points, multiplication points...etc. Furthermore, Matlab stiff function (ode15s) solver has been used for numerical solution of the model during identification or verification simulations. Fig.2 shows the model blocks diagram with all combined subsystems and the symbols are listed in Table.II. The procedure of parameter identification is summarized in the next section.

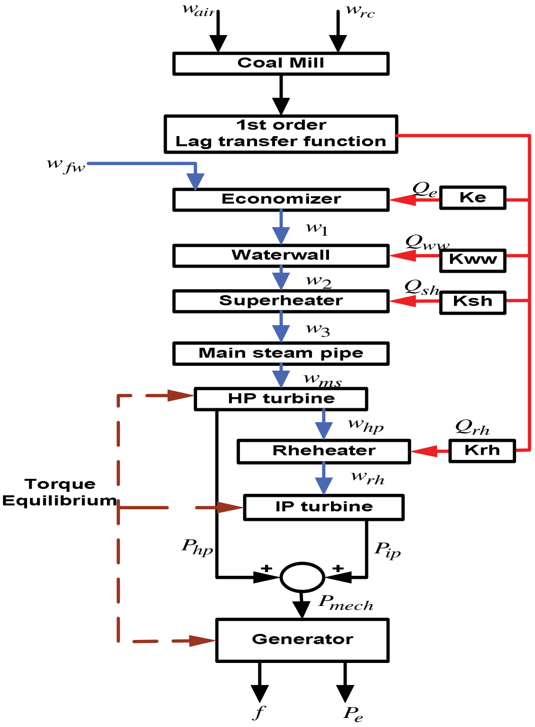


Fig.2 Mathematical model of supercritical power plan

TABLE II. LIST OF SYMBOLS IN FIG.2

| | |
|---|---|
| $w_{rc}$ : Raw coal flow rate (Kg/s). | $w_{ms}$ : Main steam flow rate (Kg/s) |
| $w_{air}$ : Primary air flow(Kg/s)<br>$w_{hp}$ : steam flow rate from HP turbine to after steam chest to the reheater (Kg/s) | $w_{rh}$ : Reheated steam flow rate (Kg/s) |
| $Q_e$ , $Q_{sh}$, $Q_{rh}$, and $Q_{ww}$: heat transferred from tube wall to the fluid (MJ/s) | $P_{mech}$ : Mechanical power (MW) |
| $w_f$ :Pulverized coal flow rate (Kg/s). | $f$ : frequency (p.u) |
| $T_{out}$ :Mill outlet temperature (C°). | $P_e$ : Electrical power (MW) |
| $w_{fw}$ : Feedwater flow rate (Kg/s). | $w_1$ , $w_2$ , $w_3$ :intermediate mass flow rates (Kg/s) |

## III. PARAMETER IDENTIFICATION

It is worth noticing that the physical model parameters are not known precisely. As described in the previous section, the GA optimization technique was adopted to identify the model unknown parameters. It should be mentioned that GA is robust optimization technique that is suitable for nonlinear system identification. Unlike conventional mathematical optimization methods, GA technique is able to tune all model parameters simultaneously with multi-objective optimization. Furthermore, GA produces global optimal solution for complex systems or functions because of parallel distributed search mechanism and mutation. The identification scheme is graphically represented in Fig.3. The work employed the coal mill model reported in [9] and the mill parameters are given in the reference. The rest of the plant model parameters are identified according to measured data responses of: 1) main steam temperature; 2) main steam pressure; 3) reheater pressure; 4) SC steam flow rate. The measured variable data for identification of turbine /generator parameter optimization are: 1) mechanical power; 2) electrical power; 3) system frequency. The onsite measurement data from a 600MW SC power plant is used for identification. Data set.1 has been used for identification which represents an increase in the load demand from 35% to 100% of load demand. Data sets 2, 3 and 4 are used for further model investigations. Fig.4 represents identification result while figs 5, 6, 7 show some verification results for different sets of data. It can be seen that the mathematical model reflects the main variation trends of the real power plant measurements over wide operating range although some assumption for simplifications were initially made.

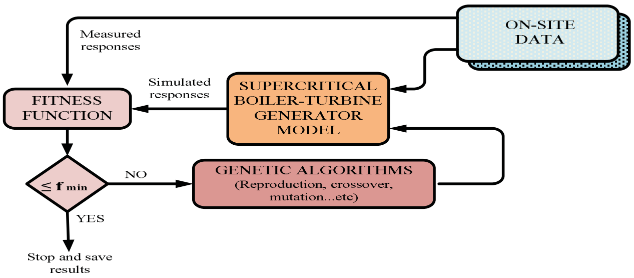

ON-SITE DATA
Measured responses
Simulated responses
FITNESS FUNCTION
SUPERCRITICAL BOILER-TURBINE GENERATOR MODEL
≤ f min
NO
GENETIC ALGORITHMS (Reproduction, crossover, mutation...etc)
YES
Stop and save results

Fig.3 Identification scheme of the model using GA.

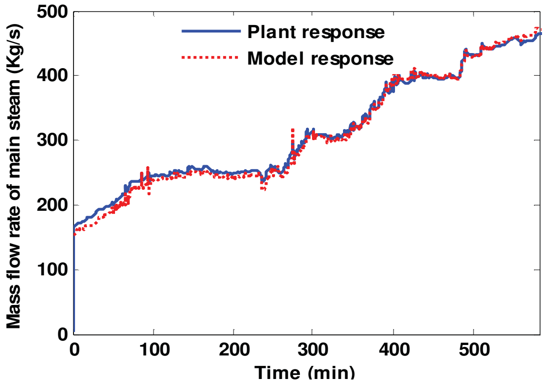


Fig.4 Mass flow (Kg/s) Data set.1 (identification)

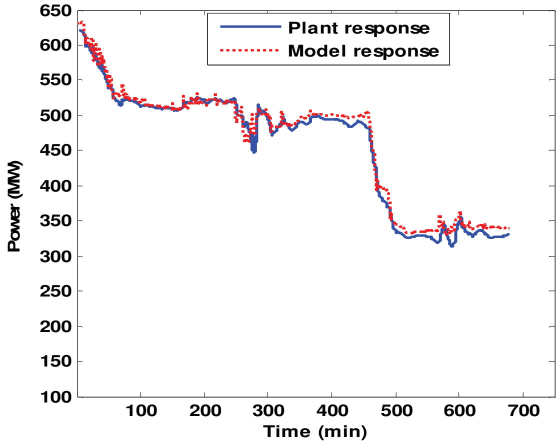


Fig.5 Electrical Power Data set.2 (verification)

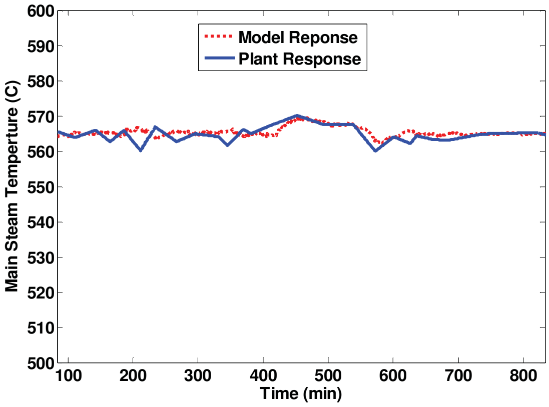


Fig.6 Main steam temperature Data set.3 (verification)

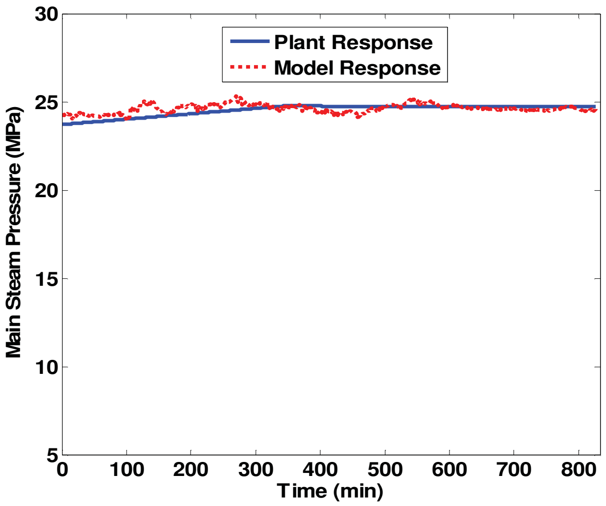


Fig.7 Main Steam pressure Data set.3 (verification)

## IV. PREDICTIVE CONTROLLER DEVELOPMENT

### A. *Reduced order linear model identification*

The generalized MPC algorithm includes an identified linear state space model, used for predicting the plant output variables. The plant identified linear model were investigated around nominal operating conditions (i.e. supercritical conditions). Again GA has been used to identify the linear model with portions of data set.1 and 3 to match the response of the original process model. The prediction model has four states three inputs and three outputs. the linear model which has the following form:

$$\boldsymbol{x}(k+1) = \boldsymbol{Ax}(k) + \boldsymbol{Bu}(k) \qquad (1)$$

$$\boldsymbol{y}(k) = \boldsymbol{Cx}(k) \qquad (2)$$

The model has four states $\boldsymbol{x}^T=[x_1 \ \ x_2 \ \ x_3 \ \ x_4]^T$, three inputs or manipulated variables $\boldsymbol{u}^T=[u_1 \ \ u_2 \ \ u_3]^T$, and three outputs.$\boldsymbol{y}^T=[y_1 \ y_2 \ \ y_3]^T = [x_2 \ \ x_3 \ \ x_4]^T$. $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$, are the normalized state space model matrices. The parameters of the digitized model are:

$$\boldsymbol{A} = \begin{bmatrix} 0.7687 & 0.05486 & 0 & 0 \\ 0.7537 & 0.05379 & 0 & 0 \\ -0.003729 & -0.0002662 & 4.041\text{e-}018 & 0 \\ 18.94 & -23.78 & 0 & 1.0001 \end{bmatrix}$$

$$\boldsymbol{B} = \begin{bmatrix} 0.008409 & 0.01259 & -0.3041 \\ 0.008181 & 0.01243 & -0.3318 \\ 0.9934 & 1.249 & 21.75 \\ 0.2056 & 0.3123 & -8.338 \end{bmatrix}, \boldsymbol{C} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

The inputs and the outputs which have been used for identification of the controlled plant are chosen as:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} \text{Main steam pressure (MPa)} \\ \text{Electrical Power (MW)} \\ \text{Main steam temperature } (\text{C}^{\text{o}}) \end{bmatrix}$$

$$\begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} \text{Feedwater flow reference (Kg/s))} \\ \text{Raw coal flow reference (Kg/s)} \\ \text{Valve Position reference(p.u)} \end{bmatrix}$$

### B. *The generalized Predictive control strategy*

A model based predictive control is developed with provisions of unmeasured disturbances and measurement noises to be used for compensation around the investigated operating conditions. Here, the linear time invariant model is used for the MPC algorithm while the process mathematical model has been used to simulate the power plant responses. The controller setup is supposed to generate such states naturally by default. In this research, the generalized predictive controller algorithm described in [13] is adopted which has been widely used for chemical or thermodynamic process control [13] [14] [15] [16]. The prediction model has been upgraded as follows:

$$\boldsymbol{x}(k+1) = \boldsymbol{Ax}(k) + \boldsymbol{B}_u\boldsymbol{u}(k) + \boldsymbol{B}_v\boldsymbol{v}(k) + \boldsymbol{B}_W\boldsymbol{w}(k) \qquad (3)$$

$$\begin{aligned} \boldsymbol{y}(k) &= \boldsymbol{y}(k) + \boldsymbol{z}(k) \\ &= \boldsymbol{Cx}(k) + \boldsymbol{D}_u\boldsymbol{u}(k) + \boldsymbol{D}_v\boldsymbol{v}(k) + \boldsymbol{D}_W\boldsymbol{w}(k) + \boldsymbol{z}(k) \end{aligned} \qquad (4)$$

Where $\boldsymbol{v}$ is the measured disturbance and $\boldsymbol{w}$ is the unmeasured disturbance vector, $\boldsymbol{z}$ is the measurement noise. The adopted predictive control algorithm is quite analogous to LQG procedure, but with implication of the operational constraints. The prediction is made over a specific prediction horizon. Then the optimization program is executed on-line to calculate the optimal values of the manipulated variables to minimize the objective function below:

$$\xi(k) = \sum_{i=H_W}^{H_p} \left\| \boldsymbol{y}(k+i|k) - \boldsymbol{r}(k+i|k) \right\| \mathbf{Q} + \sum_{i=0}^{H_c-1} \left\| \boldsymbol{\Delta u}(k+i|k) \right\|^2 \mathbf{R} \ (5)$$

The weighting coefficients (**Q** and **R**), control interval ($H_w$), prediction horizon ($H_p$) and control horizon ($H_C$) of the performance objective function will affect the performance of the controller and computation time demands. The terms $\boldsymbol{r}$ represents the demand outputs used as a reference for MPC model and $\boldsymbol{\Delta u}$ is the change in control values for $H_C$ number of steps. Zero-order hold method is then used to convert the control signals from discrete time to continuous time to be fed to the plant. The inputs/outputs constraints are determined according to the power plant operation restrictions, which are expressed as the maximum and the minimum allowable inputs:

$$\boldsymbol{u}_{min} \leqslant \boldsymbol{u} \leqslant \boldsymbol{u}_{max} \qquad (6)$$

$$\Delta\boldsymbol{u}_{min} \leqslant \Delta\boldsymbol{u} \leqslant \Delta\boldsymbol{u}_{max} \qquad (7)$$

The optimization problem is to find the control moves for each manipulated variable, i.e. the MPC control law:

$$\min_{\Delta\boldsymbol{u},\dots,\Delta\boldsymbol{u}(k+1+H_C)} \xi(k) \qquad \text{subject to (6) and (7)}$$

The quadratic programming (QP) solver, with active set method or interior point method, is commonly used to solve control law problem of the MPC. In the interest of predictive controllers for thermal power stations, the generalized MPC approaches and DMC algorithms are reported for control of power plants once-through and drum type units. [3][4][5][7][14][15][16][17]. To show the influences of coal mill control on the plant output responses, a controller is implemented to regulate the primary air fan and the other is implemented to regulate the coal feeder speed. Both receive the MPC coal flow signal as adjuster for their reference. With the MPC strategy described above, simulations have been conducted. The whole package of the proposed strategy is shown in fig.8. Simulation results are presented in the next section.

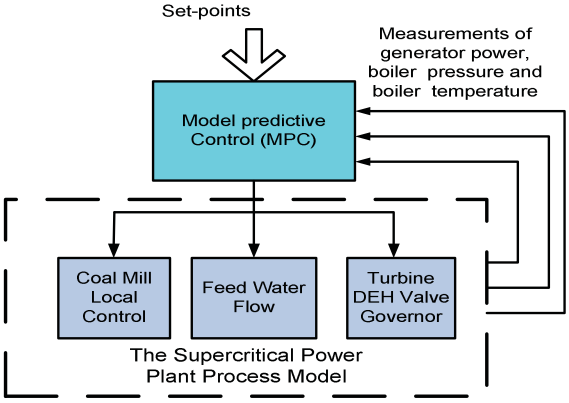

Set-points
Measurements of generator power, boiler pressure and boiler temperature
Model predictive Control (MPC)
Coal Mill Local Control
Feed Water Flow
Turbine DEH Valve Governor
The Supercritical Power Plant Process Model

Fig.8 Predictive controller scheme

## V. SIMULATION STUDIES

MPC tuning is finalized by selection appropriate values for the prediction horizon $H_p$, control horizon $H_C$, and weighting matrices **Q** and **R.** The control interval, prediction, and control horizons are found to be 1, 35, 5seconds respectively. **Q**= [1 1 1] and **R**=[0.1 0.1 0.1]. Simulating different scenarios have lead to this selection. In this scenario, a step change of ±20MW in the power is assumed as set-point signal, the pressure set-point is rescheduled from look-up table which relates the power set-point to the pressure, and the temperature set-point is constant of 570C°. In the reported results for Case A represent the improved case with using MPC as correction to the mill local control, boiler feedwater flow, and turbine valve controller and Case B represent existing milling and plant performance. From the reported results, the improvements are obvious in case of using the MPC without violating the practical constrains of the various plant variables. Thus the primary air fan and feeder speed can be regarded as other supplementary means to improve the power primary response, not only acting the turbine expansion valves.

Furthermore, the boiler steam pressure and temperature have less fluctuation around the set-point which helps in extending the life of the equipment. The pulverized coal flow to the furnace, the feedwater flow, and valve position are mentioned in fig.10, more pulverized coal is discharged to the furnace from the mills per time unit which means more coal is combusted and more energy is delivered from the boiler to give quicker responses. Hence, the MPC and its associated strategy for reference correction control play an important role in improving the plant responses and satisfying the regulations of the national grid code. Especially when increasing the grinding capability of the mills and pulverized coal discharging speed. Fig.11 shows major mill variables. High mill differential pressure and primary air pressure are created to carry more coal flow to the burners. Also higher raw coal is initially dropped in the mill because of the improved feeder speed response. The mass of raw coal and pulverized coal in the mill is higher to provide the required flow of pulverized coal in a timely manner. The only penalty which has been paid is that more current, and consequently, power is consumed from the mills to increase the grinding capability of the mills.

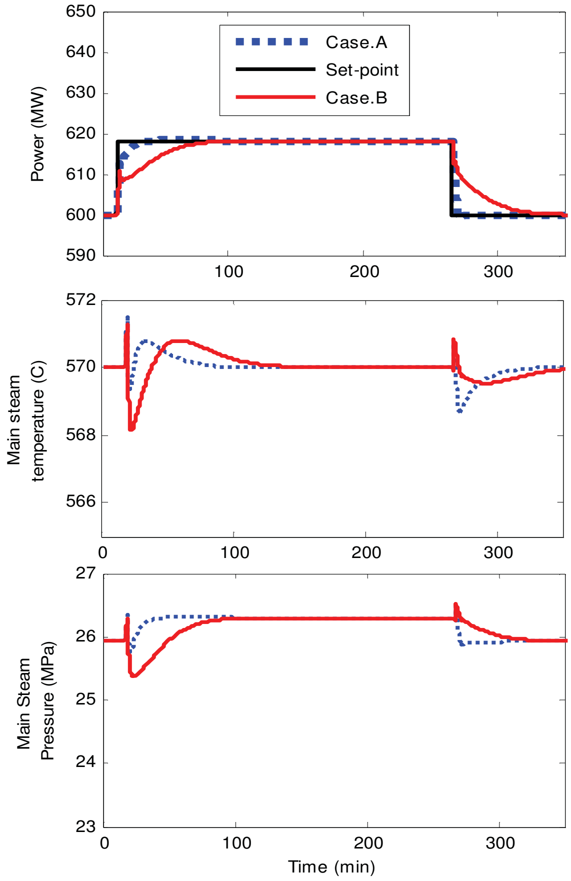


Fig.9 Controlled variables of the SCPP

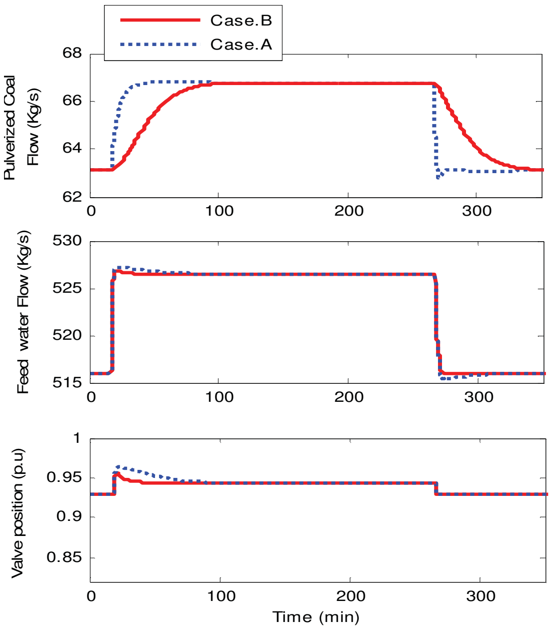


Fig.10 input variables to the boiler-turbine-generator system

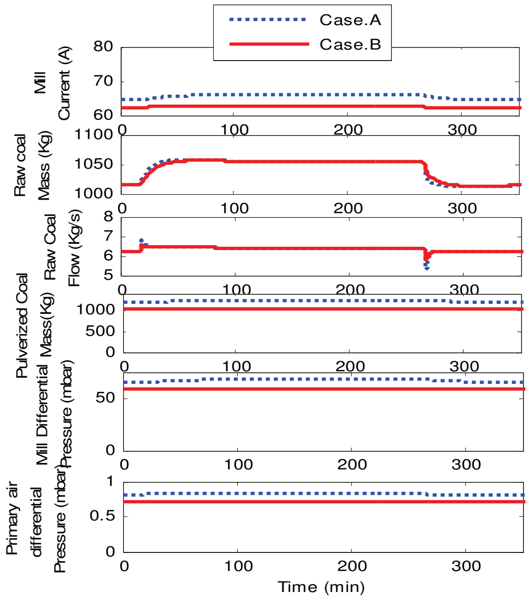


Fig.11 variables of each mill in service

## VI. CONCLUSION AND FUTURE RESEARCH

In this paper, a complete power plant process model including fuel preparation milling process. The model prepared a platform for us to investigate the influences of mill control to the whole power plant responses. A new strategy of applying model predictive control is reported and the new contribution for the strategy is to use the MPC to update the control set-up/desired values instead of tuning the individual local loop controllers. This improved the power plant responses over the existing control strategy. As a future recommendation, it is suggested to extend the method of modeling and control to ultra-supercritical power plant.

## ACKNOWLEDGEMENT

The authors would like to thank EPSRC (UK) for the research grant support (EP/G062889/1) and AWM/ERDF Birmingham Science City Energy Efficiency and Demand Reduction project.

## REFERENCE

[1] National Grid (UK), The Grid Code, Issue 4 Revision.5, 2010. http://www.nationalgrid.com/uk/Electricity/Codes/gridcode/gridcodedocs/ on 12th Feb 2012.

[2] H. Nakamura, H. Akaike "Statistical Identification for Optimal Control for Supercritical Thermal Power Plants". Automatica. Vol.17, No.1, pp 143-155, 1981.

[3] B.P. Gibbs , D.S. Weber, D.W. Porter, "Application of Nonlinear Model Predictive Control to Fossil Power Plants" proceedings of 30th conference on decision and control, Dec 1991. pp.1850-1856.

[4] J. A. Rovnak, R. Corlis, "Dynamic Matrix Based Control of Fossil Power Plant", IEEE Transactions on Energy conversions. Vol. 6, No. 2, pp. 320–326, 1991.

[5] T. Inoue, H. Taniguchi, Y. Ikeguchi "A Model of Fossil Fueled Plant with Once-through Boiler for Power System Frequency Simulation Studies", IEEE Transactions on Power Systems, vol. 15, No. 4, pp1322-1328, 2000.

[6] J. Kwang ,Y. Lee, J.S. Heo, J. A. Hoffman, S-H Kim, and W-H Jung, "Neural Network Based Modeling for Large Scale Power Plant" Power Engineering Society General Meeting, 2007. IEEE Volume, Issue 24-28, June 2007, pp.1 – 8.

[7] K. Lee , J.S. Heo, J.A. Hoffman, S-H. Kim, W-H. Jung, "Modified Predictive Optimal Control Using Neural Network-based Combined Model for Large-Scale Power Plants" Power Engineering Society General meeting, June 2007, pp. 1-8.

[8] K. Lee , J. H. Van Sickel, J. A. Hoffman,W-H. Jung, and S-H. Kim, "Controller Design for Large-Scale Ultrasupercritical Once-through Boiler Power Plant", IEEE Transaction on Energy Conversions, vol.25,No.4,pp1063-1070,2010.

[9] Y.G., Zhang, Q.H. Wu, J. Wang, G. Oluwanda, D. Matts, D., and X.X. Zhou, "Coal mill modelling by machine learning based on on-site measurement", IEEE Transactions on Energy Conversion, Vol.17. No.4, pp549-555, 2002.

[10] O. Mohamed, J. Wang "Modeling and simulation study of coal fired power generation with supercritical boiler for grid code compliance" Internal technical report, School of Engineering University of Birmingham, 2010.

[11] O. Mohamed, J. Wang, B. Al-Duri " Mathematical Modeling of Coal Fired Supercritical Power Plant and Model Parameter Identification Using Genetic Algorithms". Lecture notes in Electrical Engineering, " Electrical Engineering and Applied Computing", Chapter.1. Springer .July. 2011

[12] Yao-Nan. Yu "Electric Power System Dynamics" Academic Press, 1983.

[13] N.L. Ricker, Model predictive control with state estimation, Industrial and Engineering Chemistry Research. Vol.29, No.3, pp374-382. 1990.

[14] G. Poncia, S. Bittani " Multivariable Model Predictive Control of a Thermal Power Plant with Built-in Classical Regulation". International Journal of Control, Vol. 74, No. 11. pp 1118-1130. 2001.

[15] G. A. Oluwande" Exploitation of advanced control techniques in power generation". Computing and Control Engineering Journal, Vol. 12, No. 2, pp 63-67, April 2001.

[16] J. A. Rossiter, B. Kouvaritakis, R.M. Dunnett " Application of Generalised Predictive Control to Boiler-Turbine Unit for Electricity Generation" IEE Proceedings, Vol. 138, No.1 , 1991.

[17] A. W. Ordys and M. J. Grimble "Predictive Control in Power Generation" Computing and Control Engineering Journal, Vol. 10, No. 5, pp 214-220, 1999.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Energy Management Effects of Integrating Regenerative Braking into a Renewable Hydrogen Vehicle

Kary Thanapalan, Fan Zhang, Giuliano Premier, Jon Maddy, Alan Guwy
Sustainable Environment Research Centre (SERC)
University of Glamorgan
Pontypridd CF37 1DL, United Kingdom
E-mail:kthanapa@glam.ac.uk

*Abstract*—**This paper discusses the design and route to implementation of a regenerative braking system for a Renewable Hydrogen Hybrid Electric Vehicle (RHHEV) that would performs regenerative energy recovery based on vehicle attributes, thereby providing improved performance. A dynamic model of a regenerative braking system for a RHHEV has been derived and implemented in MATLAB/Simulink. The model is then incorporated into a simulation model of a RHHEV to analyze the energy management effects of integrating the regenerative braking into the system. Simulations are carried out and the results are discussed in this paper. The results indicate that this approach provides an improvement in performance and fuel efficiency to the RHHEV system.**

***Keywords – Hydrogen vehicles, Renewable energy, Regenerative braking, Energy sources, Energy recovery***

## I. INTRODUCTION

The utilization of regenerative braking (RB) makes the hydrogen fuel cell vehicles more attractive by better utilising the on board storage capacity for hydrogen, through raised energy efficiency. RB systems could recover the energy that would otherwise be lost through braking [1]. The results from previous research efforts show that, significant amount of energy can be recovered by the integration of a regenerative braking mechanism [2]. Further research effort should focus on the development of a systematic and integrated regenerative braking system to recover as much energy as possible, while maintaining suitable power flows in the other subsystems. To this end, several researchers have developed many different configurations of energy recovery mechanisms to continuously recover otherwise dissipated energy that occurs due to road irregularities, vehicle acceleration, and braking; see for example [3]-[9].

In this paper, developments in energy recovery mechanisms for renewable hydrogen hybrid electric vehicle (RHHEV) systems are described. Essentially and in general terms there are two approaches to such energy recovery mechanisms; implementation of a better energy management system by providing optimal powertrain topologies applied to the various energy sources and incorporating an energy recovery mechanism [10]. This work addresses both approaches by describing an optimal power management strategy for RHHEV systems for better energy management and the development of a regenerative braking system (RBS) for energy recovery. This paper begins with a discussion of RB techniques and this is followed by the incorporation of the RB system into an optimal power management strategy for RHHEV and analyses.

## II. REGENERATIVE BRAKING SYSTEM

The energy efficiency and driving range of the RHHEV system may be increased through the development and incorporation of RBS. The regenerative braking methods have the capability to increase hybrid electric vehicle (HEV) driving range by approximately10-30% [11]. Further benefits of RBS are that they can reduce the drawdown of battery charge, extend the overall life of the battery pack and reduce fuel consumption. However, the main problem with RBS is that at low speed, the electric motor may not be able to produce enough torque to stop the vehicle [12], [13]. The goal of the RBS is to recover some of the energy otherwise wasted as braking heat and store it, and then later use it for vehicular motion. The principle of a RBS circuit is shown in Fig.1. This is commonly known to represent the conditions for four quadrants of possible operation for the regenerative braking [14], [15].

In forward motoring (quadrant I) $v_a$ , $E_g$ , and $I_a$ are all positive. The torque and speed are also positive in this quadrant. During forward braking (quadrant II), the motor runs in the forward direction and the induced emf $E_g$ will continue to be positive. For the torque to be negative and the direction of the energy flow to reverse, the armature current must be negative. The supply voltage $v_a$ should be keep less than $E_g$ . In reverse motoring (quadrant III), $v_a$ , $E_g$ , and $I_a$

are all negative. The torque and speed are also negative in this quadrant. To keep the torque negative and the energy flow from the source to the motor, the back emf $E_g$ must satisfy the condition $|v_a| > |E_g|$.

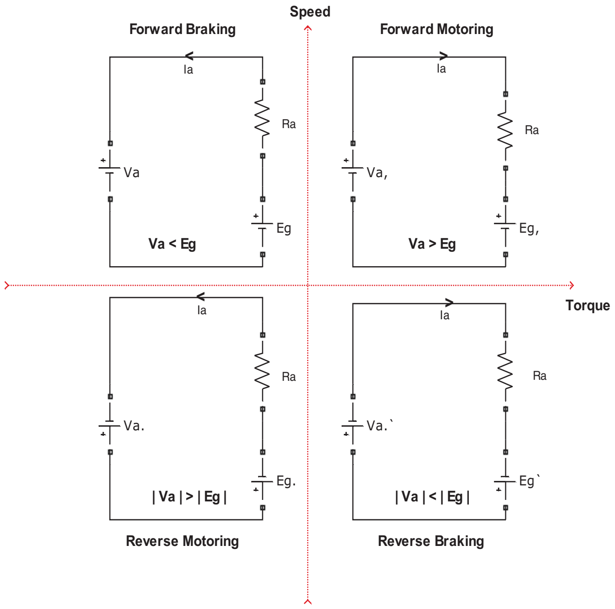


Fig.1 Principle of RBS

The polarity of $E_g$ can be reversed by changing the direction of field current or by reversing the armature terminals. During reverse braking (quadrant IV), the motor runs in the reverse direction. $v_a$ and $E_g$ will continue to be negative. For the torque to be positive and the energy to flow from the motor to the source, the armature current must be positive. The induced emf $E_g$ must satisfy the condition $|v_a| < |E_g|$.

Using this fundamental principle, there are many different types of energy recovery system that have been developed. These range from electro-hydraulic braking (EHB) systems to energy harvesting shock absorber [16]-[20]. In this paper, a simulation model of a simple regenerative braking system is developed and implemented in MATLAB/Simulink™ and incorporated into a vehicle simulator previously developed [21] to analyse the effect of integrating the RBS. The vehicle model used here is for the University of Glamorgan hydrogen bus (UoGHB). The powertrain of UoGHB consist of 12kW PEM fuel cell stack developed by Hydrogenics, a 288v, 132 Amp-hr battery pack, 375v, 63F Maxwell ultracapacitor and 70kW DC motor. Simulations are carried out on the New European Drive Cycle (NEDC) shown in Fig.2.

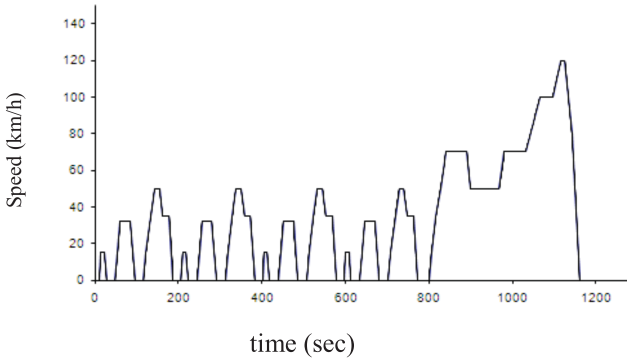


Fig.2 New European Drive Cycle

The combined power sources (fuel cell, battery and ultracapacitor) in a RHHEV provide a more viable solution than a single power source renewable vehicle [10]. The fuel cell system provides low to medium constant power delivery, whereas the lead acid battery pack provides medium constant power. Finally the ultracapacitors (UC) deliver a large, instantaneous power demand, which is ideal for high load acceleration and regeneration from braking power [22].

The regenerative braking is only possible, if there is enough torque applied to the wheels, in accordance with the driving conditions. These requirements may be expressed in and governed by the following equations [2], [5], [6]:

The regenerative torque applied to the front wheels $\tau_R$ is given by;

$$\tau_R = \eta\phi\tau_{RB} \qquad (1)$$

Where, $\eta$ is the efficiency, $\phi$ is the continuously variable transmission speed ratio and by considering that the motor has a reversible performance characteristic curve; $\tau_{RB}$ is the regenerative torque provided by the motor which is determined from the motor characteristic curve for the given speed [2] and is described as follows;

$$\tau_{RB} = \varpi_f(b_{soc}, v)\tau_m \qquad (2)$$

Where, $\varpi_f$ is the weighting factor and it depends on the battery state of charge $b_{soc}$ and vehicle velocity $v$. $\tau_m$ is the electric motor torque. The regenerative braking force at the wheels $F_{RB}$ is obtained as;

$$F_{RB} = \tau_R . r_t^{-1} \qquad (3)$$

Where, $\tau_R$ is the regenerative torque applied to the front wheel, $r_t$ is the tyre radius and $F_{bf}$ is the required front wheel braking force. If $F_{RB} > F_{bf}$ then the front wheel is braked only by the regenerative brake. Otherwise the regenerative braking needs to be combined with conventional friction braking. Suppose, the hydraulic friction brake works with the regenerative brake, then in this case the change in hydraulic pressure in the front wheel cylinder $\Delta P$ is given by;

$$\Delta P = \tau_{RB}(\lambda A \times 2r_e)^{-1} \quad (4)$$

Where, $\lambda$ is the friction coefficient, $A$ is the front wheel cylinder area and $r_e$ is the effective radius. So, the supplied hydraulic pressure $P_h$ is calculated as;

$$P_h = P_s - \Delta P \quad (5)$$

Where, $P_s$ is the master cylinder pressure. Finally, the total braking torque $\tau_{BK}$ at the front wheel is given by;

$$\tau_{BK} = \tau_R + \tau_{fr} \quad (6)$$

Where, $\tau_{fr}$ is the torque provided by conventional friction braking.

From the above analyses it is clear that the amount of regenerative brake energy depends on available motor torque, which will vary according to battery state of charge (SOC), motor speed and corresponding driver demands. It is desirable that the system use the least possible stored energy to provide the required motive or ancillary demands. This can be achieved by using energy saving mechanisms or power management algorithms [23]. Therefore, in this paper an optimal power management strategy is also integrated into the configuration of a RHHEV system with RBS. The system analyses and performances are described in the following sections.

## III. Power management of RHHEV with RBS

RHHEV has become increasingly of interest in the powertrain industry due to the finite supply of fossil fuels and the effects fossil fuels are having on the environment [10]. Hybrid electric vehicle (HEV) developments play an important role in the realisation of the renewable hydrogen hybrid electric vehicles. $CO_2$ emission reduction and possible fuel economy benefits are attractive functions of HEV systems. RHHEV's are expected to make a significant contribution to the environmental needs/demands of the premium vehicle sector. However, combining a polymer electrolyte membrane fuel cell (PEM FC) stack, UC module, with a battery pack; while managing power flow and meeting the high expectations of the market, presents challenges in the area of system configuration and controller design [10]. The system configuration and controller design for HEV systems can be complex and challenging. The control system design requires a high level of integration with existing systems on the vehicle.

The literature contains several publications concerning various aspects of HEV system power distribution methodologies and controller design; see for example, [9], [18], [19]. Broader details of the subject can be found in various reports, which summarise the power management aspects and problems experienced with HEV [22]-[24]. These reports also suggest some technical solutions and analytical methods for application to some of the problems. These range from control algorithms for global optimization, based on *a priori* knowledge of a scheduled driving cycle, to real-time power management based on optimal control theory [23].

Previous power management design efforts focussed on the design methods to determine the component sizes that minimized the cost of the power system elements in HEV [25]. Furthermore, power management strategies and component sizing are often coupled together, which implies that different component sizing might be associated with different power management strategies. Therefore both should be considered as a combined package. Many other research efforts on power distribution in HHEV have been expanded. These include; the use of an equivalent consumption minimization strategy to determine the real time optimal power distribution of a fuel cell/UC hybrid vehicle [26]. In this study, an optimal power management strategy for HHEV systems described in [23] is adopted and integrated with the RBS, in order to maximize the benefit of the RBS. The power source configuration of the RHHEV is shown in Fig.3.

The DC-link is the connection mechanism between the electrochemical power source and the motor. The DC-link is kept at a constantly high voltage, thereby assuring the highest possible motor torque over the whole speed range. Since the voltage of the power sources may vary, a DC/DC converter is needed in the configuration (see Fig.3).

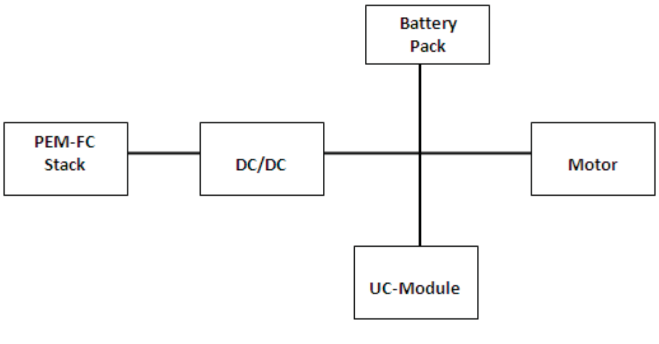


Fig.3 Power source configuration of a HHEV system

The power balance equation for the overall RHHEV system is given by;

$$p_{fc} + p_{uc} + p_b = p_d \qquad (7)$$

where $p_{fc}$, is the output power of the PEM fuel cell stack, $p_{uc}$, the output power of the ultracapacitor, $p_b$ the output power of the battery pack and $p_d$, the power demand. The regenerated power of the battery is calculated as;

$$p_{RB} = p_b - i^2 R \qquad (8)$$

The combined system with RBS and optimal power management (PM) is implemented in MATLAB/Simulink™ and simulated using the UoGHB vehicle simulator [21].

## IV. Results and Discussion

The simulation results show that by using this regenerative braking (RBS) mechanism with an optimal power management strategy, a considerable amount of energy can be recovered. The results indicate that the combined optimal power management strategy with RBS has significantly improved the RHHEV system efficiency and total power recovery.

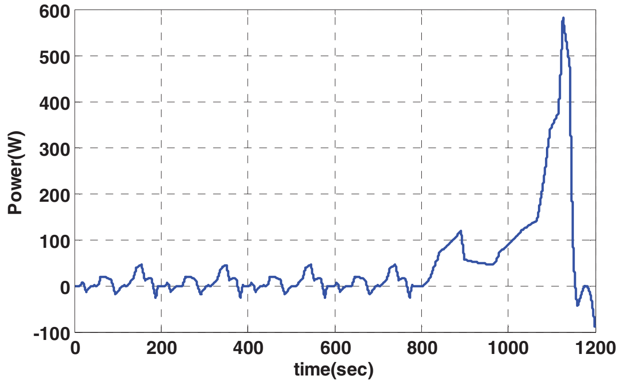


Fig.4 power requirements

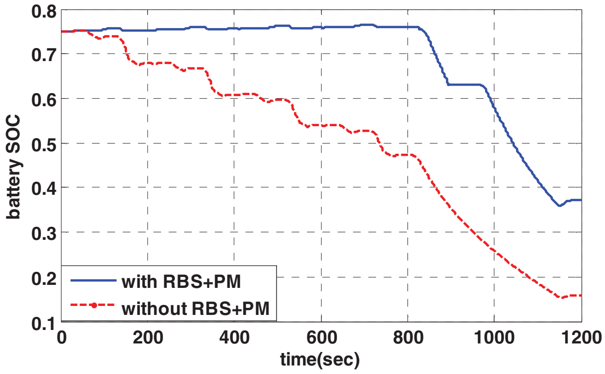


Fig.5 Compared responses of battery SOC

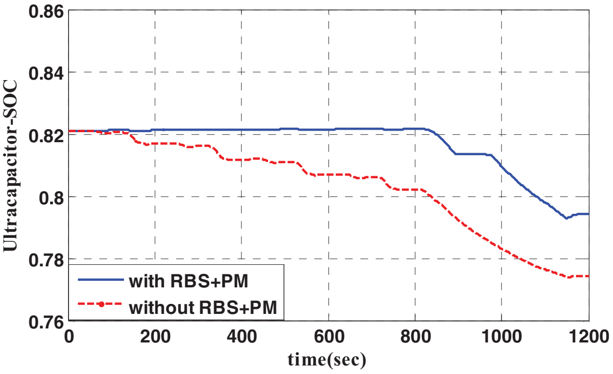


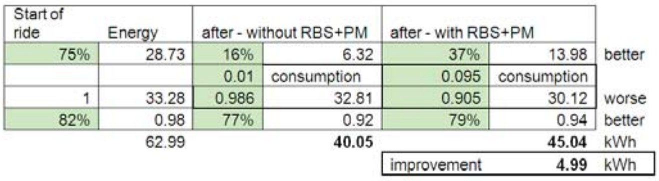
Fig.6. Compared responses of Ultracapacitor- SOC

In Fig.4 the total power requirements for the whole NEDC is shown. The combined power sources (battery, ultracapacitor and fuel cell) will provide the power to RHHEV for the required driver demand. Fig.5 & 6 shows a comparison of the charge/discharge characteristics of battery pack SOC and ultracapacitors respectively. It may be realised from these results that, due to the regenerative energy recovery and the inclusion of the optimal power management strategy the state of charge of these power sources are kept high (apart from at higher power demand) and able to provide constant power. The energy required in any given HEV system depends on the systems design and operation characteristics and the quantity and type of losses encountered during vehicular motion. In this paper the energy requirement for the HHEV system is analysed with reference to the UoGHB. The system model includes the regenerative energy recovery mechanism, thus it is of interest from an energy system point of view. It is therefore, essential to search for design and operating conditions which lead to reduction of energy dissipation and consequently lower production cost to promote the HHEV system. To this end, the UoGHB model is used to show the improvement of energy consumption via system analysis, which includes the regenerative energy recovery and power management.

Summary of energy consumption for both cases (with and without RBS plus PM) are shown in Table 4.1. In energy terms the maximum energy capacity of the battery pack and UC modules installed in the UoGHB is 38.3 kWh and 1.2 kWh respectively and energy capacity of the hydrogen fuel tank is 33.28 kWh. Usually battery/UC states of charge are kept to particular limits (or constant) to increase the life of the battery. This leads to variable usage of hydrogen and it can be controlled by incorporating an optimal control algorithm and energy saving mechanisms. In this paper for the simulation analysis, the NEDC is used and at the start of the journey it is assumed that the hydrogen tank is full and the battery SOC and UC-SOC are at 75 and 82 % respectively (see Fig.5 & 6). In energy term this is about 28.73 kWh and 0.98 kWh respectively. If the RBS and PM are not included, at the end of the journey, energy left in the battery is about 6.32 kWh (16%) and in UC is about 0.92 kWh (77%). With the inclusion of the RBS and PM the energy remaining in battery and UC is about

13.98 kWh (37%) and 0.94 kWh (79%) respectively. However, the hydrogen fuel usage was high in comparison with the system without RBS and PM. This is due to the arrangements of power management and system configuration. In this case the FC-stack is considered to be the main power source to supply the power. But, it should be noted that overall a significant improvement is achieved in energy saving by incorporating RBS and PM. The improvement is about 4.9 kWh energy saving. Thus, the system efficiency is improved by about 27%.

Table 4.1 summary of energy consumption



| Start of ride | Energy | after - without RBS+PM | | after - with RBS+PM | | |
|---|---|---|---|---|---|---|
| 75% | 28.73 | 16% | 6.32 | 37% | 13.98 | better |
| | | 0.01 | consumption | 0.095 | consumption | |
| 1 | 33.28 | 0.986 | 32.81 | 0.905 | 30.12 | worse |
| 82% | 0.98 | 77% | 0.92 | 79% | 0.94 | better |
| | 62.99 | | **40.05** | | **45.04** | kWh |
| | | | | improvement | **4.99** | kWh |

The energy consumption in any HEV system is one of the important parameters that dictate the choice of the HEV. In addition to the energy consumption the cost of the technology and the final cost of the vehicle determine the choice of the HEV system.

## V. Concluding Remarks

In this paper, an investigation of the development of a RBS with an optimal PM strategy for a RHHEV system was carried out. A dynamic model of a RBS with an optimal PM strategy has been described and implemented in MATLAB/Simulink. The model was then incorporated into a UoGHB simulator to analyse the effect of the integration of the regenerative braking mechanism. The results show that if such mechanisms are used in the UoGHB, about 4.9kWh energy could be saved during the whole NEDC. Thus, it can be concluded that in general using such a mechanism in any HHEV system is likely to a significant amount of energy being recovered.

## Acknowledgment

This work was supported by the CymruH2Wales project, part of the Low Carbon Research Institute Convergence Programme. The CymruH2Wales project has been supported by the European Regional Development Fund through the Wales Government.

## References

[1] J. Guo, J. Wang, and B. Cao, "Regenerative Braking Strategy for Electric Vehicles", In the Proceedings of the IEEE- Intelligent Vehicles Symposium, Xi'an, China, pp.864-868, 2009

[2] H. Yeo and H. Kim, "Hardware in the loop simulation of regenerative braking for a hybrid electric vehicle", *Proceedings of the Institution of Mechanical Engineering, Part D: Journal of Automobile Engineering,* vol.220, pp.855-864, 2002

[3] P. Karthikeyan, CHS. Chaitanya, NJ. Raju and SC. Subramanian, "Modelling an electropneumatic brake system for commercial vehicles", *IET Electrical Systems in Transportation*, vol.1, No.1, pp.41-48, 2011

[4] H. Bai, Y. Zhang, C. Semanson, C. Luo, and CC. Mi, "Modelling, design and optimisation of a battery charger for plug-in hybrid electric vehicles", *IET Electrical Systems in Transportation*, vol.1, No.1, pp.3-10, 2011

[5] M. Ye, Z. Bai, and B. Cao, "Robust control for regenerative braking of battery electric vehicle", *IET Control Theory and Applications*, vol.2, No.12, pp1105-1114, 2008

[6] A. Rousseau, P. Sharer, and R. Ahluwalia, "Energy storage requirements for fuel cell vehicles", *SAE International*, Paper No. 2004-01-1302, 2004

[7] JK. Ahn, KH. Jung, DH. Kim, HB. Jin, HS. Kim and SH. Hwang, "Analysis of a regenerative braking system for hybrid electric vehicles using an electro-mechanical brake", International Journal of Automotive Technology, vol.10, No.2, pp.229-234, 2009

[8] MD. Galus and G. Andersson, "Power system considerations of plug-in hybrid electric vehicles based on a multi energy carrier model", *IEEE-Power and Energy Society General Meeting*, Calgary, Alberta, Canada, pp.1-8, 2009

[9] CC. Chan, "The state of the art of electric, hybrid, and fuel cell vehicles, *In Proceedings of the IEEE*, vol.95, No.4, pp.704-718, 2007

[10] KKT. Thanapalan, F. Zhang, AD. Procter, SJW. Carr, GC. Premier, and AJ Guwy, and J. Maddy "Development of energy saving mechanism for renewable hydrogen vehicles", *Renewable Energy and Power Quality Journal*, vol.1, No.10, 353, 2012

[11] J. Zhang, X. Lu, J. Xue, and B. Li, "Regenerative braking system for series hybrid electric city bus", *The World Electric Vehicle Journal*, vol.2, No.4, pp.128-134, 2008

[12] SR. Cikanek and KE. Bailey, "Regenerative braking system for a hybrid electric vehicle", *In the Proc of the American Control Conference, Anchorage, AK, USA*, pp.3129-3134, 2002

[13] F. Wang, and B. Zhuo "Regenerative braking strategy for hybrid electric vehicles based on regenerative torque optimization control", *Journal of Automobile Engineering*, vol.222, pp.499-513, 2008

[14] JD. Halderman, "Automotive brake system" 5th Edition, Pearson Education Inc., Prentice Hall, USA, 2010

[15] MH. Rashid (Ed), "Power Electronics handbook: Devices, circuits, and applications", 3rd Edition, Elsevier Inc., 2011

[16] J. Yao, Z. Zhong, and Z. Sun, "A Fuzzy logic based regenerative braking regulation for a fuel cell bus" *In the Proc of the IEEE International Conference on Vehicular Electronics and Safety*, pp.22-25, 2006

[17] J. Paterson, and M. Ramsay, "Electric vehicle braking by fuzzy logic control, *IEEE Industry Applications Society Annual Meeting*, pp2200-2204, 1993

[18] D. Prokhorov, "Toyota Prius HEV Neurocontrol and diagnostics", *Neutral Networks*, vol.21(3), 2008, pp.458-465

[19] Y. Gao, L. Chen, and M. Ehsani, "Investigation of the effectiveness of regenerative braking for EV and HEV", SAE Technical paper 1999-01-2910, 1999

[20] SM. Lukic, J. Cao, RC. Bansal, F. Rodriguez, and A. Emadi., "Energy Storage Systems for Automotive Applications", *IEEE Transactions on Industrial Electronics*, vol.55, No.6, 2008

[21] KKT. Thanapalan, JG. Williams, GC. Premier, AJ. Guwy, "Design and implementation of renewable hydrogen fuel cell vehicles", *Renewable Energy & Power Quality Journal*, vol.1, No.9, 310, 2011

[22] H. Douglas and P. Pillay, "Sizing ultracapacitors for hybrid electric vehicles" *In the Proceedings of 31st Annual Conference of IEEE-Industrial Electronics Society*, 2005

[23] KKT. Thanapalan, F. Zhang, GC. Premier and AJ. Guwy "Optimal power management of hydrogen fuel cell vehicles", *In the Proc. of the World Renewable Energy Congress, Denver, Colorado, USA.* 2012

[24] KT. Chau and YS. Wong, "Overview of power management in hybrid electric vehicles", *Energy Conversion and Management*, vol.43, pp.1953-1968, 2002

[25] Y. Wu, and H. Gao, "Optimization of fuel cell and supercapacitor for fuel cell electric vehicles", *IEEE Transactions on vehicular Technology*, 55(6), 1748-1755, 2006.

[26] P. Rodatz, G. Paganelli, A. Sciarretta, and L. Guzzalla, "Optimal power management of an experimental fuel cell/supercapacitor powerd hybrid vehicle", *Control Engineering Practice,* 13(1), 41-53, 2005.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Recursive Filtering for a Class of Nonlinear Systems with Missing Measurements

Jun Hu
Research Institute of
Intelligent Control and Systems,
Harbin Institute of Technology,
Harbin 150001, China.

Zidong Wang
Department of Information Systems
and Computing, Brunel University,
London, UB8 3PH, UK.
Email: Zidong.Wang@brunel.ac.uk

Bo Shen
Institute for Automatic
Control and Complex Systems,
University of Duisburg-Essen,
D-47057 Duisburg, Germany.

Chenxiao Cai
School of Automation, Nanjing
University of Science and Technology,
Nanjing 210094, China.

James Lam
Department of Mechanical Engineering,
The University of Hong Kong,
Hong Kong, China.

*Abstract*—**This paper is concerned with the finite-horizon recursive filtering problem for a class of nonlinear time-varying systems with missing measurements. The missing measurements are modeled by a series of mutually independent random variables obeying Bernoulli distributions with possibly different occurrence probabilities. Attention is focused on the design of a recursive filter such that, for the missing measurements, an upper bound for the filtering error covariance is guaranteed and such an upper bound is subsequently minimized by properly designing the filter parameters at each sampling instant. The desired filter parameters are obtained by solving two Riccati-like difference equations that are of a recursive form suitable for online applications. A simulation example is exploited to demonstrate the effectiveness of the proposed filter design scheme.**

## I. Introduction

The past few decades have seen a surge of research interest on the filtering or state estimation theories due to their extensive applications in a variety of practical areas including weather forecasting, economics, radar tracker and global positioning system. Up to now, a great deal of efforts has been delivered to the design issues of various kinds of filters, for example, Kalman filters [2], [16], [22], extended Kalman filters [9], [11], [24], [25] and $H_\infty$ filters [1], [6], [10], [12], [15], [23], [28]. Among them, the traditional Kalman filter has been shown to be an optimal one in the sense of minimum variance for the linear systems, and the extended Kalman filter has been developed to serve as an effective way for handling the nonlinear estimation problems. Recently, the robust extended Kalman filtering problem has been tackled in [24] for a class of nonlinear systems, and a filtering algorithm has been presented in a recursive form suitable for online applications.

Most traditional filter design approaches rely on the assumption that the measurement signals are perfectly transmitted. Such an assumption, however, is conservative in many engineering practice presented with unreliable communication channels. For example, due to temporal sensor failures or network congestions, the system measurements may contain noise only at certain time points and the true signals are simply missing. As such, the control and filtering problems with missing measurements have received considerable research attention, see e.g. [5], [8], [13], [14], [18]–[21], [27]. A common way for modeling the data missing is to introduce a random variable satisfying the Bernoulli binary distribution taking values on either 1 or 0, where 1 is for the perfect signal delivery and 0 represents the measurement missing. Most of the aforementioned results have been based on the hypothesis that all sensors have identical failure characteristics [8]. However, such a hypothesis may not be true in the case that the signals are observed by multiple sensors and each individual sensor may have different failure rate.

It is worth mentioning that most existing results regarding the missing measurements have concentrated on *linear* systems. It is well known that the nonlinearity is a ubiquitous feature in almost all practical systems, and the occurrence of the nonlinearity inevitably degrades the system performance and even leads to instability [25], [26]. However, so far, the filtering problem for *general nonlinear* stochastic systems with missing measurements has not been thoroughly investigated yet, not to mention the case where multiple sensors undergo probabilistic missing measurements. It is, therefore, our aim of this paper to shorten the gap by initiating a study on such a challenging issue.

Motivated by the above discussions, we aim to investigate the recursive filtering problem for a class of nonlinear time-varying systems with missing measurements. A series of mutually independent random variables are introduced to describe the phenomenon of missing measurements where individual sensor is allowed to have different missing probability. The finite-horizon filter is designed such that an upper bound on the filtering error covariance is guaranteed and such an upper bound is subsequently minimized by the designed filter at each sampling instant. The proposed filter scheme is given in terms of the solutions to two Riccati-like difference equations, and therefore the algorithm is suitable for recursive computations.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

**Notations.** The notations used throughout the paper are standard. $R^n$ and $R^{n\times m}$ denote the $n$-dimensional Euclidean space and the set of all $n \times m$ matrices, respectively. For a matrix $P$, $P^T$ and $P^{-1}$ represent its transpose and inverse, respectively. $\mathrm{tr}(\cdot)$ stands for the trace of a matrix. $\circ$ is the Hadamard product defined as $[A \circ B]_{ij} = A_{ij} \cdot B_{ij}$. $E\{x\}$ stands for the expectation of the stochastic variable $x$. $I$ and $0$ represent the identity matrix and the zero matrix with appropriate dimensions, respectively. $\mathrm{diag}\{X_1, X_2, \ldots, X_n\}$ stands for a block-diagonal matrix with matrices $X_1, X_2, \ldots, X_n$ on the diagonal. Matrices, if their dimensions are not explicitly stated, are assumed to be compatible for algebraic operations.

## II. Problem Formulation and Preliminaries

Consider the following class of time-varying nonlinear systems:

$$\begin{aligned} x_{k+1} &= f(x_k) + D_k\omega_k && (1) \\ y_k &= \Xi_k C_k x_k + \nu_k && (2) \end{aligned}$$

where $x_k \in R^n$ is the system state to be estimated, the initial value $x_0$ has mean $\bar{x}_0$ and covariance $P_{0|0}$, $y_k \in R^m$ is the output vector, $\omega_k \in R^r$ is the process noise with zero-mean and covariance $Q > 0$, and $\nu_k \in R^m$ is the zero-mean measurement noise with covariance $V > 0$. The nonlinear function $f(x_k)$ is analytic everywhere with known form, $C_k$ and $D_k$ are known and bounded matrices with appropriate dimensions. $\Xi_k = \mathrm{diag}\{\xi_k^1, \xi_k^2, \ldots, \xi_k^m\}$ is to account for the missing measurements where the mutually uncorrelated (in $k$ and $i$) random variables $\xi_k^i \in R$ $(i = 1, 2, \ldots, m)$ take values of 1 and 0 with

$$\begin{aligned} \mathrm{Prob}\left\{\xi_k^i = 1\right\} &= E\left\{\xi_k^i\right\} := \vartheta_k^i, && (3) \\ \mathrm{Prob}\left\{\xi_k^i = 0\right\} &= 1 - E\left\{\xi_k^i\right\} := 1 - \vartheta_k^i. && (4) \end{aligned}$$

Here, $\vartheta_k^i \in [0, 1]$ is a known constant, $\xi_k^i$ is assumed to be independent with $\omega_k$, $\nu_k$ and $x_0$. Also, the noise signals mentioned above are uncorrelated with each other.

In this paper, we design a filter of the following form:

$$\begin{aligned} &\hat{x}_{k+1|k} = f\left(\hat{x}_{k|k}\right), && (5) \\ &\hat{x}_{k+1|k+1} = \hat{x}_{k+1|k} + K_{k+1}\left(y_{k+1} - \bar{\Xi}_{k+1}C_{k+1}\hat{x}_{k+1|k}\right) && (6) \end{aligned}$$

where $\hat{x}_{k|k}$ is the estimate of $x_k$ at time $k$ with $\hat{x}_{0|0} = \bar{x}_0$, $\hat{x}_{k+1|k}$ is the one-step prediction at time $k$, $K_{k+1}$ is the filter parameter to be determined, and $\bar{\Xi}_{k+1} := E\{\Xi_{k+1}\} := \mathrm{diag}\{\vartheta_{k+1}^1, \vartheta_{k+1}^2, \ldots, \vartheta_{k+1}^m\}$.

The objective of this paper is twofold. First, we aim to design a finite-horizon filter of form (5)-(6) such that, for the missing measurements, an upper bound for the filtering error covariance is guaranteed, i.e., there exists a sequence of positive-definite matrices $\Sigma_{k+1|k+1}$ $(0 \le k \le N)$ satisfying

$$\begin{aligned} &E\left\{\left(x_{k+1} - \hat{x}_{k+1|k+1}\right)\left(x_{k+1} - \hat{x}_{k+1|k+1}\right)^T\right\} \\ &\le \Sigma_{k+1|k+1} \end{aligned} \qquad (7)$$

Second, we shall minimize such an upper bound $\Sigma_{k+1|k+1}$ by appropriately designing the filter parameter at each sampling instant.

**Remark 1** In the model (2), $C_k x_k$ represents the measurement output subject to probabilistic information loss characterized by the matrix $\Xi_k$, and $\nu_k$ is the random exogenous noise acting on the measurement output. In other words, the model (2) is comprehensive to include the practical cases of probabilistic missing measurements and external additive disturbances, thereby reflecting the engineering practice in a more realistic way.

**Remark 2** In this paper, the phenomena of measurements missing is considered. Owing to the sensors aging and/or sensor temporal failure, the missing measurements may occur intermittently. In (2), $\Xi_k$ is introduced to characterize the missing measurements where the random variable $\xi_k^i$ $(i = 1, 2, \ldots, m)$ corresponds to the $i$ sensor operating at the $k$th sampling time point. For different sensors, it would be more reasonable to allow multiple sensors to have different missing probabilities (or failure rates [8]).

Before ending this section, we recall the following lemmas which will be frequently used in subsequent developments.

**Lemma 1** [7] Let $A = [a_{ij}]_{n\times n}$ be a real-valued matrix and $B = \mathrm{diag}\{b_1, b_2, \ldots, b_n\}$ be a diagonal stochastic matrix. Then

$$E\{BAB^T\} = \begin{bmatrix} E\{b_1^2\} & E\{b_1b_2\} & \cdots & E\{b_1b_n\} \\ E\{b_2b_1\} & E\{b_2^2\} & \cdots & E\{b_2b_n\} \\ \vdots & \vdots & \ddots & \vdots \\ E\{b_nb_1\} & E\{b_nb_2\} & \cdots & E\{b_n^2\} \end{bmatrix} \circ A$$

where $\circ$ is the Hadamard product.

**Lemma 2** [22] Given matrices $A$, $H$, $G$ and $F$ with appropriate dimensions such that $FF^T \le I$. Let $X$ be a symmetric positive definite matrix and $\gamma$ be an arbitrary positive constant such that $\gamma^{-1}I - GXG^T > 0$. Then the following inequality holds

$$\begin{aligned} &(A + HFG)X(A + HFG)^T \\ \le\ &A\left(X^{-1} - \gamma G^TG\right)^{-1}A^T + \gamma^{-1}HH^T. \end{aligned} \qquad (8)$$

**Lemma 3** [17] For $0 \le k \le N$, suppose that $X = X^T > 0$, $S_k(X) = S_k^T(X) \in R^{n\times n}$ and $H_k(X) = H_k^T(X) \in R^{n\times n}$. If

$$S_k(Y) \ge S_k(X), \ \ \forall\ X \le Y = Y^T \qquad (9)$$

and

$$H_k(Y) \ge S_k(Y), \qquad (10)$$

then the solutions $M_k$ and $N_k$ to the following difference equations

$$M_{k+1} = S_k(M_k), \ \ N_{k+1} = H_k(N_k), \ \ M_0 = N_0 > 0 \qquad (11)$$

satisfy $M_k \le N_k$.

## III. Main results

In this section, a sufficient condition for the design of filter parameters is established by solving two Riccati-like difference equations.

To proceed, denote the one-step prediction error as $\tilde{x}_{k+1|k} = x_{k+1} - \hat{x}_{k+1|k}$ and the filtering error as $\tilde{x}_{k+1|k+1} = x_{k+1} - \hat{x}_{k+1|k+1}$. Subtracting (5) from (1), we obtain

$$\tilde{x}_{k+1|k} = f\left(x_k\right) - f\left(\hat{x}_{k|k}\right) + D_k\omega_k. \tag{12}$$

By using the Taylor series expansion around $\hat{x}_{k|k}$, we linearize $f\left(x_k\right)$ as follows:

$$f\left(x_k\right) = f\left(\hat{x}_{k|k}\right) + A_k\tilde{x}_{k|k} + o(|\tilde{x}_{k|k}|) \tag{13}$$

where

$$A_k = \frac{\partial f\left(x_k\right)}{\partial x_k}|_{x_k=\hat{x}_{k|k}},$$

and $o(|\tilde{x}_{k|k}|)$ stands for the high-order terms of the Taylor series expansion. For presentation convenience, along the similar line of [3], [25], the high-order terms are transformed into the following easy-to-handle formulation:

$$o(|\tilde{x}_{k|k}|) = B_k\Omega_k L_k\tilde{x}_{k|k}, \tag{14}$$

where $B_k$ is a bounded problem-dependent scaling matrix, $L_k$ is a bounded matrix for providing an extra degree of freedom to tune the filter, and $\Omega_k$ is an unknown time-varying matrix accounting for the linearization errors of the dynamical model and satisfies

$$\Omega_k\Omega_k^T \leq I. \tag{15}$$

**Remark 3** In traditional extended Kalman filter algorithms, the Taylor series expansion is employed to linearize the nonlinearity $f(x_k)$, and the linearization errors are simply neglected which would inevitably lead to conservatism in certain cases. Recently, a new approach has been proposed in [3] to describe the higher-order terms in the Taylor series in terms of parameter uncertainties. In this paper, as in [3], [24], we use the deterministic matrix $\Omega_k$ and the scaling matrix $B_k$ in (14)-(15) to account for the linearization errors in obtaining the matrix $A_k$. For more details we refer the reader to Appendix C of [3] where a nice interpretation has been given. It is worthwhile to further mention that, in practice, the high-order terms in the Taylor series expansion are commonly bounded and it is reasonable to regard them as deterministic uncertainties affecting the system matrix $A_k$.

It follows from (12)-(14) that the one-step prediction error is given by

$$\tilde{x}_{k+1|k} = \left(A_k + B_k\Omega_k L_k\right)\tilde{x}_{k|k} + D_k\omega_k. \tag{16}$$

On the other hand, it follows from (6) that the filtering error $\tilde{x}_{k+1|k+1}$ can be described by

$$\begin{aligned}&\tilde{x}_{k+1|k+1}\\ &= \left(I - K_{k+1}\bar{\Xi}_{k+1}C_{k+1}\right)\tilde{x}_{k+1|k} - K_{k+1}\nu_{k+1}\\ &\quad - K_{k+1}\left(\Xi_{k+1} - \bar{\Xi}_{k+1}\right)C_{k+1}x_{k+1}\end{aligned} \tag{17}$$

Based on (16) and (17), we are ready to present the following lemmas which give the recursion of the one-step prediction error covariance and filtering error covariance, respectively.

**Lemma 4** The one-step prediction error covariance $P_{k+1|k}$ obeys the following recursion:

$$\begin{aligned}P_{k+1|k} =& \left(A_k + B_k\Omega_k L_k\right)P_{k|k}\left(A_k + B_k\Omega_k L_k\right)^T\\ &+ D_kQD_k^T\end{aligned} \tag{18}$$

where $P_{k|k} = E\{\tilde{x}_{k|k}\tilde{x}_{k|k}^T\}$ is the filtering error covariance.

**Proof** Since (18) follows from (16) directly, the proof is omitted for brevity.

**Lemma 5** The filtering error covariance $P_{k+1|k+1}$ is given as follows:

$$\begin{aligned}&P_{k+1|k+1}\\ &= \left(I - K_{k+1}\bar{\Xi}_{k+1}C_{k+1}\right)P_{k+1|k}\left(I - K_{k+1}\bar{\Xi}_{k+1}C_{k+1}\right)^T\\ &\quad + K_{k+1}(J_{k+1} + V)K_{k+1}^T\end{aligned} \tag{19}$$

where

$$\begin{aligned}J_{k+1} :=& \breve{\Xi}_{k+1} \circ \left(C_{k+1}E\left\{x_{k+1}x_{k+1}^T\right\}C_{k+1}^T\right),\\ \breve{\Xi}_{k+1} :=& \mathrm{diag}\{\vartheta_{k+1}^1\left(1 - \vartheta_{k+1}^1\right), \vartheta_{k+1}^2\left(1 - \vartheta_{k+1}^2\right),\\ &\ldots, \vartheta_{k+1}^m\left(1 - \vartheta_{k+1}^m\right)\}.\end{aligned} \tag{20}$$

**Proof** According to (17), we have

$$\begin{aligned}&P_{k+1|k+1}\\ &= \left(I - K_{k+1}\bar{\Xi}_{k+1}C_{k+1}\right)P_{k+1|k}\left(I - K_{k+1}\bar{\Xi}_{k+1}C_{k+1}\right)^T\\ &\quad + K_{k+1}E\{\left(\Xi_{k+1} - \bar{\Xi}_{k+1}\right)C_{k+1}x_{k+1}x_{k+1}^TC_{k+1}^T\\ &\quad \times \left(\Xi_{k+1} - \bar{\Xi}_{k+1}\right)\}K_{k+1}^T + K_{k+1}VK_{k+1}^T.\end{aligned} \tag{21}$$

Next, applying Lemma 1 and together with the property of conditional expectation, we obtain

$$\begin{aligned}&E\{\left(\Xi_{k+1} - \bar{\Xi}_{k+1}\right)C_{k+1}x_{k+1}x_{k+1}^T\\ &\times C_{k+1}^T\left(\Xi_{k+1} - \bar{\Xi}_{k+1}\right)\}\\ =&\breve{\Xi}_{k+1} \circ \left(C_{k+1}E\left\{x_{k+1}x_{k+1}^T\right\}C_{k+1}^T\right)\end{aligned} \tag{22}$$

where $\breve{\Xi}_{k+1}$ is defined in (20). Therefore, (19) follows directly from (21) and (22), and the proof of this Lemma is complete.

**Remark 4** It can be seen that the linearization has been enforced to facilitate the recursive filtering algorithm developments. From Lemmas 4-5, the filtering error covariance can be obtained for the missing measurements provided that the matrix equations (18) and (19) are solvable. Unfortunately, due to the consideration of the nonlinearity, (18) and (19) are contaminated by some uncertain terms $\Omega_k$ and $E\left\{x_{k+1}x_{k+1}^T\right\}$, which lead to essential difficulty in determining the accurate value of the filtering error covariance $P_{k+1|k+1}$. In the following, an alternatively way is employed to design an appropriate filter parameter $K_{k+1}$ such that there exists an upper bound for the filtering error covariance.

Now, we are in a position to present our main results. In view of Lemmas 2-5, the filter parameter is designed such that an optimized upper bound for the filtering error covariance is achieved at each sampling instant.

**Theorem 1** Consider the one-step prediction error covariance and the filtering error covariance in (18)-(19), respectively. Assume that (15) holds. Let $\gamma_k$ and $\varepsilon$ be positive scalars.

If the following two Riccati-like difference equations

$$\begin{aligned}&\Sigma_{k+1|k}\\ &=A_k\left(\Sigma_{k|k}^{-1}-\gamma_k L_k^T L_k\right)^{-1}A_k^T+\gamma_k^{-1}B_kB_k^T\\ &\quad+D_kQD_k^T,\end{aligned} \tag{23}$$

and

$$\begin{aligned}&\Sigma_{k+1|k+1}\\ &=\left(I-K_{k+1}\bar{\Xi}_{k+1}C_{k+1}\right)\Sigma_{k+1|k}\\ &\quad\times\left(I-K_{k+1}\bar{\Xi}_{k+1}C_{k+1}\right)^T+K_{k+1}\\ &\quad\times\left[\check{\Xi}_{k+1}\circ\left(C_{k+1}\Phi_{k+1|k}C_{k+1}^T\right)+V\right]K_{k+1}^T\end{aligned} \tag{24}$$

with initial condition $\Sigma_{0|0}=P_{0|0}>0$ have positive-definite solutions $\Sigma_{k+1|k}$ and $\Sigma_{k+1|k+1}$ such that, for all $0\le k\le N$, the following constraint

$$\gamma_k^{-1}I-L_k\Sigma_{k|k}L_k^T>0, \tag{25}$$

are satisfied where

$$\Phi_{k+1|k}:=(1+\varepsilon)\,\Sigma_{k+1|k}+\left(1+\varepsilon^{-1}\right)\hat{x}_{k+1|k}\hat{x}_{k+1|k}^T, \tag{26}$$

then with the filter parameter $K_{k+1}$ given by

$$\begin{aligned}&K_{k+1}\\ &=\Sigma_{k+1|k}C_{k+1}^T\bar{\Xi}_{k+1}\Big[\bar{\Xi}_{k+1}C_{k+1}\Sigma_{k+1|k}C_{k+1}^T\bar{\Xi}_{k+1}\\ &\quad+\check{\Xi}_{k+1}\circ\left(C_{k+1}\Phi_{k+1|k}C_{k+1}^T\right)+V\Big]^{-1}\end{aligned} \tag{27}$$

the matrix $\Sigma_{k+1|k+1}$ is an upper bound for $P_{k+1|k+1}$, i.e.,

$$P_{k+1|k+1}\le\Sigma_{k+1|k+1}. \tag{28}$$

Moreover, the filter parameter $K_{k+1}$ given by (27) minimizes the upper bound $\Sigma_{k+1|k+1}$.

**Proof** Note that the covariance matrices $P_{k+1|k}$ and $P_{k+1|k+1}$ can be rewritten as the functions of $P_{k|k}$ and $P_{k+1|k}$, respectively. Then, it is not difficult to verify that the condition (9) in Lemma 3 is satisfied.

Now, we are ready to deal with the terms of the right-hand side of (19). Considering the following elementary inequality

$$\left(\varepsilon^{\frac{1}{2}}\tilde{x}_{k+1|k}-\varepsilon^{-\frac{1}{2}}\hat{x}_{k+1|k}\right)\left(\varepsilon^{\frac{1}{2}}\tilde{x}_{k+1|k}-\varepsilon^{-\frac{1}{2}}\hat{x}_{k+1|k}\right)^T\ge0,$$

we can obtain the following inequality

$$\begin{aligned}&\tilde{x}_{k+1|k}\hat{x}_{k+1|k}^T+\hat{x}_{k+1|k}\tilde{x}_{k+1|k}^T\\ &\le\varepsilon\tilde{x}_{k+1|k}\tilde{x}_{k+1|k}^T+\varepsilon^{-1}\hat{x}_{k+1|k}\hat{x}_{k+1|k}^T\end{aligned}$$

with $\varepsilon>0$ being a scalar, which yields

$$\begin{aligned}&E\left\{x_{k+1}x_{k+1}^T\right\}\\ &\le E\left\{(1+\varepsilon)\,\tilde{x}_{k+1|k}\tilde{x}_{k+1|k}^T+\left(1+\varepsilon^{-1}\right)\hat{x}_{k+1|k}\hat{x}_{k+1|k}^T\right\}\\ &=(1+\varepsilon)\,P_{k+1|k}+\left(1+\varepsilon^{-1}\right)\hat{x}_{k+1|k}\hat{x}_{k+1|k}^T.\end{aligned} \tag{29}$$

Then, the last term of the right-hand side of (19) can be determined as

$$\begin{aligned}&K_{k+1}(J_{k+1}+V)K_{k+1}^T\\ &\le K_{k+1}\left[\check{\Xi}_{k+1}\circ\left(C_{k+1}M_{k+1|k}C_{k+1}^T\right)+V\right]K_{k+1}^T\end{aligned} \tag{30}$$

where where

$$M_{k+1|k}:=(1+\varepsilon)\,P_{k+1|k}+\left(1+\varepsilon^{-1}\right)\hat{x}_{k+1|k}\hat{x}_{k+1|k}^T.$$

It then follows from (19) and (30) that

$$\begin{aligned}&P_{k+1|k+1}\\ &\le\left(I-K_{k+1}\bar{\Xi}_{k+1}C_{k+1}\right)P_{k+1|k}\\ &\quad\times\left(I-K_{k+1}\bar{\Xi}_{k+1}C_{k+1}\right)^T+K_{k+1}\\ &\quad\times\left[\check{\Xi}_{k+1}\circ\left(C_{k+1}M_{k+1|k}C_{k+1}^T\right)+V\right]K_{k+1}^T\end{aligned} \tag{31}$$

Combining (23), (24) and (31), we can show that the condition (10) in Lemma 3 is satisfied. Therefore, it follows directly from Lemmas 2-3 that

$$P_{k+1|k+1}\le\Sigma_{k+1|k+1}.$$

Having determined the upper bound $\Sigma_{k+1|k+1}$, we are now ready to show that the filter parameter given by (27) is optimal in the sense that it minimizes the upper bound $\Sigma_{k+1|k+1}$. Taking the partial derivative of (24) with respect to $K_{k+1}$ and letting the derivative be zero, we have

$$\begin{aligned}&\frac{\partial\mathrm{tr}\left(\Sigma_{k+1|k+1}\right)}{\partial K_{k+1}}\\ &=-2\left(I-K_{k+1}\bar{\Xi}_{k+1}C_{k+1}\right)\Sigma_{k+1|k}C_{k+1}^T\bar{\Xi}_{k+1}\\ &\quad+2K_{k+1}\left[\check{\Xi}_{k+1}\circ\left(C_{k+1}\Phi_{k+1|k}C_{k+1}^T\right)+V\right]\\ &=0.\end{aligned} \tag{32}$$

From (32), and through straightforward algebraic manipulations, the optimal filter parameter $K_{k+1}$ can be determined as follows:

$$\begin{aligned}&K_{k+1}\\ &=\Sigma_{k+1|k}C_{k+1}^T\bar{\Xi}_{k+1}\Big[\bar{\Xi}_{k+1}C_{k+1}\Sigma_{k+1|k}C_{k+1}^T\bar{\Xi}_{k+1}\\ &\quad+\check{\Xi}_{k+1}\circ\left(C_{k+1}\Phi_{k+1|k}C_{k+1}^T\right)+V\Big]^{-1}\end{aligned} \tag{33}$$

Obviously, the filter parameter $K_{k+1}$ in (33) is identical to (27). To this end, the optimal filter gain $K_{k+1}$ is designed in the sense of minimizing the upper bound $\Sigma_{k+1|k+1}$ for the filtering error covariance and, therefore, the proof of this theorem is complete.

**Remark 5** At each sampling instant, the filter parameter $K_{k+1}$ is designed in Theorem 1 to minimize the upper bound of filtering error covariance. The consideration of the multiple missing measurements constitutes the main difference between our work and the work of [24]. In our main results, the constants $\vartheta_k^i$ $(i=1,2,\ldots,m)$ are there for the missing measurements where all sensors are allowed to have different missing probabilities. Furthermore, the proposed filter is derived in terms of the solutions to two Riccati-like difference equations, which is recursive and therefore suitable for online applications.

## IV. An illustrative example

Consider the following nonlinear system with missing measurements:

$$\begin{cases} x_{k+1} = f(x_k) + D_k\omega_k \\ y_k = \Xi_k C_k x_k + \nu_k \end{cases}$$

where

$$\begin{aligned} f(x_k) &= \begin{bmatrix} 0.8x_{1,k} + x_{1,k}x_{2,k} \\ 1.5x_{2,k} - x_{1,k}x_{2,k} \end{bmatrix}, \\ D_k &= \begin{bmatrix} 0.06 \\ 0.03 + 0.5e^{-5k} \end{bmatrix}, \\ C_k &= \begin{bmatrix} 0.85 & 0 \\ 0 & -1.5 \end{bmatrix} \end{aligned}$$

and $x_k = \begin{bmatrix} x_{1,k} & x_{2,k} \end{bmatrix}^T$ is the state vector with $x_{i,k}$ $(i = 1, 2)$ being the $i$-th element of the system state, $\omega_k \in R$ and $\nu_k \in R^2$ are zero-mean Gaussian white noises with covariances $0.5$ and $0.02I_2$, respectively.

In the simulation, set the initial value of estimation as $\hat{x}_{0|0} = \bar{x}_0 = \begin{bmatrix} 0.8 & 0.2 \end{bmatrix}^T$ and $\Sigma_{0|0} = 10I$. Assume that $\bar{\Xi}_k = \text{diag}\{0.95, 0.90\}$. The other parameters are chosen as $B_k = \text{diag}\{0.1, 0.2\}$, $L_k = 0.1I_2$, $\gamma_k = 0.005$, and $\varepsilon = 0.35$. By solving (23) and (24), the filter parameter can be obtained recursively and the simulation results are shown in Figs. 1-4. Here, MSE-$i$ $(i = 1, 2)$ denotes the mean square error (MSE) for the estimation of the state.

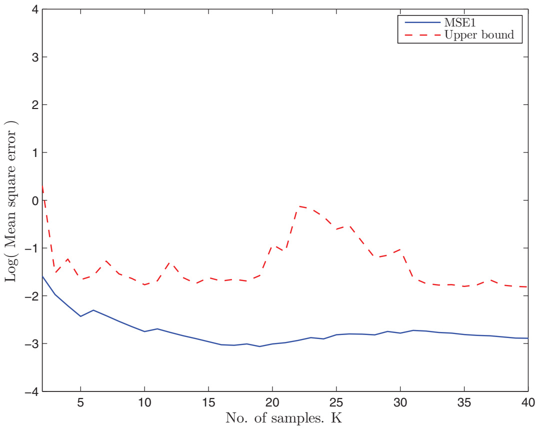


Fig. 1. MSE1 and its upper bound

In the figures, Figs. 1-2 show the upper bounds $\Sigma^{11}_{k|k}$ and $\Sigma^{22}_{k|k}$ as well as the MSE for the states $x_{1,k}$ and $x_{2,k}$, which confirm that the MSE stay below their upper bounds. The trajectories of the actual states $x_{i,k}$ and their estimates $\hat{x}_{i,k}$ $(i = 1, 2)$ are plotted in Figs. 3-4, which illustrate that the

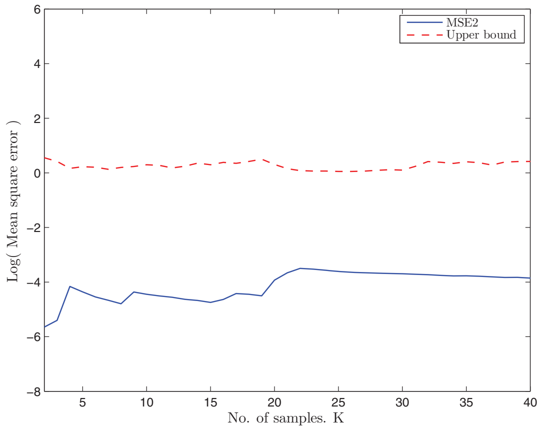


Fig. 2. MSE2 and its upper bound

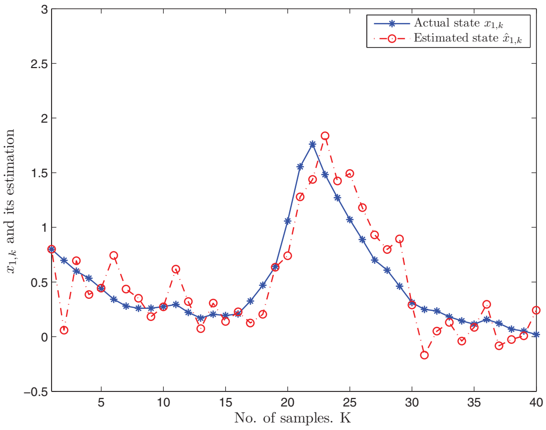


Fig. 3. The actual state $x_{1,k}$ and its estimation $\hat{x}_{1,k}$

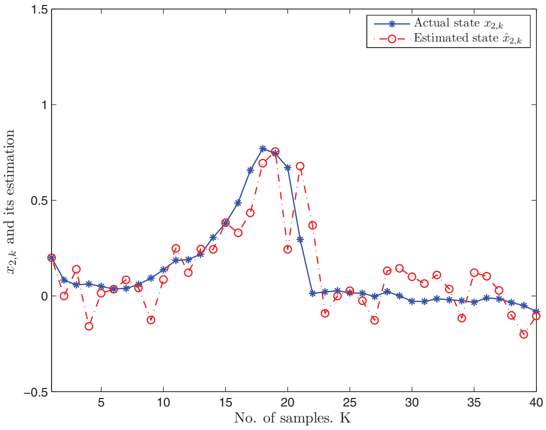


Fig. 4. The actual state $x_{2,k}$ and its estimation $\hat{x}_{2,k}$

presented scheme can perform well to estimate the system states.

## V. Conclusions

In this paper, the finite-horizon filter design problem has been investigated for a class of time-varying nonlinear systems with missing measurements. A series of mutually independent random variables that obeys Bernoulli distribution has been introduced to describe the missing measurement phenomenon. A filter has been designed to guarantee an optimized upper bound on the filtering error covariance by means of solving two Riccati-like difference equations. Finally, a numerical example has been provided to illustrate the effectiveness of the main results. Further research topics include the extension of the main results to the recursive filtering problem for general nonlinear stochastic systems, to the finite-horizon $H_\infty$ filtering problem with fading measurements, and so on.

## Acknowledgments

This work was supported in part by the National 973 Project under Grant 2009CB320600, National Natural Science Foundation of China under Grants 61028008, 61134009 and 60825303, the State Key Laboratory of Integrated Automation for the Process Industry (Northeastern University) of China, the Engineering and Physical Sciences Research Council (EPSRC) of the U.K. under Grant GR/S27658/01, the Royal Society of the U.K., the GRF HKU 7140/11E, and the Alexander von Humboldt Foundation of Germany.

## References

[1] M. Basin, P. Shi, and D. Calderon-Alvarez, Central suboptimal $H_\infty$ filter design for linear time-varying systems with state and measurement delays, *International Journal of Systems Science*, vol. 41, no. 4, pp. 411–421, Apr. 2010.

[2] R. Caballero-Águila, A. Hermoso-Carazo, J. D. Jiménez-López, J. Linares-Peréz, and S. Nakamori, Signal estimation with multiple delayed sensors using covariance information, *Digital Signal Processing*, vol. 20, no. 2, pp. 528–540, Mar. 2010.

[3] G. Calafiore, Reliable localization using set-valued nonlinear filters, *IEEE Trans. Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 35, no. 2, pp. 189–197, Mar. 2005.

[4] H. Dong, Z. Wang, J. Lam, and H. Gao, Fuzzy-model-based robust fault detection with stochastic mixed time delays and successive packet dropouts, *IEEE Trans. Systems, Man, and Cybernetics-Part B: Cybernetics*, vol. 41, no. 2, pp. 365–376, Apr. 2012.

[5] J. Feng, Z. Wang, and M. Zeng, Recursive robust filtering with finite-step correlated process noises and missing measurements, *Circuits Syst. Signal Process.*, vol. 30, no. 6, pp. 1355–1368, Dec. 2011.

[6] H. Gao, Y. Zhao, J. Lam, and K. Chen, $H_\infty$ fuzzy filtering of nonlinear systems with intermittent measurements, *IEEE Trans. Fuzzy Syst.*, vol. 17, no. 2, pp. 291–300, Apr. 2009.

[7] R. A. Horn and C. R. Johnson, *Topic in Matrix Analysis*, New York: Cambridge University Press, 1991.

[8] F. O. Hounkpevi and E. Yaz, Robust minimum variance linear state estimators for multiple sensors with different failure rates, *Automatica*, vol. 43, no. 7, pp. 1274–1280, Jul. 2007.

[9] A. G. Kallapur, I. R. Petersen, and S. G. Anavatti, A discrete-time robust extended Kalman filter for uncertain systems with sum quadratic constraints, *IEEE Trans. Autom. Control*, vol. 54, no. 4, pp. 850–854, Apr. 2009.

[10] H. R. Karimi, Robust $H_\infty$ filter design for uncertain linear systems over network with network-induced delays and output quantization, *Modeling, Identification and Control*, vol. 30, vo. 1, pp. 27–37, 2009.

[11] S. Kluge, K. Reif, and M. Brokate, Stochastic stability of the extended Kalman filter with intermittent observations, *IEEE Trans. Autom. Control*, vol. 55, no. 2, pp. 514–518, Feb. 2010.

[12] P. Li, J. Lam, and Z. Shu, $H_\infty$ positive filtering for positive linear discrete-time systems: an augmentation approach, *IEEE Trans. Autom. Control*, vol. 55, no. 10, pp. 2337–2342, Oct. 2010.

[13] M. Moayedi, Y. K. Foo, and Y. C. Soh, Adaptive Kalman filtering in networked systems with random sensor delays, multiple packet dropouts and missing measurements, *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1577–1588, Jul. 2010.

[14] B. Piccoli, K. Zadarnowska and M. Gaeta, Stochastic algorithms for robustness of control performances, *Automatica*, vol. 45, no. 6, pp. 1407–1414, 2009.

[15] P. Shi, Filtering on sampled-data systems with parametric uncertainty, *IEEE Trans. Autom. Control*, vol. 43, no. 7, pp. 1022–1027, Jul. 1998.

[16] P. Shi, E. K. Boukas, and R. K. Agarwal, Kalman filtering for continuous-time uncertain systems with Markovian jumping parameters, *IEEE Trans. Autom. Control*, vol. 44, no. 8, pp. 1592–1597, Aug. 1999.

[17] Y. Theodor and U. Shaked, Robust discrete-time minimum-variance filtering, *IEEE Trans. Signal Process.*, vol. 44, no. 2, pp. 181–189, Feb. 1996.

[18] Z. Wang, J. Lam, L. Ma, Y. Bo, and Z. Guo, Variance-constrained dissipative observer-based control for a class of nonlinear stochastic systems with degraded measurements, *Journal of Mathematical Analysis and Applications*, vol. 377, no. 2, pp. 645–658, May 2011.

[19] Z. Wang, F. Yang, D. W. C. Ho, and X. Liu, Robust finite-horizon filtering for stochastic systems with missing measurements, *IEEE Signal Process. Letters*, vol. 12, no. 6, pp. 437–440, Jun. 2005.

[20] Z. Wang, F. Yang, D. W. C. Ho, and X. Liu, Robust $H_2/H_\infty$ filtering for stochastic time-delay systems with missing measurements, *IEEE Trans. Signal Process.*, vol. 54, no. 7, pp. 2579–2587, Jul. 2006.

[21] Y. Xia, J. Shang, J. Chen, and G. P. Liu, Networked data fusion with packet losses and variable delays, *IEEE Trans. Systems, Man, and Cybernetics-Part B: Cybernetics*, vol. 39, no. 5, pp. 1107–1120, Oct. 2009.

[22] L. Xie, Y. C. Soh, and C. E. de Souza, Robust Kalman filtering for uncertain discre-time systems, *IEEE Trans. Autom. Control*, vol. 39, no. 6, pp. 1310–1314, Jun. 1994.

[23] J. Xiong and J. Lam, Fixed-order robust $H_\infty$ filter design for Markovian jump systems with uncertain switching probabilities, *IEEE Trans. Signal Process.*, vol. 54, no. 4, pp. 1421-1430, Apr. 2006.

[24] K. Xiong, L. Liu, and Y. Liu, Robust extended Kalman filtering for nonlinear systems with multiplicative noises, *Optimal Control Applications and Methods*, vol. 32, no. 1, pp. 47–63, Jan-Feb. 2011.

[25] K. Xiong, C. Wei, and L. Liu, Robust extended Kalman filtering for nonlinear systems with stochastic uncertainties, *IEEE Trans. Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 40, no. 2, pp. 399–405, Mar. 2010.

[26] R. Yang, P. Shi, and G. Liu, Filtering for discrete-time networked nonlinear systems with mixed random delays and packet dropouts, *IEEE Trans. Autom. Control*, vol. 56, no. 11, pp. 2655–2660, Nov. 2011.

[27] X. Yao, L. Wu, and W. Zheng, Fault detection filter design for Markovian jump singular systems with intermittent measurements, *IEEE Trans. Signal Process.*, vol. 59, no. 7, pp. 3099–3109, Jul. 2011.

[28] H. Zhang, G. Feng, G. Duan, and X. Lu, $H_\infty$ filtering for multiple-time-delay measurements, *IEEE Trans. Signal Process.*, vol. 54, no. 5, pp. 1681–1688, Oct. 2006.

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# Renewable Hydrogen Hybrid Electric Vehicles and Optimal Energy Recovery Systems

Kary Thanapalan, Fan Zhang, Giuliano Premier, Alan Guwy, Jon Maddy
Sustainable Environment Research Centre (SERC)
University of Glamorgan
Pontypridd CF37 1DL, United Kingdom
E-mail:kthanapa@glam.ac.uk

*Abstract*—**This paper investigates the design of an optimal energy recovery system for a hydrogen hybrid electric vehicle to allow increased range. The proposed system includes two different energy recovery sub-systems. These are; a regenerative braking sub-system and an active suspension sub-system. A dynamic model of this optimal energy recovery sub-system is implemented in MATLAB/Simulink$^{TM}$ and integrated with a hydrogen hybrid electric vehicle model to investigate the effect of the optimal energy recovery mechanism. The simulation results indicate the energy recovery and hence, lead to the extended range capabilities.**

***Keywords – Hydrogen hybrid vehicles, Regenerative braking, Active suspension system, Energy recovery***

## I. INTRODUCTION

The increasing global demand of fuels together with  and green house gas (GHG) environmental concerns, have progressively leads to the global trend towards low-emission and fuel efficient vehicles in recent years [1].

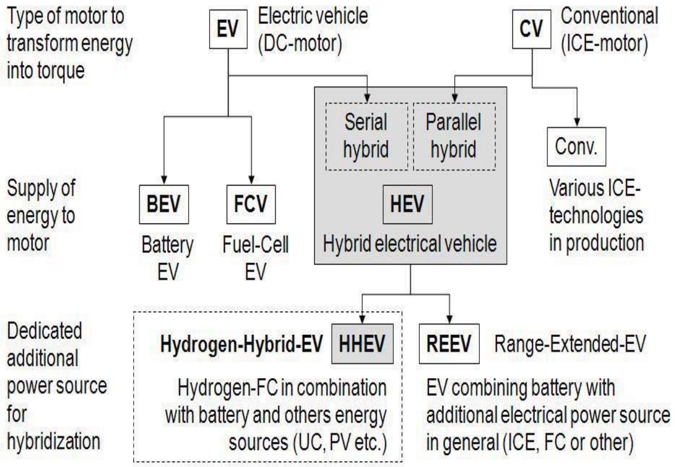

Type of motor to transform energy into torque
EV
Electric vehicle (DC-motor)
CV
Conventional (ICE-motor)
Serial hybrid
Parallel hybrid
Conv.
Supply of energy to motor
BEV
Battery EV
FCV
Fuel-Cell EV
HEV
Hybrid electrical vehicle
Various ICE-technologies in production
Dedicated additional power source for hybridization
Hydrogen-Hybrid-EV
HHEV
Hydrogen-FC in combination with battery and others energy sources (UC, PV etc.)
REEV
Range-Extended-EV
EV combining battery with additional electrical power source in general (ICE, FC or other)

Fig.1 Classification of HEV systems

Several variations and definitions have recently been under some scrutiny for use in alternative car concepts (see Fig.1). Among these vehicles, renewable hydrogen hybrid electric vehicles (HHEV) are considered to be one of the potential alternative candidates to replace conventional fossil fuelled internal combustion engine vehicles. Renewable hydrogen fuel cell vehicles are able to provide better fuel economy and comparatively reduced environmental pollution [2]. By incorporating energy recovery mechanisms, improved performance through energy recovery can be achieved [3]. Energy loss can also be minimized and fuel consumption rate be kept low. In such circumstances, an efficient economic fuel cell vehicle can be realised [3], [4]. The utilization of energy recovery systems makes the hydrogen fuel cell vehicles more attractive by better utilising the on board storage capacity for hydrogen, through raised energy efficiency. Model-based simulation has been used to assist in making an assessment of the benefits from such a system and can be used in exploring further options for improving energy efficiency.

Recent research efforts show that the energy recovery via an active suspension system in hydrogen hybrid electric vehicle may play a part in improving efficiency and extended range capabilities [5], [6]. Energy recovery suspension can achieve a damping function and energy recovery by changing the suspension input vibration produced by road roughness into electrical energy, so removing energy and effecting damping. The control structure of a plausible suspension system is shown in Fig.2. The control unit will detect the available energy supply from the suspension subsystem via the sensors circuit. The recovered energy will then be stored in the vehicle's main electrical energy storage (ESS) such as a battery, to be utilised subsequently to extend the range of the electric vehicle.

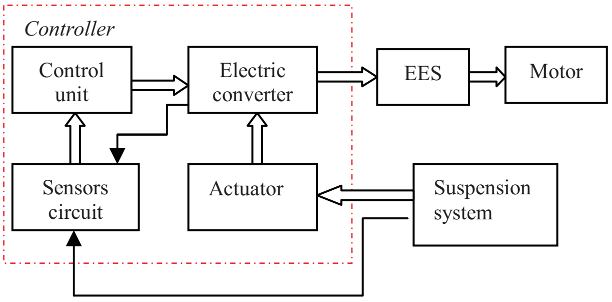


Fig.2 Control operation of a suspension system

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

Results from previous research show that, significant amount of energy can also be recovered by the integration of the regenerative braking mechanism [4]. Further research effort should focus on the development of a systematic and integrated energy recovery system to recover as much kinetic energy as possible, while maintaining suitable power flows in the other subsystems and effecting a suspension function. To this end, several researchers have developed many different configurations of energy recovery mechanisms to continuously recover energy otherwise dissipated, that results from road irregularities, vehicle acceleration, and braking; see for example [7]-[12]. This recovered energy can be used to reduce fuel consumption.

This work addresses the issues relating to the development of a new mechanism for optimal energy recovery. The main purpose of the paper is to employ a simulation model of a renewable hydrogen fuel cell vehicle together with an optimal energy recovery system (OERS) to carry out simulation analysis to investigate the effect of the OERS.

## II. Energy Recovery Systems for HEV

The energy efficiency and driving range of HHEV system can be increased through the development and incorporation of OERS. Many different types of energy recovery system have been developed over the last two decades. These range from electro-hydraulic braking systems (EHBS) to regenerative shock absorbers [13]-[18].

EHBS can be configured in either series or parallel regenerative mode. In a series configuration, the hydraulic control unit manages the brake cylinder pressures, as well as the front-rear axle brake balance. It requires active brake management to achieve total braking to all four wheels. However, it should be noted that in most modern hybrid electric vehicles the regenerative braking system is applied only on the drive wheel, consequently the system has to use thermal/friction braking systems at the same time, to equilibrate the braking. Parallel configuration is less complex because the thermal brakes are used along with energy recovery by the reversible motor/generator. Front and rear brake balance is retained because the thermal brakes are in use during the entire braking event. The amount of energy captured by a parallel regenerative braking system is thus less than from a series system [19].

Hydraulic power assist is another way of managing the energy usage. In the hydraulic power assist, when the driver steps on the brake, the vehicle's kinetic energy is used to power a reversible pump, which sends hydraulic fluid from a low pressure accumulator into a high pressure accumulator. The pressure is created by nitrogen gas in the accumulator, which is compressed as the fluid is pumped into the space the bagged or isolated gas formerly occupied. This slows the vehicle and helps bring it to stop. The fluid remains under pressure in the accumulator until the driver pushes the accelerator again, at which point the pump is reversed and the pressurized fluid is used to accelerate the vehicle.

Regenerative shock absorber mechanisms are a recent development for energy recovery and are essentially regenerative suspension systems technologies. In such regenerative suspension systems, the function is, as in a conventional automotive shock absorber, mechanisms to dampen suspension movements to produce a controlled action that keeps the tire firmly on the road and filters road irregularities for comfort and vehicle integrity. This is normally done by converting the kinetic energy into heat energy, which is then dissipated through the hydraulic oil and cylinder by heat transfer to atmosphere. The power generating shock absorber converts this kinetic energy into electricity through the use of a linear electric generator or other device. The electricity generated by each power generating shock absorber stroke (see Fig.3) can then be combined with electricity from other power generation systems such as regenerative braking and stored in the vehicle's energy storage systems [20].

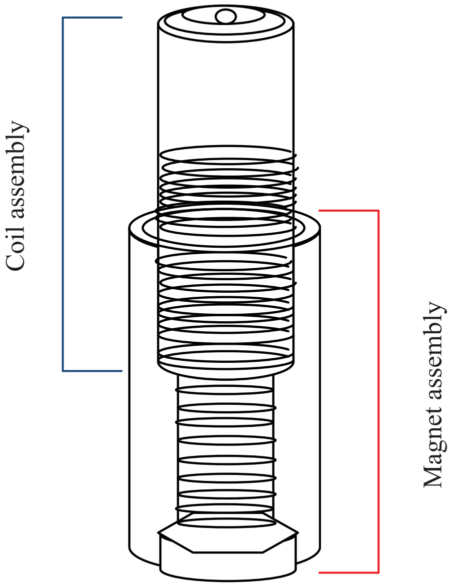


Fig.3 Shock absorbers

For the proposed OERS system, two different energy recovery mechanisms have been considered. These are; the suspension energy recovery mechanism and regenerative braking subsystem. The combined energy recovery sources (suspension energy recovery mechanism and regenerative braking) in the HHEV system may provide a more viable solution if increased capital costs can be rendered economic. The suspension energy recovery mechanism will recover some energy from suspension system and through regeneration from braking, the vehicle is allowed to recapture and store a part of the kinetic energy as opposed to wasting it as heat. Regenerative braking and suspension both essentially convert kinetic energy back into usable energy. Hence, they will be most beneficial in circumstances where most energy is otherwise lost.

## III. OERS Model

In order to develop an optimal energy recovery system model, it is hypothesised that utilization of a regenerative braking system and an active energy recovery suspension processes in a single system, make the proposed optimal energy recovery system more attractive from an energy system point of view.

*Regenerative braking system:* In recent literature, hybrid electric vehicle systems, regenerative braking is applied only on the drive wheels, therefore a standard front wheel-drive configuration [21], [22] is adopted in this paper. The regenerative braking operation can be described by the following equations.

The braking force required $F_{req}$ is given by;

$$F_{req} = F_{acc} - F_{roll} - F_{drag} - F_{hill} \qquad (1)$$

Where $F_{acc}$ is the force required to give the rate of change of velocity (i.e., acceleration/deceleration), $F_{roll}$ the rolling resistance force, $F_{drag}$ is the aerodynamic drag, $F_{hill}$ is the hill climbing force.

The power required to slow the vehicle $P_{req}$ is calculated by;

$$P_{req} = F_{req} v \qquad (2)$$

Where $v$ is the vehicle velocity.

$P_{req}$ is the total power required for the front and rear axles. Thus,

$$P_{req} = P_f + P_r \qquad (3)$$

Where $P_f$ and $P_r$ are the power provided by the front and rear wheels respectively. It should be noted that $P_f > P_m$ due to the efficiency of the motor sub-system, where $P_m$ is the motor electric power and is given by;

$$P_m = \eta_m . P_f \qquad (4)$$

Where $\eta_m$ is the efficiency of the motor and is calculated by;

$$\eta_m = \tau_m \omega_m / (\tau_m \omega_m + k_1 \tau_m^2 + k_2 \omega_m + k_3 \omega_m^3 + k_s) \qquad (5)$$

Where $\tau_m$ is the motor braking torque, $\omega_m$ is the motor angular speed, $k_1, k_2$ and $k_3$ are the material losses coefficients. $k_s$ is the constant losses that apply at any velocity

The power that charges the battery $P_{ch}$ is given by;

$$P_{ch} = P_m - P_{av} \qquad (6)$$

When $P_{ch} < 0$ it becomes the discharge power, where $P_{av}$ is the average power of the accessories, i.e., the electric power needed to run the other electrical systems. While the HHEV is in motion; when braking is applied, a certain amount of power is dissipated into the battery. In this paper an internal resistance battery model [5] is adopted, which characterizes the battery with a voltage source and an internal resistance.

Let the current $I$ be flowing into the battery, so the charge power of the battery can be obtained as;

$$P_{ch} = EI + I^2 R \qquad (7)$$

Where $E$ is the open circuit voltage, and it changes with the state of charge ( $S$ ) of the battery, $R$ is the internal resistance of the battery. Now, the charge current $I$ is obtained by solving equation (7) and is given by;

$$I = (-E + \sqrt{E^2 + 4RP_{ch}}) / 2R \qquad (8)$$

and $(S)$ is represented as follows;

$$(S) = (S_0) + \Delta t \times I . C_P^{-1} \qquad (9)$$

Where $(S_0)$ is the battery initial state of charge, $\Delta t$ is the sampling time and $C_P$ is the Peukert capacity.

The regenerated power of the battery is calculated as;

$$P_{regen} = P_{ch} - I^2 R \qquad (10)$$

With the regenerated energy efficiency $\eta_{regen}$ defined as;

$$\eta_{regen} = \frac{\int P_{regen}}{\int P_{req}} \qquad (11)$$

From the above regenerative braking system (RGBS) model it is clear that the amount of regenerated brake energy depends on multiple factors in HHEV.

*Suspension energy recovery system*: Energy recovery suspension processes can achieve a significant amount of energy recovery by converting suspension vibration damping required for road roughness, into electrical energy. Design studies of suspension energy recovery systems for hybrid electric vehicles have been conducted recently by several researchers, see for example, [3], [6]. Hybrid electric vehicle configuration with active suspension systems would allow the recovery of energy and hence, lead to the extended range capabilities. The purpose of a suspension system is to support the vehicle body, damp vibration and increase ride comfort. Currently, there are three different types of suspension systems used in automobile industry: passive, semi-active and active [6]. The traditional passive suspensions use springs and dampers to absorb the oscillation while in active cases the suspension is controlled by an external controller. Semi-active suspensions include devices such as springs and shock absorbers together with other systems such as hydro-pneumatic and electromagnetic suspension. On the one hand, the use of semi-active/active suspension achieves a better isolation performance for various vibration modes and improves the riding comfort, but on the other hand, they increase the cost, weight and energy consumption of the car. The suspension energy recovery mechanism developed here is expected to recovery some energy from the suspension system to improve the energy efficiency. The dynamic model with seven degree of freedom (7-DOF); four wheel suspension systems can be expressed using differential equations [5]. Fig. 4 shows a simplified version of a suspension system for a complex full car model.

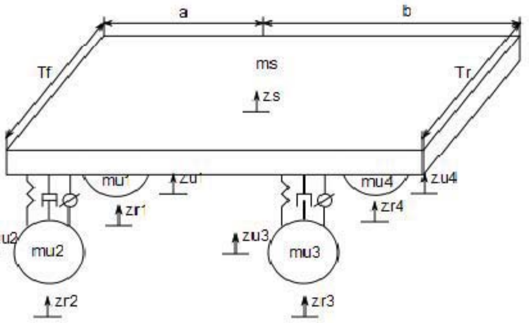


Fig.4. Simplified version of a suspension system for a complex full car model

There are roll, pitch and vertical displacement of sprung mass and four unsprung mass are included. The dynamic model is divided into seven main parts:

Bouncing of the sprung mass

$$\begin{aligned} m_s\ddot{Z}_s = & -b_f\left(\dot{Z}_{s_1}-\dot{Z}_{u_1}\right)-bf\left(\dot{Z}_{s_2}-\dot{Z}_{u_2}\right)-br\left(\dot{Z}_{s_3}-\dot{Z}_{u_3}\right) \\ & -br\left(\dot{Z}_{s_4}-\dot{Z}_{u_4}\right)-kf\left(Z_{s_1}-Z_{u_1}\right)-kf\left(Z_{s_2}-Z_{u_2}\right) \\ & -kr\left(Z_{s_3}-Z_{u_3}\right)-kr\left(Z_{s_4}-Z_{u_4}\right)+u_1+u_2+u_3+u_4 \end{aligned} \tag{12}$$

Pitching of the sprung mass

$$\begin{aligned} I_p\ddot{\theta}_s = & -b_f a\left(\dot{Z}_{s_1}-\dot{Z}_{u_1}\right)-b_f a\left(\dot{Z}_{s_2}-\dot{Z}_{u_2}\right)-b_r b\left(\dot{Z}_{s_3}-\dot{Z}_{u_3}\right) \\ & +b_r b\left(\dot{Z}_{s_4}-\dot{Z}_{u_4}\right)-k_f a\left(Z_{s_1}-Z_{u_1}\right)+k_f a\left(Z_{s_2}-Z_{u_2}\right) \\ & +k_r b\left(Z_{s_3}-Z_{u_3}\right)+k_r b\left(Z_{s_4}-Z_{u_4}\right)+au_1+au_2-bu_3-bu_4 \end{aligned} \tag{13}$$

Rolling of the sprung mass

$$\begin{aligned} I_r\ddot{\varphi}_s = & -b_f T_f\left(\dot{Z}_{s_1}-\dot{Z}_{u_1}\right)-b_f T_f\left(\dot{Z}_{s_2}-\dot{Z}_{u_2}\right)-b_r T_r\left(\dot{Z}_{s_3}-\dot{Z}_{u_3}\right) \\ & +b_r T_r\left(\dot{Z}_{s_4}-\dot{Z}_{u_4}\right)-k_f T_f\left(Z_{s_1}-Z_{u_1}\right)+k_f T_f\left(Z_{s_2}-Z_{u_2}\right) \\ & -k_r T_r\left(Z_{s_3}-Z_{u_3}\right)+k_r T_r\left(Z_{s_4}-Z_{u_4}\right)+T_f u_1-T_f u_2+T_r u_3-T_r u_4 \end{aligned} \tag{14}$$

Vertical direction for each wheel

$$m_{uf}\ddot{Z}_{u_i}=b_f\left(\dot{Z}_{s_i}-\dot{Z}_{u_i}\right)+k_f\left(Z_{s_i}-Z_{u_i}\right)-k_{tf}Z_{u_i}-u_i+k_{tf}Z_{r_i} \tag{15}$$

with $i=1,2,3,4$

where;

$$Z_{s_1}=T_f\varphi_s+a\theta_s+Z_s\,,\ \dot{Z}_{s_1}=T_f\dot{\varphi}_s+a\dot{\theta}_s+\dot{Z}_s$$

$$Z_{s_2}=-T_f\varphi_s+a\theta_s+Z_s\,,\ \dot{Z}_{s_2}=-T_f\dot{\varphi}_s+a\dot{\theta}_s+\dot{Z}_s$$

$$Z_{s_3}=-T_r\varphi_s-b\theta_s+Z_s\,,\dot{Z}_{s_3}=-T_r\dot{\varphi}_s-b\dot{\theta}_s+Z_s$$

$$Z_{s_4}=-T_r\varphi_s-b\theta_s+Z_s\,,\ \dot{Z}_{s_4}=-T_r\dot{\varphi}_s-b\dot{\theta}_s+\dot{Z}_s$$

where $Z_s$ is the vertical displacement, $Z_{u_i}$ $(i=1,2,3,4)$ is the vertical displacement of each wheel, $\theta_s$ pitch angle and $\varphi_s$ roll angle.

So far, a regenerative braking system model and a suspension system for a complex full car model have been established. These models are then used for the design of a combined OERS system to improve HHEV performance.

The combined OERS system is implemented in MATLAB/Simulink$^{TM}$ and incorporated into a vehicle simulator [5] to analysis the effect of the integration of the OERS system. In the following section of the paper discusses results from the OERS system developments.

## IV. Results and Discussion

The simulation results show that by using this combined optimal energy recovery mechanism, a considerable amount of energy can be recovered. For illustrative purposes, example simulation results are shown here. For this simulation, the New European drive cycle (NEDC) (see Fig.5) is used here and the suspension system parameters are listed in Table 4.1.

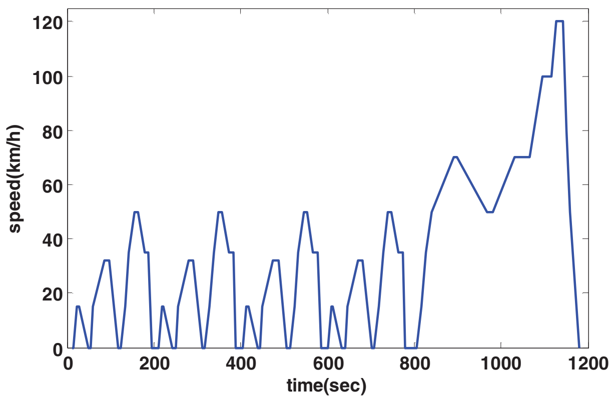


Fig.5 New European  Drive cycle

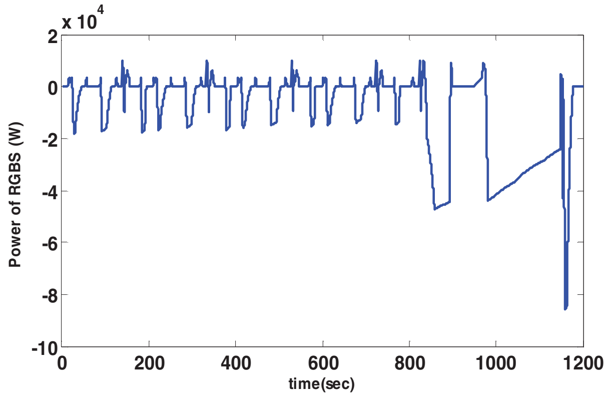


Fig.6 Power of RGBS energy recovery

Table 4.1: Suspension system parameters

| Symbol | Quantity | Unit |
|---|---|---|
| Suspension system parameters | | |
| $m_s$ | 464 | Kg |
| $m_u$ | 35 | Kg |
| $K_s$ | 16812 | N/m |
| $K_t$ | 190000 | N/m |
| $b_s$ | 1000 | N/m |
| $b_t$ | 1000 | N/m |

where $m_s$ is the mass of a vehicle, $m_u$ is the mass of a wheel components, $K_s$ is the spring stiffness, $K_t$ is the tire stiffness, $b_s$ is the damper coefficient, and $b_t$ is the tire damping coefficient.

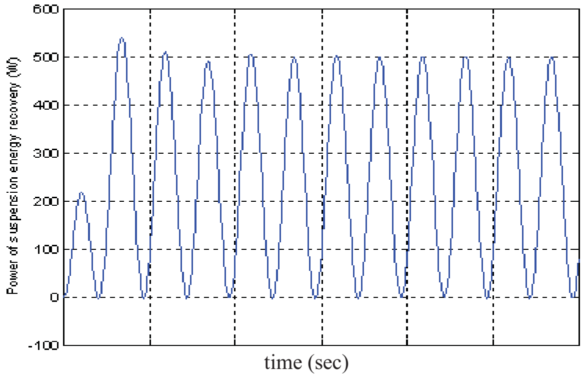


Fig.7 Power of suspension energy recovery

Suppose, the energy recovered by the use of the suspension energy recovery mechanism is used to produce hydrogen, then the average hydrogen production rate is about $4\times10^{-3}kgh^{-1}$ (see Fig.8).

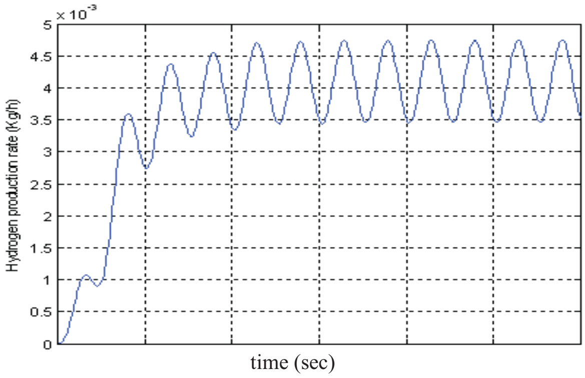


Fig.8 Hydrogen production rate from the suspension energy recovery mechanism

Fig. 6 and 7 shows the power of RGBS energy recovery and power of suspension energy recovery respectively. From the results it is clearly shown that the RGBS recover more energy than the suspension energy recovery system. However, it should be noted that the total energy recovered by the use of OERS is higher than using only the suspension recovery mechanism or RGBS (see Fig.9 and 10). In Fig. 9 the total energy recovery by the OERS is shown. Figure 10 shows the enlarged view of compared responses of Fig.6 and 9 for the first 30 sec in order to clearly show the power of OERS (solid line) and RGBS (dashed line).

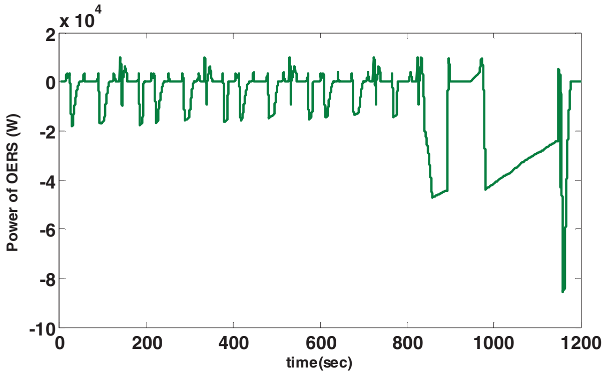


Fig.9 Power of OERS system

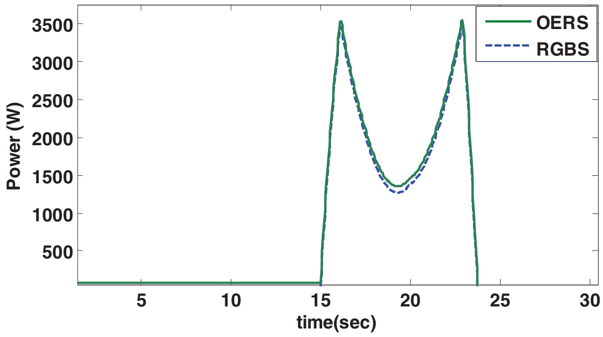


Fig.10 Comparison of power of OERS and RGBS

## V. CONCLUDING REMARKS

In this paper, an investigation of the development of an optimal energy recovery system for a renewable hydrogen fuel cell vehicle is carried out. A dynamic model of a regenerative braking system and an active suspension energy recovery system has been developed and combined as a single energy recovery mechanism. This optimal energy recovery subsystem is implemented in MATLAB/Simulink$^{TM}$. The model is then incorporated into a vehicle simulator to analysis the effect of the integration of the energy recovery mechanism. The results show that by using this mechanism, significant amount of energy can be recovered. It is expected that the modelling will provide a suitable platform for investigating the effects of system architecture and vehicle drive cycles analysis.

## ACKNOWLEDGMENT

This work was supported by the CymruH2Wales project, part of the Low Carbon Research Institute Convergence Programme. The CymruH2Wales project has been supported by the European Regional Development Fund through the Wales Government.

## REFERENCES

[1] KKT. Thanapalan and GP. Liu, “Modelling and control of fuel cell hybrid electric vehicle systems”, *In the Proc of the UKACC Int. Conference on Control, Coventry, UK,* pp.1100-1105, 2010.

[2] A. Emadi., ad SS.Williamson “Fuel cell vehicles: opportunities and challenges,” in *the Proc. IEEE Power Engineering Society General Meeting, Denver, Colorado, USA*, vol.2, pp. 1640–1645, 2004.

[3] KKT. Thanapalan, F. Zhang, AD. Procter, SJW. Carr, GC. Premier, AJ Guwy, and J. Maddy “Development of energy saving mechanism for renewable hydrogen vehicles”, *Renewable Energy and Power Quality Journal*, vol.1, No.10, 353, 2012

[4] F. Wang, and B. Zhuo “Regenerative braking strategy for hybrid electric vehicles based on regenerative torque optimization control”, *Journal of Automobile Engineering*, vol.222, pp.499-513, 2008

[5] KKT. Thanapalan, F. Zhang, GC. Premier, J Maddy and AJ. Guwy “On-board renewable hydrogen production system for hydrogen hybrid vehicles”, *In the Proc. of the World Renewable Energy Congress, Denver, Colorado, USA.* 2012

[6] R. Rajamani, and KJ. Hedrick, “Adaptive observers for active automotive suspensions theory and experiment”, IEEE Transactions on control systems technology, vol.3, No.1, 1995

[7] H. Yeo, S. Hwang, and H. Kim, “Regenerative braking algorithm for a hybrid electric vehicle with CVT ratio control”, *Proceedings of the Institution of Mechanical Engineering, Part D: Journal of Automobile Engineering,* vol.220, pp.1589-1600, 2006

[8] J. Zhang, X. Lu, J. Xue, and B. Li, “Regenerative braking system for series hybrid electric city bus”, *The World Electric Vehicle Journal*, vol.2, No.4, pp.128-134, 2008

[9] J. Yao, Z. Zhong, and Z. Sun, “A Fuzzy logic based regenerative braking regulation for a fuel cell bus” *In the Proc of the IEEE International Conference on Vehicular Electronics and Safety*, pp.22-25, 2006

[10] SR. Cikanek and KE. Bailey, “Regenerative braking system for a hybrid electric vehicle”, *In the Proc of the American Control Conference, Anchorage, AK, USA*, pp.3129-3134, 2002

[11] KT. Chau and YS. Wong, “Overview of power management in hybrid electric vehicles”, *Energy Conversion and Management*, vol.43, pp.1953-1968, 2002

[12] J. Paterson, and M. Ramsay, “Electric vehicle braking by fuzzy logic control, *IEEE Industry Applications Society Annual Meeting*, pp2200-2204, 1993

[13] M. Assaf, D. Seshsachalam, D. Chandra, and RK Tripathi, “DC-DC Converters via MATLAB/Simulink”, *In the Proc of WSEAS Conference on Automatic Control, Modelling and Simulation (ACMOS’05), Prague, Czech Republic*, pp.464-471, 2005

[14] JD. Halderman, “Automotive brake system” 5th Edition, Pearson Education Inc., Prentice Hall, USA, 2010

[15] D. Prokhorov, “Toyota Prius HEV Neurocontrol and diagnostics”, *Neutral Networks*, vol.21(3), pp.458-465, 2008

[16] MH. Rashid (Ed), “Power Electronics handbook: Devices, circuits, and applications”, 3rd Edition, Elsevier Inc., 2011

[17] B. Cao, Z. Bai, and W. Zhang, “Research on control for regenerative braking of electric vehicles”, *In the Proc of the IEEE International Conference on Vehicular Electronics and Safety*, pp.92-97, 2005

[18] Y. Gao, L. Chen, and M. Ehsani, “Investigation of the effectiveness of regenerative braking for EV and HEV”, SAE Technical paper 1999-01-2910, 1999

[19] Electro-Hydraulic Brake System (EHBS), www.autoelectronics.com, 2012

[20] shock absorber, www.gizmag.com, 2012

[21] L. James, and L. John, “Electric vehicle technology explained”, England, UK: John Wiley & Sons, pp183-292, 2003

[22] J. Guo, J. Wang, and B.Cao, “Regenerative braking strategy for electric vehicles”, *In the proc of the IEEE Intelligent vehicles symposium*, pp864-868, 2009

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# Experimental Verification of Best Linear Approximation of a Wiener System for Binary Excitations

Hin Kwan Wong
School of Engineering
University of Warwick
Coventry, UK CV4 7AL
Email: Hin.Wong@warwick.ac.uk
Telephone: (44) 2476 572904

Johan Schoukens
ELEC Dept.
Vrije Universiteit Brussel
1050 Brussels, Belgium
Email: Johan.Schoukens@vub.ac.be
Telephone: (32) 2629 2944

Keith Godfrey
School of Engineering
University of Warwick
Coventry, UK CV4 7AL
Email: K.Godfrey@warwick.ac.uk
Telephone: (44) 2476 523144

***Abstract*—The Best Linear Approximation (BLA) is a linearisation of the transfer characteristics of a nonlinear system in a least squares sense. The BLA is known to depend on the statistical properties of the input signal used to identify it. The theory for Gaussian input sequences has been known for several years, but the corresponding theory for binary input sequences has only recently been developed. In this paper, experiments on a physical electronic Wiener system, aimed at verification of predictions made for the differences between BLA's estimated using Gaussian sequences and those estimated using binary sequences, are described. The results were found to be a good match with the theory but difficulties encountered during the experiment highlight a need for further work in extending the discrete-time theory to the continuous-time domain.**

## I. Introduction

In system identification, the frequency response function or the impulse response of a system contains information about the system. Based on this information, one can model, make prediction and control the system to produce desired behaviour. All systems are nonlinear to some extent, and in some systems the nonlinearity plays a significant role. Even so, linearising a nonlinear system has merits in modelling and control, and the Best Linear Approximation (BLA) [1–5], which is a linear model minimising the expected value of mean squared difference between the actual output of the system and the modelled output is particularly useful for this. The BLA however, depends on the power and amplitude distribution of the excitation signals used to identify it [6]. This paper verifies the theory developed by Wong et al. [6] for binary input signals through experiments on a physical electronic system with a linear low-pass filter and a cubic nonlinearity in a Wiener configuration.

## II. Experiment Setup

The system was set up using the equipment listed below, and following the system schematic shown in Figure 1. The HP VXI mainframe was connected to a desktop computer with the MATLAB software. Data analysis was performed through MATLAB.

### A. List of equipment

- HP E1401B VXI mainframe with:
  - VXI-MXI-2 interface card
  - 2 × HP E1430A 10 MSa/s 23-bit ADCs, with filtering and memory (henceforth referred to as 'acquisition card')
  - HP E1445A arbitrary function generator card
- Desktop computer with PCI-MXI-2 interface card
- Non-inverting pre-buffer with AD8610A op-amp
- Non-inverting post-buffer with TL071CP op-amp
- 2 × 50 Ω matched impedance buffers
- RC filter circuit with changeable resistors and capacitors
- 50 Ω matched impedance measurement buffers
- Pre-built cubic nonlinearity circuit
  - based on AD532JH four-quadrant multipliers
- Tektronix TDS 2001C Oscilloscope
- ±12 V and ±15 V power supplies
- A 1.5 nF capacitor with either 2.7 kΩ, 27 kΩ or 110 kΩ resistors in the RC filter,
  - giving cut-off frequency values of $f_{\text{co}} = 39.3\,\text{kHz}$, 3.93 kHz and 0.965 kHz respectively.

### B. Methodology

The objective is to verify theoretical difference between the BLA's obtained from the use of Maximum Length Binary Sequences (m-sequences) [7], and from Gaussian signals, more specifically random phase multisines [2]. The linearity was a simple RC filter circuit, and the nonlinearity was a static cubic power function. Three sets of experiments were performed, each with a different combination of time constants for the linearity. The resistor values used were 2.7 kΩ, 27 kΩ and 110 kΩ. The capacitor value was fixed at 1.5 nF, for a reason which will be explained in Section II-D.

Table I lists parameters and their values used in the experimental work. Both types of input were subjected to supersampling (see Section II-D). However, the bandwidth of the multisine was set equal to the clock frequency of the binary

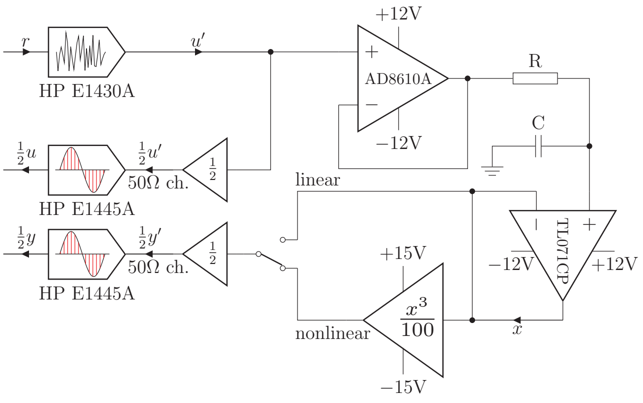


Figure 1. System schematic

sequence. This was performed so that after downsampling (also see Section II-D) at the measurement, both types of signal will have identical bandwidth and the spectral ‘whiteness’ of the two signal types could be preserved. The whiteness of the input spectrum constitutes one of the assumptions of the original discrete time BLA theory in Wong et al. [6].

The general procedure of data collection of the experiment for both the linear and nonlinear cases was as follows:

1) Generate the reference signal $r$, either:
   a) a discrete random-phase multisine for the Gaussian case, or:
   b) a Maximum Length Binary Sequence of period $N_{\text{base}} = 511$ samples for the binary case.
2) Realise the periodic signal using the HP E1445A arbitrary function generator card. The excitation is uninterrupted and continuously turned on from this point onwards.
3) Pause for 5 seconds so that transient effects in the measurements are expected to be negligible.
4) Initiate measurements with the acquisition cards and collect $P$ periods of data. The measurement intervals are internally synchronised with the generator.
5) Due to internal attenuation of the matched impedance buffers, measurements are normalised by a factor of two to obtain $u$ and $y$ (in multiple periods).
6) Go to Step 1 and repeat for a different realisation of input until $M$ data sets of different input realisations are obtained.

Note that across the $M$ sub-experiments, the same input sequence realisation was never used twice.

### C. Robust non-parametric identification procedure

Given a set of input and output data from $M$ sub-experiments of independent input realisations, each with $P$ periods of steady-state measurements, robust estimator for the BLA is given as:

$$\hat{G}_{\text{BLA}}(\mathrm{j}\omega) = \frac{\sum_{m=1}^{M}\sum_{p=1}^{P} Y_{[m,p]}U^{\star}_{[m,p]}}{\sum_{m=1}^{M}\sum_{p=1}^{P} U_{[m,p]}U^{\star}_{[m,p]}} \tag{1}$$

where $U_{[m,p]}$ and $Y_{[m,p]}$ contain the $m^{\text{th}}$ sub-experiment and $p^{\text{th}}$ period of the measured input spectrum $U(\mathrm{j}\omega)$ and the measured output spectrum $Y(\mathrm{j}\omega)$. The $^{\star}$ symbol denotes complex conjugate. For reasons stated in II-D, the measured input spectrum for the binary excitation case, $U_{[m,p]}$ is taken as the reference input spectrum $R_{[m]}(\mathrm{j}\omega)$.

The estimator is robust against noise disturbances and is unbiased if input noise levels are small.

### D. Supersampling

The HP VXI system is capable of any sampling frequency up to and including 10 MHz. In the experiment, the measurements were oversampled by a factor of $\mu$ above the frequency of the binary sequence input $f_{\text{c}}$. The upper frequency of the bandwidth of the random phase multisine $f_{\text{w}}$ was set equal to $f_{\text{c}}$. The measurement data were then subjected to manual downsampling by the same factor $\mu$. This results in both input signal types having identical bandwidth. The supersampling and the subsequent downsampling were performed for two main reasons detailed in the next two subsections (Ringing and Overshoot; Anti-alias).

The RC filter has low-pass (smoothing) characteristics. To downsample (or subsample) the measurement, a location of the highest peak (or lowest trough) of the output signal was taken as the reference point. From this reference point onwards and backwards every $\mu^{\text{th}}$ sample was taken as an idealised zero-order-hold (ZOH) measurement, with the ZOH clock frequency a factor $\mu$ lower than the original sampling (hence a downsampling). The reference signal $r$ and the measured output $y$ were also aligned through this reference point, so that the peaks of the output after RC filtering would then occur directly after the switching points of the binary reference input. The procedure is more easily appreciated by referring to Figure 2 by comparing the reference input (dotted line) and the two red solid lines representing the original high frequency sampling and the subsequent downsampled and aligned ZOH data. If there were no ringing, overshoot, nonlinear effects or noise, this subsampling procedure would result in a perfect reconstruction of the behaviour of an ideal ZOH sampler according to discrete-time theory. This had been verified by simulation. This subsampling procedure is performed for both linear and nonlinear measurements. Since this procedure can only be reliably performed through the easily visible binary switching points, the same alignment amount and reference point time coordinate were used for the corresponding case with multisine input.

*Ringing and overshoot:* During testing with binary excitation signals, it was observed through the oscilloscope that

Table I
Table of parameters and settings

| Symbol | Description | Value (units) | |
|---|---|---|---|
| $f_s$ | Sampling frequency for the arbitrary waveform generator, and acquisition cards. The Nyquist frequency is then $f_s/2$. | 312.5 | kHz |
| $T_s$ | Sampling interval $= 1/f_s$. | 3.2 | µs |
| $\mu$ | Over-sampling ratio for m-sequences (see Section II-D). | 8 | |
| $f_c$ | Clock frequency of the m-sequences $(= f_s/\mu)$. | $39\tfrac{1}{16}$ | kHz |
| $T_b$ | Bit interval for the m-sequences $(= 1/f_c)$. | 25.6 | µs |
| $f_w$ | Bandwidth of the multisine sequence before downsampling $(= f_c = f_s/\mu)$. | $39\tfrac{1}{16}$ | kHz |
| $f_{aa}$ | Anti-aliasing filter cut-off frequency $\equiv 0.4 f_s$. This coupling with the sampling frequency value is internally enforced by the HP1430A acquisition cards. | 125 | kHz |
| $N_{\text{base}}$ | Base length of sequence after subsampling, (= length of a 9-tap m-sequence) | 511 | Sa$^\dagger$ |
| $N$ | Length of a data record $(= \mu N_{\text{base}})$. | 4088 | Sa$^\dagger$ |
| $P$ | Number of periods measured (linear case; nonlinear case). | 12; 4 | |
| $M$ | Number of independent realisations (linear case; nonlinear case). | 5; 16 | |
| $V_{\text{rms}}$ | RMS voltage of the input signals. | 1.5 | V |

$^\dagger$Sa = samples

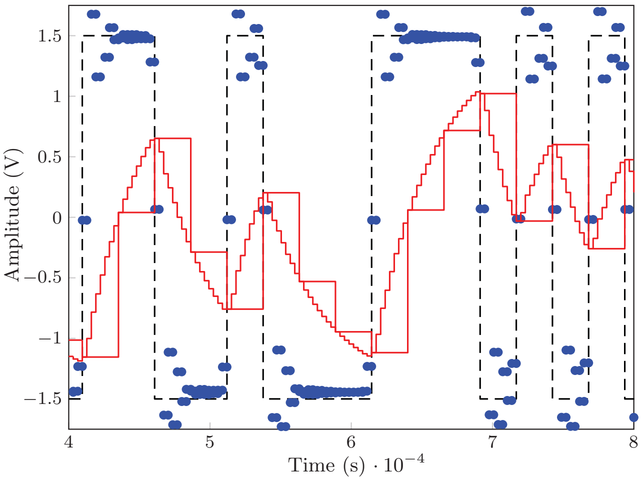


Figure 2. The use of supersampling and subsampling – Reference input (black dotted line), Measured input (blue dots), Measured output: a) supersampled and b) subsampled (red solid lines).

all operational amplifier (op-amp) based electronic buffers introduce high frequency oscillations in form of ringing to a varying extent. This is caused by non-ideal step-response characteristics when load or parasitic capacitances at output of op-amps introduce unintended poles in the transfer characteristics of the op-amps through feedback. The datasheets of many op-amps have step-response graphs which illustrate this.

In this experiment setup, the overshoots and undershoots were especially large, up to 20% with the pre-buffer due to the capacitive load at the RC circuit, even when a higher quality op-amp (with regards to its ability in driving capacitive loads) was used [8]. The overshoot depends on the load or parasitic capacitance hence the load capacitor $C$ of the RC circuit was fixed at 1.5nF for consistency.

Moreover, the HP E1430A acquisition cards themselves have significant overshoots that can be seen in the measurement data, although the oscilloscope suggested the actual acquisition inputs $u'$ and $y'$ were relatively free of such effects. This may be caused by the high order high cut-off frequency anti-aliasing filter having oscillatory step responses. The ringing at the measured input channel from an acquisition card can be seen in Figure 2. This phenomenon persisted with an Agilent 33120A waveform generator directly driving the acquisition cards, isolated completely from the system in question.

While the RC passively forms a low pass filter and is capable of minimising the effect of ringing from the pre-buffer, overshoot and ringing from the acquisition cards are inevitable. Due to the nature of sample-and-hold at the acquisition cards, the use of supersampling is necessary to obtain measurements of acceptable accuracy. The BLA theory developed is incapable of modelling in continuous time domain of such effects at the moment.

For multisine input sequences, there are no noticeable ringing or overshoot effects.

Because of the overshoot and ringing present in the measurement data from the HP E1430A acquisition cards, the signal sequence $u$ is no longer reliable and accurate representation of $u'$ in the m-sequence case. Henceforth in dealing with binary sequences, the reference signal $r$ is used as the basis for identification.

In addition, manual alignment of the measured input and output signals can be performed.

*Anti-alias:* It is necessary to minimise the effect of anti-aliasing filters on the measurements because of the use of the ideal reference signal $r$ instead of measured input $u$ in the case of binary excitations. In addition, the nonlinearity broadens the bandwidth of the output, which then may be interfered with by the anti-aliasing filter if action is not taken. Supersampling allows the internal anti-aliasing filter to be bypassed since the internal anti-aliasing filter of the HP E1430A acquisition cards have their cut-off frequencies $f_{\text{aa}}$ dependent upon the sampling frequency $f_{\text{s}}$ (see Table I). The combination of the specified low bandwidth of the multisine, the discrete nature of binary excitation signals and the low pass characteristics of RC mean that any real aliasing effect was negligible. It has been shown that broadening of spectrum due to nonlinearity would result in aliased components that are never coherent with the original input component [2, Theorem 3.21], hence the lack of anti-aliasing filter would only act as additional uncorrelated noise in the BLA measurement.

### *E. Linear measurements*

Measurements were performed to identify either a parametric or a non-parametric model for the linearity. The rms signal amplitudes for the Gaussian and binary signals were both set to 1.5 V. The non-parametric model was obtained using (1), and a parametric model was fitted where suitable using the iterative weighted nonlinear least squares procedure provided by ELiS in the *fdident* toolbox for MATLAB [9]. The weighting factors were proportional to the reciprocal of the variances at each frequency point. The isolation provided by the pre-buffer and post-buffer for the RC circuit introduced some additional linear dynamics, and hence suitable single pole models could not be fitted to the data. When a parametric model of order four is not sufficient to describe the transfer characteristics of the linearity in both the z-domain and the s-domain, the non-parametric model is used. This was the case for when the resistor value was 110 kΩ, and hence Figure 6 does not contain the results from the parametric model.

Figure 3 shows an example of the result of a non-parametric linearity identification. The noise variances indicate levels of exogenous additive noise from the environment whereas the total variances indicate the levels of nonlinear distortions plus environment noise. There is a discrepancy between the result obtained with multisine sequences and that obtained from m-sequences. This suggests input dependent nonlinear characteristics which include some effect from ringing oscillations.

As an example, for $R = 27\,\text{k}\Omega$ and with $C = 1.5\,\text{nF}$, the time constant $T_{\text{p}} = 27{\times}10^{3} \cdot 1.5{\times}10^{-9} = 4.05{\times}10^{-5}$ seconds. With a sampling interval $T_{\text{c}}$ given by $1/f_{\text{c}} = (39\,{}^{1}/_{16} \times 10^{3})^{-1}$ seconds, $T_{\text{c}}/T_{\text{p}} = 0.6321$, and therefore the theoretical transfer function is:

$$G(\text{z}) = \frac{\text{z}}{\text{z} - e^{-0.6321}} = \frac{\text{z}}{\text{z} - 0.5315}. \tag{2}$$

The parametric model identified for the m-sequence case with sampling time $T_{\text{c}}$ was:

$$\hat{G}(\text{z}) = \frac{0.01011(\text{z}+41.23)(\text{z}+0.05967)}{(\text{z}-0.5481)(\text{z}+0.01818)}. \tag{3}$$

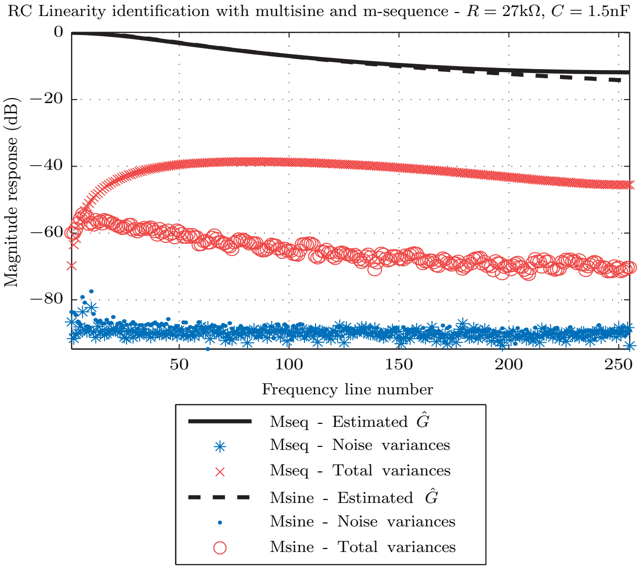


Figure 3. RC Linearity identification with op-amp based pre- and post-buffers. (To convert to frequency (Hz), line number should be multiplied by $f_{\text{s}}/(8 \times 511)$)

It can be seen that the estimated positive pole is very close to the theoretical value, but as noted above, the pre-buffer and the post-buffer to the RC circuit introduced some additional dynamics, with a negative zero and a negative pole very close to the origin, and a further negative zero that is so large that it can be regarded as a constant over the frequency range of interest.

Despite the fact that the system under test was linear, there were nonlinear distortions in both input cases and the level was higher for the binary input. This was due to to be non-linear effects from the unity-gain op-amp buffers, especially from the pre-buffer which had to drive the capacitive load. Ringing oscillations were especially noticeable with binary inputs (see II-D). If the buffers were not used, the nonlinear distortions disappear regardless of the input signal. However, due to current driving limitation of the signal generator and the capacitive load, there was unacceptable distortion of the realised input for the binary case, hence the buffers were necessary.

### *F. Nonlinear measurements and BLA theory*

The nonlinear measurements were obtained in a similar manner to the linear measurement case. The non-parametric BLA was obtained using (1).

To enable comparison with the theory, additional information is required. This includes the even higher order moments of the input signals, the signal power (or the rms value $V_{\text{rms}}$), the impulse response of the linearity and the polynomial coefficients of the nonlinearity. For the Gaussian case, the even order moments of $u$ (i.e., $\text{E}[u^{n}]$) were measured and averaged for a single experiment, for even $n$. For the binary case, $u$ was replaced by $r$ hence $\text{E}[r^{n}] = V_{\text{rms}}^{n}$. The impulse response of the linearity was taken from the parametric model

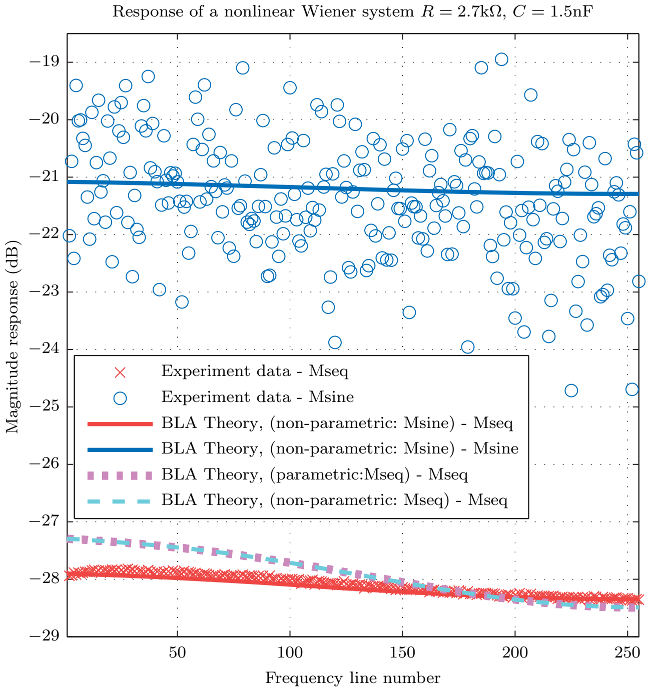


Figure 4. Experiment result of the identification of the BLA with Gaussian and binary inputs for an electronic Wiener system with non-ideal cubic nonlinearity and a RC linearity with $R = 2.7\text{k}\Omega$, $C = 1.5\text{nF}$ giving corner frequency $f_{\text{co}} = 39.3\text{kHz}$. (For conversion from frequency line number to Hz, see caption of Figure 3.)

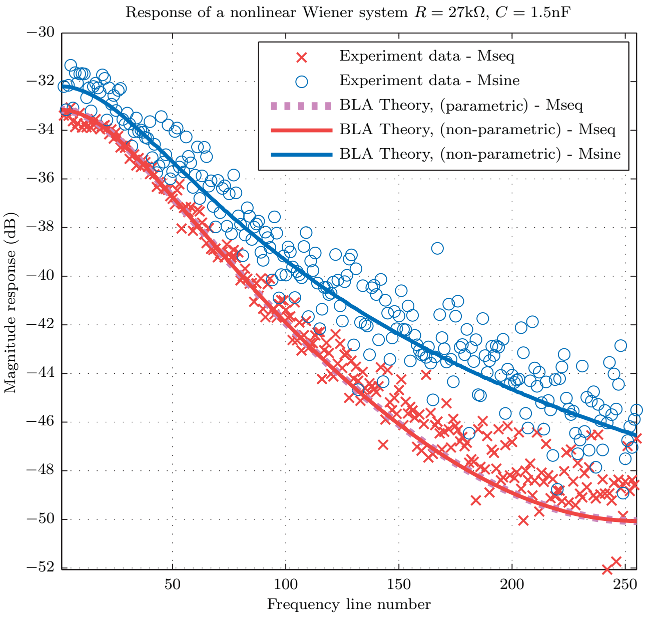


Figure 5. As Figure 4, but with $R = 27\text{k}\Omega$, giving corner frequency $f_{\text{co}} = 3.93\text{kHz}$.

if available, and the non-parametric model by inverse Fourier transform. Finally, the nonlinearity was identified by simple least squares polynomial regression performed on 20 periods of output data obtained from the nonlinearity with a multisine excitation as a direct input to the nonlinearity. The polynomial fitted to the nonlinearity was:

$$\begin{aligned} \text{f}_{\text{NL}}(x) = 0.01088x^3 - 0.001356x^2 \\ +0.008169x + 0.05816. \end{aligned} \tag{4}$$

The cubic electronic circuit had a transfer characteristic of $\text{f}_{\text{NL}}(x) = 0.01x^3$ as shown in Figure 1. Due to non-ideal characteristics there was a non-negligible quadratic term together with a linear term and a dc component in the fitted characteristic. Neither the quadratic term nor the dc offset enter into the theoretical calculations, but the linear component does, and it was taken into account in the comparisons between theory and the practical results described in Section III.

## III. Results and Analysis

Figure 4, 5 and 6 show the comparison of the BLA obtained through experiment results and those obtained from theory, for a Wiener system with (non-ideal, see (4)) cubic nonlinearity and RC filter linearity with $C = 1.5\text{nF}$ for all three cases and $R = 2.7\,\text{k}\Omega$, $27\,\text{k}\Omega$ and $110\,\text{k}\Omega$ respectively.

With $R = 2.7\,\text{k}\Omega$, the RC filter has a cut-off (or corner) frequency of $f_{\text{co}} = 1/2\pi RC \approx 39.3\,\text{kHz}$ and acts as an all-pass filter since the binary signal clock frequency was $f_{\text{c}} = 39\,{}^{1}\!/\!_{16}\,\text{kHz}$. Unfortunately this means ringing and overshoot effects (see Section II-D) were significant immediately after the RC filter stage. The linearity identification using m-sequences would yield unreliable results despite subsampling techniques. Here the use of non-parametric models of the linearity identified by a multisine was more suitable for the BLA theory. This can be seen by the fact that in Figure 4 the solid red line, representing the BLA theory based on a non-parametric linearity model identified with a multisine sequence, was able to match the experiment data represented by crosses more closely than that based on a linearity model (parametric or non-parametric) identified with a m-sequence. There are minimal differences between results derived from the non-parametric and parametric linearity models—the plots (cyan dashed and magenta heavy-dotted, respectively) are very close to each other.

For $R = 27\,\text{k}\Omega$, the RC filter had a corner frequency of approximately $3.93\,\text{kHz}$. Ringing and overshoot effects were then negligible immediately after the RC filter stage. Here the non-parametric models of the linearity for the m-sequence and multisine were used for their respective counterparts. In addition, the parametric model of (3) from Section II-E was used in the BLA theory to calculate the biased theoretical BLA for binary sequences. There were no discernible differences in the BLA theory calculated from the non-parametric and parametric linearity models as shown by the overlapping of the heavy-dotted magenta line and the solid red line in Figure 5.

When $R = 110\,\text{k}\Omega$, the RC filter had a corner frequency of about $0.965\,\text{kHz}$. The result is illustrated in Figure 6. This time parametric models up to order four produced by ELiS could not produce adequate quality fit to the transfer function of the linearity. Nevertheless the BLA theory based on the

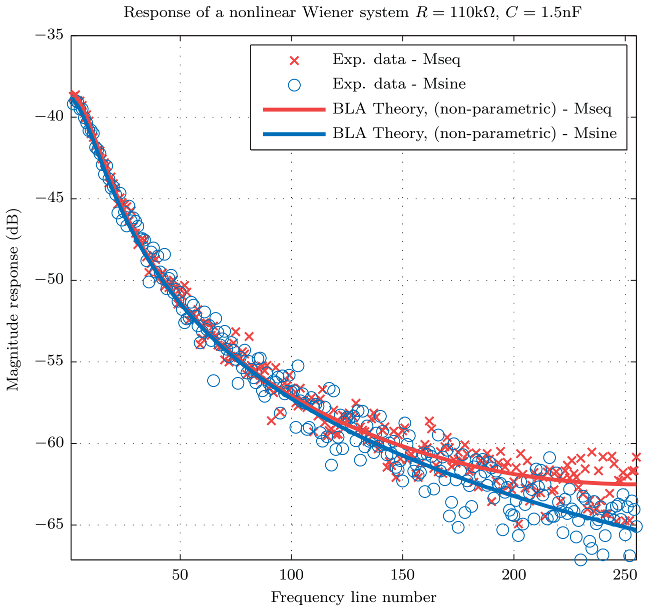


Figure 6. As Figure 4, but with $R = 110\text{k}\Omega$, giving corner frequency $f_{\text{co}} = 0.965\,\text{kHz}$.

non-parametric models was able to match the experiment data in both the gain and the shape of the transfer characteristics.

As the time constant of the system increases, the length of the impulse response of the system increases. It has been shown in Wong et al. [6] that this results in a BLA estimated by a signal with an arbitrary amplitude distribution converging to that obtained from Gaussian signal. This is also observed in Figures 4 to 6.

## IV. Conclusions

For all three sets of experiments investigated, it can be seen that the BLA theory prediction and experiment result are in good agreement. This is despite the fact that the BLA theory was based on impulse responses of the linearity modelled as finite-impulse-response (FIR) filters, whereas here the RC filter circuit is an infinite-impulse-response (IIR) filter.

The difficulties encountered with the experiment, mainly the ringing and overshoot effects illustrate a weakness in the z-domain discrete-time theory. It may therefore be beneficial to extend the theory to the continuous-time s-domain.

## Acknowledgement

H. K. Wong was supported by a Doctoral Training Grant from the U.K. Engineering and Physical Sciences Research Council (EPSRC). The work was supported in part by the Fund for Scientific Research (FWO-Vlaanderen), by the Flemish Government (Methusalem), and by the Belgian Government through the Inter-University Poles of Attraction (IUAP VI/4) Program.

## References

[1] M. Enqvist and L. Ljung, ‘Linear approximations of nonlinear FIR systems for separable input processes’, Department of Electrical Engineering, Linköping University, SE-581 83 Linköping, Sweden, Tech. Rep. LiTH-ISY-R-2718, Dec. 2005.

[2] R. Pintelon and J. Schoukens, *System identification: a frequency domain approach*, 2nd ed. Hoboken, NJ: Wiley-IEEE Press, 2012.

[3] P. M. Mäkilä and J. R. Partington, ‘Least-squares LTI approximation of nonlinear systems and quasistationarity analysis’, *Automatica*, vol. 40, pp. 1157–1169, Jul. 2004.

[4] P. M. Mäkilä, ‘On optimal LTI approximation of nonlinear systems’, *IEEE Trans. Autom. Control*, vol. 49, no. 7, pp. 1178–1182, Jul. 2004.

[5] ——, ‘LTI approximation of nonlinear systems via signal distribution theory’, *Automatica*, vol. 42, pp. 917–928, Jun. 2006.

[6] H. K. Wong, J. Schoukens and K. R. Godfrey, ‘Analysis of best linear approximation of a Wiener-Hammerstein system for arbitrary amplitude distributions’, *IEEE Trans. Instrum. Meas.*, vol. 61, no. 3, pp. 645–654, Mar. 2012.

[7] K. R. Godfrey, ‘Introduction to perturbation signals for time-domain system identification’, in *Perturbation Signals for System Identification*, K. R. Godfrey, Ed., Hemel Hempstead: Prentice-Hall, 1993, ch. 1.

[8] *AD8610A datasheet*, One Technology Way, P.O. Box 9106, Norwood, MA 02062-9106, U.S.A.: Analog Devices.

[9] I. Kollár, *Frequency Domain System Identification Toolbox for use with MATLAB*. Natick, MA: MathWorks Inc., 1994.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# *An Algorithm for Generating Real Describing Functions*

G.F.Page, J.B.Gomm and S.S.Douglas
School of Engineering, Technology & Maritime Operations
Liverpool John Moores University
Liverpool, UK
g.f.page@ljmu.ac.uk

*Abstract*— A classical method of dealing with the non-linear elements in a transfer function is to assume that they can be separated from the linear section and can then be represented by describing functions. The standard method of calculating these describing functions has been to use a graphical method which breaks the non-linear characteristic into a series of linear sections and super-imposes their effects onto a sinusoidal input. The output is assumed to be a Fourier series and the Fourier transform for the first coefficient of this series is calculated in a piecemeal fashion. The process is not difficult but there is a considerable amount of calculation involved. In this paper an algorithm is presented which enables the describing functions of real non-linearities, with any number of linearized sections, to be simply written down without the usual onerous calculations. Additionally, a method of quickly sketching the general shape of describing functions is outlined.

***Keywords- non-linearity; describing function; limit-cycle.***

## I. INTRODUCTION

This paper begins by outlining the standard graphical method for obtaining describing functions and then proceeds to use the technique to develop a general solution for obtaining the family of real describing functions. These are the describing functions whose non-linear characteristics are the same irrespective of whether the magnitude of the input signal is increasing or decreasing. Obtaining a general solution also permits the creation of algorithms for its implementation and one such algorithm is presented. Although the general methods of deriving complex describing functions are well known, the real describing functions have usually been overlooked in the literature or else have been considered as trivial. This has meant that some simplifying approaches, and their associated algorithms, have been overlooked.

The purpose of developing a general method is to enable the easy and rapid delineation of the describing functions for real non-linearities – these are non-linearities which do not possess memory. To demonstrate the effectiveness of the general method, and the associated algorithm that has also been developed, the describing functions for two common linearities have been calculated: one which caused a single limit-cycle to be produced and one which caused two limit-cycles. The predicted limit-cycle results obtained by applying the algorithm were compared with simulations using SIMULINK.

## II. THE GRAPHICAL METHOD

The early development of the graphical describing function technique can be traced to several groups working independently [1], [2], [3], [4]. However these wartime developments did not come into general use until the mid - 1950s [4], [5], [8]. It is basically an harmonic balance approach modified for feedback control. This meant that only the principal harmonic was used and higher-order oscillations were considered to be negligible due to the filtering influence of the inertia inherent in the overall process which was being controlled. The basic graphical approach is shown in Figure 1. It is assumed that the input is sinusoidal (bottom left-hand corner of the diagram) and this is mapped via the non-linearity (represented at the top left-hand corner of the diagram) to an output. The input has an input magnitude *x* plotted against time *t*. The non-linear transformation translates this input magnitude *x*, on the horizontal axis, to the output magnitude *y* on the vertical axis. The output is a plot of output *y* against time *t*. This is the same time period as for the input signal and hence the output signal *y*, due to the non-linear transformation is correlated with the input.

Assuming a sinusoidal input, the input equation will be given by:

$$x = X.\sin(\omega t) \tag{1}$$

Since the non-linearity is symmetric about the origin, and since most systems behave like low-pass filters because of inertia, the output equation will be given by:

$$y = A_1 \sin(\omega t) \tag{2}$$

Hence the describing function is

$$N(X,\omega) = \frac{A_1}{X} \tag{3}$$

and there is only a need to calculate $A_1$.

Also, because of the quarter–wave symmetry:

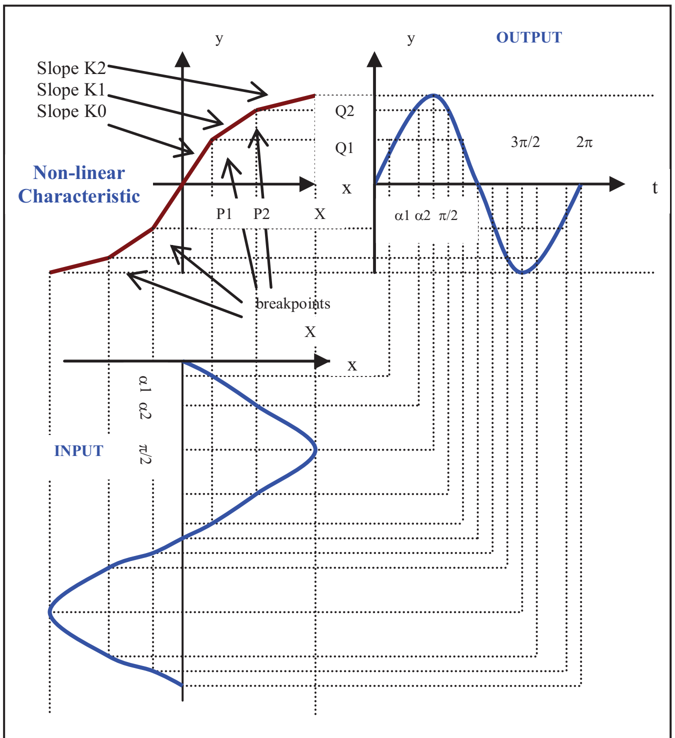


Fig. 1: The basic describing function graphical approach

$$A_1 = \frac{4}{\pi}\int_0^{\frac{\pi}{2}} y.\sin(\omega t).d(\omega t) \qquad (4)$$

The method is a quasi-linearization process in which a section of a static non-linearity is represented by a gain which depends on the magnitude of the input signal [6], [7], [9]. For this reason it is assumed that the non-linearity consists of lines of constant slope to each side of the break-points. The integrations can be calculated in a piecewise fashion to give:

$$A_1 = \frac{4}{\pi}\left[\int_0^{\alpha_1} y.\sin(\omega t).d(\omega t) \quad + \int_{\alpha_1}^{\alpha_2} y.\sin(\omega t).d(\omega t) \quad + \int_{\alpha_2}^{\frac{\pi}{2}} y.\sin(\omega t).d(\omega t)\right] \qquad (5)$$

where the individual values of $y$ in each of the separate linear sections have the form

$$y = Kx + c \qquad (6)$$

in which $K$ is the slope of the relevant section. The positions at which the slopes $K$ abruptly change value have been termed ***break-points*** in this investigation.

### *A. A General Solution*

Gibson and subsequent authors [7], [8], [9] showed how to obtain a general approach to the calculation of describing function by using this piecewise linear approach but aimed for an overall solution involving both real and imaginary parts. Also, although the early authors developed a general approach they only applied it case-by-case and did not present a general algorithm It is our opinion that a useful general algorithm can be obtained by considering single-valued non-linearities separately from double, or multi-valued, non-linearities and then specifically formulating a general solution. Single-valued non-linearities will produce real, as opposed to complex, describing functions and they can often be described by polynomial functions. By using this approach a general method for generating the describing functions of real non-linearities has been obtained. This work has resulted in a straightforward and relatively simple method of generating describing functions. An added advantage is that since there will be no phase-shifts, the superposition of the inverse Nyquist locus onto the describing function diagram is simplified since only one value of the inverse Nyquist locus, that at which it crosses the real axis, will need to be considered.

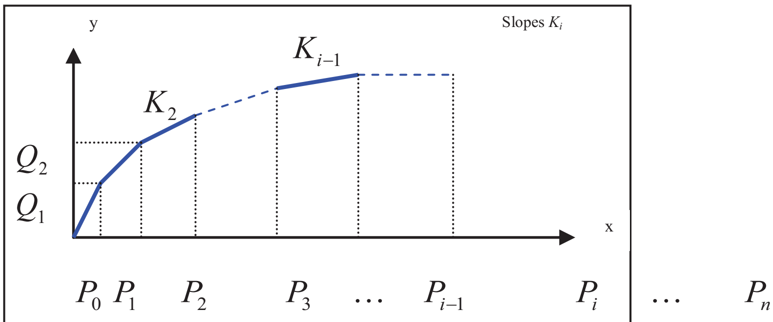


Fig. 2: Non-linear function with $n$ breakpoints

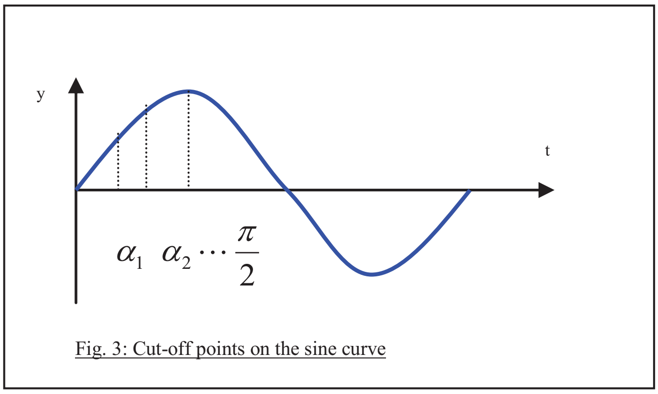


Fig. 3: Cut-off points on the sine curve

By taking Gibson's initial construction, in Fig. 1, with two breakpoints and extending it to $n$ breakpoints the graph will have $(n$ - 1) linear sections with slopes $K_0K_1 \ldots K_n$, and breakpoints occurring at horizontal positions $P_1P_2 \ldots P_n$ (with $P_0$ at the origin if necessary), jumps in the vertical plane (y-direction) at $Q_1Q_2 \ldots$ and angles on the sinusoidal input of $\alpha_1\alpha_2 \ldots \alpha_i \ldots \alpha_n, \frac{\pi}{2}$. The non-linearity will have the form shown in Fig. 2 (replacing the on-linear characteristic, top left-

GP would like to thank Prof. Ian Jenkinson, School of Engineering, Liverpool John Moores University for the facilities and support which enabled this work to be undertaken.

hand corner of Fig. 1 which maps the input signal to the output signal). The cut-off points on the sine curve will occur

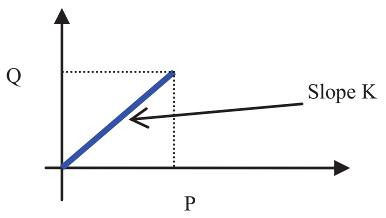


Fig. 4: Relationship between P and Q

as in Fig. 3 (replacing the sinusoidal-type of output in the top right-hand corner of Fig. 1.) The calculation of $A_1$ will be given:

$$A_1=\frac{4}{\pi}\left[\begin{array}{l}\int_0^{\alpha_1} y.\sin(\omega t).d(\omega t)+\int_{\alpha_1}^{\alpha_2} y.\sin(\omega t).d(\omega t)+\dots\\ \dots+\int_{\alpha_{i-1}}^{\alpha_i} y.\sin(\omega t).d(\omega t)+\dots+\int_{\alpha_n}^{\pi/2} y.\sin(\omega t).d(\omega t)\end{array}\right] \qquad (6)$$

which produces a general solution for the describing function:

$$N=\frac{2}{\pi}\left[K_n.\frac{\pi}{2}+\sum_{i=1}^{n}\left(K_{i-1}-K_i\right)\left(\sin^{-1}\left(\frac{P_i}{X}\right)+\left(\frac{P_i}{X}\right)\sqrt{\left[1-\left(\frac{P_i}{X}\right)^2\right]}\right)\right] \qquad \mathbf{(7)}$$

If Coulomb friction or relay action is present at initial amplitudes then equation (7) has to be adjusted.

Consider the case where only Coulomb friction is present: as $P\to 0\ \ y=Q$ , Fig. 4, $K_0=\infty\ \ and\ K_1=0$.

In this case, as $P\to 0,\ \sin^{-1}\left(\frac{P}{X}\right)\to\frac{P}{X}$ and

$$\left(\frac{P}{X}\right)\sqrt{\left[1-\left(\frac{P}{X}\right)^2\right]}\to\frac{P}{X}$$

Also from Figure (4), $P=\dfrac{Q}{K}$

Hence equation (7) reduces to $N=\dfrac{4Q}{\pi X}$ (8).

*B. An Algorithm*

An algorithm for using the general solution given in equation (7) and the special case of equation (8) to generate particular real describing functions is now presented. Although it specifically deals with discrete cases it can easily be extended to deal with continuous functions.

If Coulomb friction or relay action is present, start at stage one, otherwise start at stage two.

**Stage One:**

(a) If Coulomb friction or relay action is present then make $\dfrac{4Q}{\pi X}$ the first term of the describing function, where Q is the value of the Coulomb friction term.

(b) If dead-zone is also present multiply the above result by $\sqrt{\left[1-\left(\frac{P}{X}\right)^2\right]}$ where P is the dead-zone break-point.

**Stage Two:**

(a) If saturation is not present make $K_n$ the first term of the describing function. ($K_n$ is the gain of the last stage of the non-linearity) or add it to the result of stage one.

(b) If saturation is present then omit this term.

**Stage Three:**

(a) If there are $n$ breakpoints then add $n$ terms of the form

$$\frac{2}{\pi}(K_{i-1}-K_i)\left(\sin^{-1}\left(\frac{P_i}{X}\right)+\left(\frac{P_i}{X}\right)\sqrt{\left[1-\left(\frac{P_i}{X}\right)^2\right]}\right)$$

$$where\ i=0\to n.$$

Go to end.

(b) If saturation is present then change the last of the terms in stage 3(a) to

$$\frac{2}{\pi}K_{n-1}\left(\sin^{-1}\left(\frac{P_{n-1}}{X}\right)+\left(\frac{P_{n-1}}{X}\right)\sqrt{\left[1-\left(\frac{P_{n-1}}{X}\right)^2\right]}\right)$$

Go to end.

**End.**

## III. Derivation of Two Describing Functions Using the Algorithm

Two examples are presented (i) dead-zone plus saturation which can cause a single limit-cycle to be produced and (ii) a non-linearity which has three break-points (four pseudo-linear regions) and so can cause two nested limit-cycles to be produced. The effects of these non-linearities were then simulated by placing them in series with a transfer function for a third-order linear system and applying unity feedback. The simulations were created using the SIMULINK package.

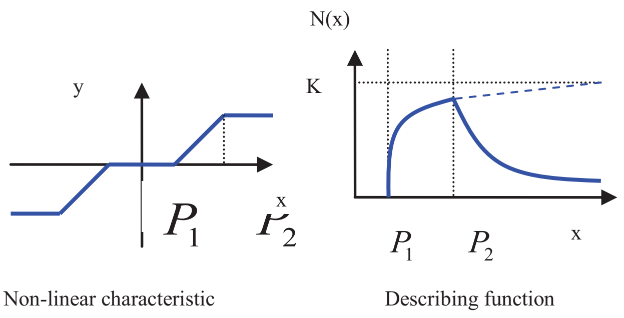


Fig. 5: dead-zone plus saturation

### A. *Dead-zone plus saturation*

The parameters are $n=2, K_0=0, K_1=K, K_2=0$ so equation (7) gives:

$$N=\frac{2K}{\pi}\left[\sin^{-1}\left(\frac{P_2}{X}\right)-\sin^{-1}\left(\frac{P_1}{X}\right)+\left(\frac{P_2}{X}\right)\sqrt{\left[1-\left(\frac{P_2}{X}\right)^2\right]}-\left(\frac{P_1}{X}\right)\sqrt{\left[1-\left(\frac{P_1}{X}\right)^2\right]}\right]$$

*when $X>P_2$; $N$ as for deadzone when $P_1<X\le P_2$ and $N=0$ when $X\le P_1$*

Using the algorithm, stages two (b), three (a) and three (b) apply and give the same result as above. This result is shown graphically in Fig. 5.

### B. *A non-linearity with three break-points (four slopes)*

In this case only one combination of slopes of the pseudo-linear sections has been considered: $K_0>K_1\ \&\ K_1<K_2\ \&\ K_2>K_3$.

In particular, $K_0=-0.7, K_1=0.2, K_2=1.5\ \&\ K_3=0.2.$ Again, the results were obtained by using equation (7), and by the algorithm, to give

$$N=\frac{2}{\pi}\left[\begin{array}{l}K_3.\frac{\pi}{2}+\left(K_2-K_3\right)\left[\sin^{-1}\left(\frac{P_3}{X}\right)+\left(\frac{P_3}{X}\right)\sqrt{\left[1-\left(\frac{P_3}{X}\right)^2\right]}\right]+\ldots\\ \left(K_1-K_2\right)\left[\sin^{-1}\left(\frac{P_2}{X}\right)+\left(\frac{P_2}{X}\right)\sqrt{\left[1-\left(\frac{P_2}{X}\right)^2\right]}\right]+\ldots\\ \left(K_0-K_1\right)\left[\sin^{-1}\left(\frac{P_1}{X}\right)+\left(\frac{P_1}{X}\right)\sqrt{\left[1-\left(\frac{P_1}{X}\right)^2\right]}\right]\end{array}\right]$$

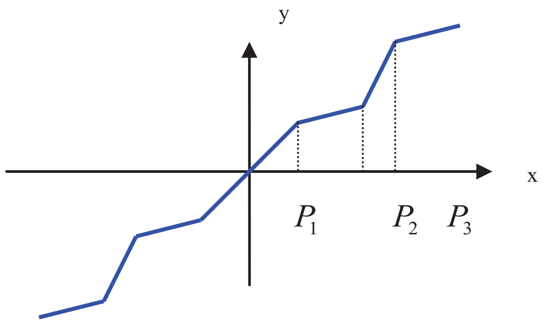


Fig.6(a): Non-linearity with three breakpoints

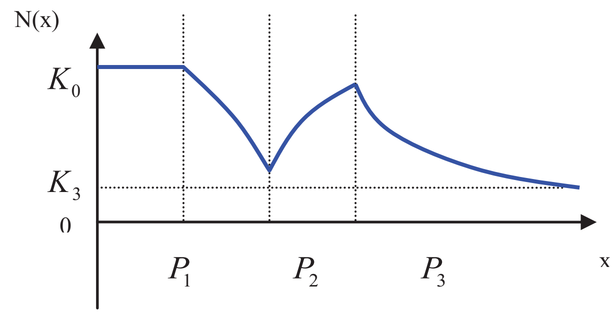


Fig. 6(b): Describing function for the three-breakpoint case

Using the algorithm, stages two (a) and three (a) apply. The graphical results are shown in Fig. 6(a) and Fig. 6(b).

## IV. USE OF REAL DESCRIBING FUNCTIONS TO PREDICT LIMIT-CYCLES

If the describing function is represented by N(X,ω) and the open-loop transfer function of a system is represented by G(jω) the Kochenburger's Stability Criterion [4] states that, in order for a system to remain stable, the locus $|G(j\omega)|$ must keep the entire locus ${-1}/{|N(X,\omega)|}$ on the right; or the inverse locus ${1}/{|G(j\omega)|}$ must keep the locus $-|N(X,\omega)|$ on the left (or must completely enclose the whole of the locus). For this work the authors found that the inverse Nyquist approach was more convenient. Furthermore, since only systems with real, as opposed to complex, describing functions were being investigated, plots with real and imaginary axes were of little use. It was better to plot the magnitude of the describing functions against the magnitude of the input signal and to superimpose on this the magnitude of the inverse Nyquist value at which it crossed the real axis. The position at which the descending describing function locus crossed the inverse Nyquist value then enabled the magnitude of the limit-cycle to be determined – as shown in Fig. 7.

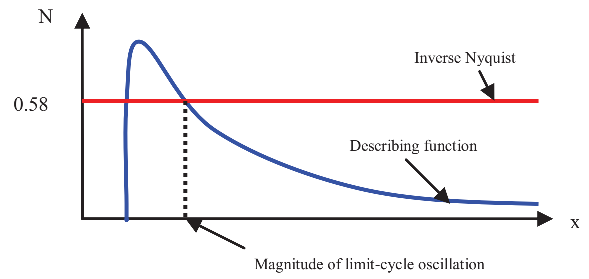


Fig. 7: A real describing function locus with an inverse Nyquist magnitude superimposed on it.

A similar approach was tried [10] using the direct Nyquist instead of the inverse function but it didn't lend itself to the same predictive opportunities.

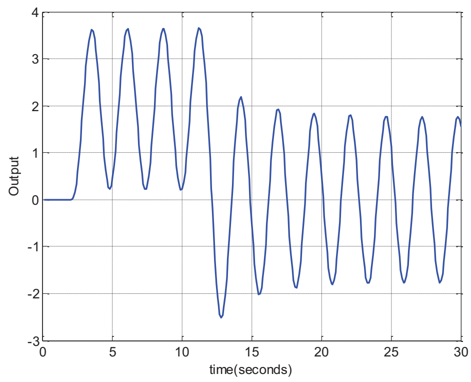


Fig. 8: Limit-cycle oscillation with saturation and dead-zone

### A. Dead-zone plus saturation

According to the basic describing function locus shown in Fig. 5 this has the potential to exhibit the limit-cycle effect. The system was simulated in series with a third-order transfer function and the oscillatory response shown in Fig. 8 was obtained. This oscillatory response was produced with the limits of the saturation non-linearity set to ±1 and dead-zone set to ±0.5.

The calculated magnitude of the limit-cycle, from Fig. 10, is 1.79 ± 0.02. The actual magnitude of the limit-cycle, from Fig. 9, is 1.78 ± 0.03. The calculated frequency of oscillation is 2.45±0.001 rad/s and, from Figure 9 the measured frequency of oscillation is 2.44 ± 0.03 rad/s

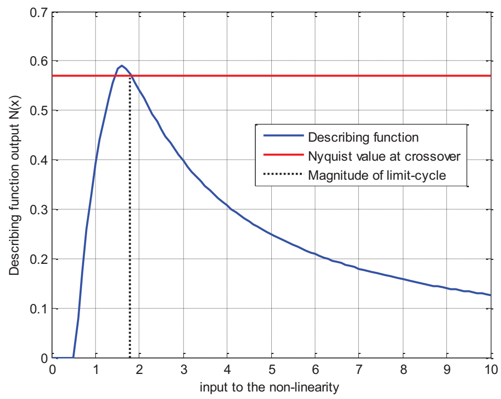


Fig. 9: Plot of describing function of dead-zone plus saturation

### B. The non-linearity with three break-points (four slopes)

This time there were two positions where limit-cycles might occur depending on the signal input magnitude.

From Fig. 10, the measured frequency of the limit-cycle oscillation was 2.41 ± 0.07 rad/s which compared with the calculated limit-cycle frequency of 2.45 ± 0.001 rad/s. From Figure 11 the calculated magnitude of the lower limit-cycle was 1.60 ± 0.05 and of the higher limit-cycle it was 9.2 ± 0.08. From Fig. 10 the actual magnitude of the lower limit-cycle was 1.52 ± 0.7 and of the higher limit-cycle it was 10.1 ± 1.2.

It was found, by successively increasing the values of the second impulse magnitude, that the second limit-cycle was reached once the impulse had been raised above about 5.6, although it took several oscillation to reach this new stable position. Again, this could be predicted from Figure 11 where the rising value of the describing function crossed the Inverse Nyquist.

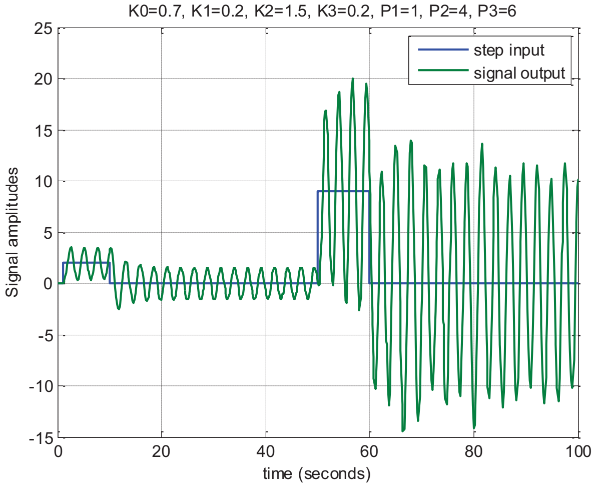


Fig. 10: The three-breakpoint response showing two limit-cycles

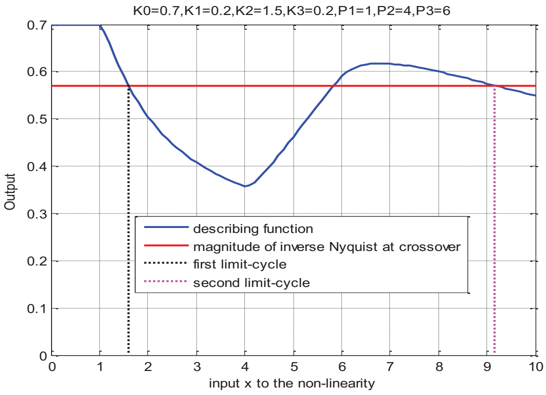


Fig.11: Describing function for the three break-point characteristic

## V. A QUICK METHOD OF SKETCHING DESCRIBING FUNCTIONS

In every case investigated, both in those shown in Figures 5 and 6(b) and in more complicated real non-linearities, the slope of the 'linear' section became the asymptotic value to which that section of the describing function tended. Also, there was an abrupt change in the describing function at the same value of x as that at which the break-value of the non-linearity occurred. When $|K_n| > |K_{n-1}|$ the locus increased in value approximately to the formula: $y = K_{n-1} + (K_n - K_{n-1})(1 - e^{-x})$ and when $|K_n| < |K_{n-1}|$ the locus decreased in value approximately to the formula: $y = K_n + (K_{n-1} - K_n)e^{-x}$. A more precise mathematical description of each type of locus was not considered necessary since the observed behaviour has only been used in subsequent work to afford a rough sketch of the shape of the describing function. A sketch can be performed more rapidly than using the algorithm to give a general outline of the shape before the algorithm itself is used to give more precise results.

## VI. CONCLUDING REMARKS

This work has used the classical method for deriving describing functions. However, by restricting the formulation to deal only with those non-linearities which produce real, not complex, describing functions it has been possible to devise an algorithm which enabled such functions to be simply written down without the usual onerous calculations. Furthermore, the method could be rapidly applied to non-linearities of considerable complexity. Also, because only real describing functions were being considered, it was possible to use a simplified graphical form of Kochenburger's criteria to derive the positions of limit-cycles and also the range of inputs needed to induce them. After the derivation of the algorithm two examples of non-linearities have been presented, one which produces a single limit-cycle and one which produces two nested limit-cycles. The expected parameters which produced these effects have been calculated and compared with the actual results obtained by simulation. Finally a quick method of producing a rough sketch of the general shape of a describing function has been included. This paper presents the first stages of a series of investigations into non-linear effects and their control.

## REFERENCES

[1] Tustin, A., The effects of backlash and of speed dependent friction on the stability of closed cycle control systems, JIEE, part II, Vol. 94, pp 143-151, 1947.

[2] Dutilh, J., Theorie des servomechanism a relais, Onde Elec., pp 438-445, 1950.

[3] Oppelt, W., Locus curve method for regulators with friction, Z. Deut. Ingr., Berlin, p 90, 1948.

[4] Kochenburger, R.J., A frequency response method for analysing and synthesising contactor servomechanisms, Trans. AIEE, Vol. 69, pp 270-283, 1950.

[5] Atherton, D.P., Nonlinear Control Engineering, Van Nostrand, 1975.

[6] Atherton, D.P., Stability of Non-Linear Systems, Research Studies Press, Wiley, 1981.

[7] Gibson, J.E., Nonlinear Automatic Control, McGraw-Hill, New York, pp 405-410, 1963.

[8] Khalil, H.K., Non-linear System 3rd Ed., Prentice-Hall, pp 54-59, 280-288, 2002.

[9] Dutton, K., Thompson, S. and Barraclough, , B., The Art of Control Enginerering, Prentice-Hall, pp 698-710 , 1997.

[10] Kim, E., Lee, H., and Park, M., Limit-Cycle Prediction of a Fuzzy Control System Based on Describing Function Method, IEEE Trans. Fuzzy Systems, Vol. , No. 1, pp 11-22, 2000.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Estimation of pulmonary elastance fuzzy model by data combination of two respiration phases

M. Nakamichi*, S. Kanae*, Z.J. Yang† and K. Wada‡
*Fukui University of Technology,
3-6-1 Gakuen, Fukui-city, Fukui 915-8505, Japan
†Ibaraki University,
4-12-1 Makanarusawa, Hitachi-city, Ibaraki 316-8511, Japan
‡Kyusyu University,
744 Motooka, Nishi-ku, Fukuoka-city, Fukuoka 819-0395, Japan
Email: kanae@fukui-ut.ac.jp

***Abstract*—Pulmonary characteristics differ in patients, and the suitable setting of ventilation condition is needed for every patient in the artificial respiration. The pulmonary elastance is one of the important features of lung, and it is a basis for deciding the airway pressure limit value. To get the pulmonary elastance of the of the patient from measurement data of the artificial respiration, the fuzzy logic technique has been proposed for estimating the pulmonary elastance and the static $P-V$ curve in our previous works. In this paper, a new technique of fuzzy modeling based on data combination of two respiration phases is proposed to improve the estimation precision, and some estimation examples using real patient data are given to illustrate the superiority of the proposed method over the previous algorithm in the precision.**

*Key Words*—**modeling, estimation, fuzzy logic, artificial respiration, elastance**

## I. INTRODUCTION

Artificial respirator is used for patients with little or no autonomous breathing ability. Using a respirator, particular attention should be paid to set suitable ventilation condition. Pulmonary characteristics differ in patients and change by the extent of illness progress or recovery. Now, the setting of the respirator is decided by the experience and the intuition of the doctor. This is a problem that there is no appropriate method for deciding ventilation condition. This problem may cause a medical accident. These circumstances motivate us to develop a method to estimate the pulmonary elastance of the patient and to set a ventilation condition of the artificial respirator.

In our previous work, we have presented an estimation technique of the pulmonary elastance by fuzzy logic, and have represented the static $P-V$ curve needed for the respirator setting. However, the estimation precision of previous technique was not satisfactory, the estimated static $P-V$ curve cannot be used for the respirator setting. In this paper, we improved the estimation precision by new technique that combine the measurement data of two phases of breath. Furthermore, this paper shows some estimation examples using real patient data to illustrate the superiority of the proposed method over the previous algorithm in the precision.

The paper is organized as follows. Section 2 introduces the static $P-V$ curve and nonlinear differential equation model of respiratory systems. A new method that estimating the pulmonary elastance by fuzzy logic is described in Section 3. Some examples of estimation are given in Section 4. Finally, Section 5 concludes this paper.

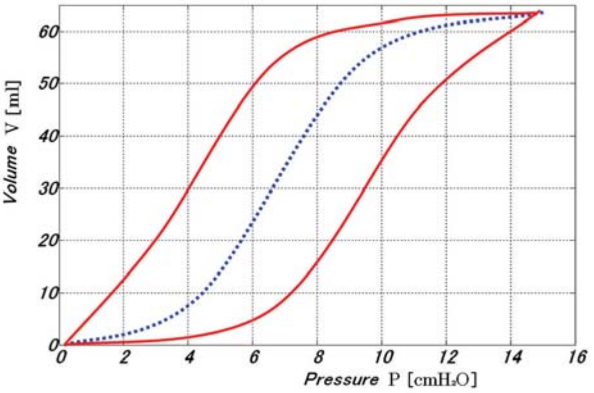


Fig.1. The dynamical $P-V$ curve (solid line) and the static $P-V$ curve (dotted line).

## II. NONLINEAR DIFFERENTIAL EQUATION MODEL OF RESPIRATORY SYSTEMS

The static state is a state without the flow of air in the respiratory system. In this state, the static $P-V$ curve is drawn by the pressure ($P_l(t)$) and the volume ($V(t)$) inside of lung. The gradient of the static $P-V$ curve is the compliance, and the inverse of the compliance is the elastance of the lung. The pulmonary elastance ($f_E(V)$) is as a nonlinear function of $V(t)$, and the static $P-V$ curve is expressed by the following equation:

$$P_l(t) = f_E(V)V(t). \qquad (1)$$

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

This static $P-V$ curve expresses the important feature of the lung, and it is a basis for deciding the air-way pressure limit value.

Fig.1 shows the dynamic $P-V$ curve (solid line) and the static $P-V$ curve (dotted line) by respiratory data over one period. The dynamic $P-V$ curve can be drawn by air-way pressure and air-volume of lung that can be measured directly. However, the static $P-V$ curve cannot be drawn directly, because inside pressure of lung cannot be measured directly. So, we must estimate the static $P-V$ curve by technique of the system identification.

For a rational setting of the respirator, a mathematical model is needed to describe the respiration. Some related works are addressed in the literature[2][3]. The lung of human is divided into right and left. Even if each lung is modeled by a simple first order differential equation, the synthesized overall respiratory system model becomes to second order differential equation [1]. Respiratory system model as the following equation is proposed by Kanae *et al*,

$$\begin{aligned} P_{ao}(t) + a_1\dot{P}_{ao}(t) &= f_E(V)V(t) \\ &+ g_R(\dot{V})\dot{V}(t) + b_2\ddot{V}(t) + P_{eea} + \epsilon(t), \end{aligned} \tag{2}$$

where $P_{ao}(t)$ is air-way opening pressure, $F(t) = \dot{V}(t)$ is air-flow, $P_{eea}$ is the end-expiratory alveolar pressure. In addition, $g_R(\dot{V})$ is so-called air-way resistance. The pressure loss $P_r$ is expressed by the following equation,

$$P_r(t) = g_R(F)F(t) = (r_1 + r_2|\dot{V}(t)|)\dot{V}(t). \tag{3}$$

The pulmonary elastance and the air-way resistance are described by polynomials of volume $V(t)$. $\epsilon(t)$ contains the modeling error and the measurement noise. In the model (2), the order of equation are the same as the conventional model[2]. Therefore, the model has the capability of describing the nonlinear dynamical characteristics of respiration. We can measure air-way opening pressure, volume, air-flow in each sampling period.

The fuzzy logic is suitable for expressing a nonlinear function, and can reduce computational complexity in the estimation by the fuzzy division. After that, using this respiratory system model (2), this study presents a method to estimate the pulmonary elastance by fuzzy logic.

## III. FUZZY MODEL OF RESPIRATORY SYSTEMS

### A. Fuzzy Logic

Mamdani's fuzzy IF-THEN rules are very famous in fuzzy logic[4]. The structure of these rules is that all inputs are set in part of the antecedent and all outputs are set in part of the consequent.

In this study, the volume $(V(t))$ is defined as the part of the antecedent and pulmonary elastance of the first order function of $(f_E(V) = k_1+k_2V)$ is defined as the part of the consequent. Fuzzy variables in part of the antecedent are defined as $small$, $medium$, $big$.

### B. Previous Method

Using technique called Functional Type SIRMs (Single Input Rule Modules) fuzzy reasoning method, We estimated the pulmonary elastance in previous study. Functional Type SIRMs fuzzy reasoning method is proposed by Seki *et al* [5]. Firstly, this method defines IF-THEN fuzzy rule modules of single input type. The overall reasoning result is weighted sum of the reasoning results of each rule modules. In this study, input term $V$ is divided into the inspiratory volume $V_{in}$ and the expiratory volume $V_{out}$. Therefore, fuzzy rules is structured as follows,

$$\begin{aligned} & Rule - V_{out}: \\ & if \;\; V_{out} = small_{out}, \\ & \;\; then \;\; f_{Eout(small)} = k_{1out(small)} + k_{2out(small)}V; \\ & if \;\; V_{out} = medium_{out}, \\ & \;\; then \;\; f_{Eout(medium)} = k_{1out(medium)} + k_{2out(medium)}V; \\ & if \;\; V_{out} = big_{out}, \\ & \;\; then \;\; f_{Eout(big)} = k_{1out(big)} + k_{2out(big)}V. \\ & \\ & Rule - V_{in}: \\ & if \;\; V_{in} = small_{in}, \\ & \;\; then \;\; f_{Ein(small)} = k_{1in(small)} + k_{2in(small)}V; \\ & if \;\; V_{in} = medium_{in}, \\ & \;\; then \;\; f_{Ein(medium)} = k_{1in(medium)} + k_{2in(medium)}V; \\ & if \;\; V_{in} = big_{in}, \\ & \;\; then \;\; f_{Ein(big)} = k_{1in(big)} + k_{2in(big)}V. \end{aligned} \tag{4}$$

The consequent part parameters of fuzzy rules (4) are estimated by the numerical integration technique and the least squares method[1][6]. The reasoning result of each rule modules is calculated by the Center of Gravity Method as follows,

$$\begin{aligned} & Rule - V_{out}: \\ & h_j^{V_{out}} = A_j^{V_{out}}(V^0), \qquad (j = small, medium, big), \\ & f_{Eout}^0 = \frac{\sum_{j=small}^{big} h_j^{V_{out}} f_{Eout(j)}(V^0)}{\sum_{j=small}^{big} h_j^{V_{out}}}; \end{aligned} \tag{5}$$

$$
\begin{aligned}
&Rule - V_{in}: \\
&h_j^{V_{in}} = A_j^{V_{in}}(V^0), \qquad (j = small, medium, big), \\
&f_{Ein}^0 = \frac{\displaystyle\sum_{j=small}^{big} h_j^{V_{in}} f_{Ein(j)}(V^0)}{\displaystyle\sum_{j=small}^{big} h_j^{Vin}},
\end{aligned}
\tag{6}
$$

where, $A$ is fuzzy set of the antecedent variable $V$, $h_j$ is the conformity degree of the antecedent.

The overall reasoning result $f_E^0$ can be calculated as follows:

$$
f_E^0 \;=\; \omega_{Vout}\, f_{Eout}^0 \;+\; \omega_{Vin}\, f_{Ein}^0, \tag{7}
$$

where, $\omega_i$ ( $i = Vin, Vout$ ) is so-called the serious consideration degree for the overall reasoning result.

Finally, $f_E^0$ is substituted for relation equation between the pressure ($P_l$) and the volume ($V$) inside of lung as follows:

$$
P_l \;=\; f_E^0\, V. \tag{8}
$$

The static $P - V$ curve is drawn by equation (8).

However, estimation precision for this technique was not satisfactory. We will explain a process to improve estimation precision later.

### C. Estimation of consequent part parameters of the fuzzy rule

This section introduces the method that estimates the function parameters of consequent part about an arbitrary rule in fuzzy rule of equation (4). Using the respiratory system model, the first order function ($f_E(V) = k_1 + k_2 V$) substitutes the relationship of

$$
\begin{aligned}
P_{ao}(t) + a_1\dot{P}_{ao}(t) &= k_1 V(t) + k_2 V^2(t) \\
&\quad + g_R(\dot{V})\dot{V}(t) + b_2\ddot{V}(t) + P_{eea} + \epsilon(t)\ .
\end{aligned}
\tag{9}
$$

Define data vector $\varphi(t)$ and parameter vector $\theta$ as follows: $\varphi^T(t) = [-\dot{P}_{ao}(t),\ V(t),\ V^2(t),\ \dot{V}(t),\ |\dot{V}(t)|\dot{V}(t),\ \ddot{V}(t),\ 1.0]$; $\theta^T = [a_1, k_1, k_2, r_1, r_2, b_2, P_{eea}]$. Then, using relationship between the volume and the flow ($F(t) = \dot{V}(t)$), data vector $\varphi(t)$ can be written as follows: $\varphi^T(t) = [-\dot{P}_{ao}(t),\ V(t),\ V^2(t),\ F(t),\ |F(t)|F(t),\ \dot{F}(t),\ 1.0]$ .

The model equation can also be written in short form as follows:

$$
P_{ao}(t) \;=\; \varphi^T(t)\theta + \epsilon(t)\ . \tag{10}
$$

Equation (9) is continuous-time model. An identification model is obtained by applying the numerical integration technique which is known as an effective approach for continuous-time model identification[7].

Measurement data are sampled data of air pressure $P_{ao}(k)$, flow $F(k)$, volume $V(k)$, where $k$ ( $k = 1, 2, \cdots, N$ ) denotes sampling instant. $N$ define the data size. $T$ define the sampling period of data collection. Then, at time instant $t = kT$, integrate both sides of equation (10) over the interval $[(k-\ell)T, kT]$. Using the numerical integration technique that is proposed by Sagara *et al*, left side of equation (10) can be calculated as follows:

$$
y(k) \;=\; \int_{(k-\ell)T}^{KT} P_{ao}(\tau)d\tau \;\doteq\; \sum_{j=0}^{\ell} g_j P_{ao}(k-j), \tag{11}
$$

where, $\ell$ is a natural number that decides the window size of numerical integration. The coefficients $g_i (i = 1, 2, \cdots, \ell)$ are determined by formulae of numerical integration. When the trapezoidal rule is taken, they are given as follows:

$$
\begin{cases}
g_0 = g_\ell = T/2, \\
g_i = T, \quad i = 1, 2, \cdots, \ell - 1.
\end{cases}
\tag{12}
$$

As calculation of equation (11), data vector $\varphi(t)$ can be calculated by

$$
\begin{aligned}
\phi(k) \;&=\; \int_{(k-\ell)T}^{KT} \varphi(\tau)d\tau \\
&= \begin{bmatrix}
-P_{ao}(k) + P_{ao}(k-\ell) \\
\displaystyle\sum_{j=0}^{\ell} g_j V(k-j) \\
\displaystyle\sum_{j=0}^{\ell} g_j V^2(k-j) \\
V(k) - V(k-\ell) \\
\displaystyle\sum_{j=0}^{\ell} g_j |F(k-j)| F(k-j) \\
F(k) - F(k-\ell) \\
\ell T
\end{bmatrix}.
\end{aligned}
\tag{13}
$$

Get together the approximation error $\Delta_E$ caused by numerical integration and the integral of original error term $\epsilon$ in $e(k)$. Namely, Let $e(k)$ be

$$
e(k) \;=\; \Delta_E \;+\; \int_{(k-\ell)T}^{KT} \epsilon(\tau)d\tau. \tag{14}
$$

Consequently, an identification model of discrete-time form is provided from equation (11), equation (13) and equation (14) as follows:

$$
y(k) \;=\; \phi^T(k)\theta \;+\; e(k)\ . \tag{15}
$$

From the measurements of air pressure $P_{ao}(k)$, flow $F(k)$, volume $V(k)$, calculates $y(k)$ and $\phi(k)$ at each time instant $k = \ell + 1, \cdots, N$. And the vector equation is structured by them as follows:

$$
\mathbf{y} = \mathbf{\Phi}\theta + \mathbf{e}, \tag{16}
$$

where, $\mathbf{y} = [y(N)\cdots y(\ell+1)]^T$, $\mathbf{\Phi} = [\phi(N)\cdots\phi(\ell+1)]^T$, $\mathbf{e} = [e(N)\cdots e(\ell+1)]^T$.

The least squares estimate that minimizes the criterion function $J = \| \mathbf{y} - \mathbf{\Phi}\theta \|^2$ is given by

$$\hat{\theta} = (\mathbf{\Phi^T \Phi})^{-1}\mathbf{\Phi^T y}. \tag{17}$$

Then, the function of consequent part about an arbitrary rule in fuzzy rules (4), in other words, the function of pulmonary elastance is estimated as

$$\hat{f}_E(V) \ = \ \hat{k}_1 + \hat{k}_2 V \ . \tag{18}$$

Executing the above-mentioned calculation on all fuzzy rules, fuzzy rules are made.

### D. New Method

It is considered that there are two reasons to result unsatisfactory precision of estimation. One is that the design of serious consideration degree has to be performed by hand-operation. Another one is that there is relatively large error in each connection area of two fuzzy rules, and the whole precision becomes worse as the partition number of variable increased. We are going to solve these two problems here.

The idea is to combine the data of two respiratory phases. Firstly, prepares the regression data in each range of fuzzy variables according to equation (16). Then, combines the data of two phases of expiratory and inspiratory respiration in every same range. Consequently, the fuzzy rules are combined as follows:

$$\begin{aligned}
&if\ V = small,\\
&\quad then\ f_{E(small)} = k_{1(small)} + k_{2(small)}V;\\
&if\ V = medium,\\
&\quad then\ f_{E(medium)} = k_{1(medium)} + k_{2(medium)}V;\\
&if\ V = big,\\
&\quad then\ f_{E(big)} = k_{1(big)} + k_{2(big)}V.
\end{aligned} \tag{19}$$

Using combined data, the parameters of the consequent part of the fuzzy rules (19) are estimated by the least squares method as mentioned above.

Using above fuzzy rules (19), the overall reasoning result of pulmonary elastanse is estimated. The reasoning result of fuzzy rules (19) is calculated by the Center of Gravity Method as follows:

$$\begin{aligned}
h_j^V &= A_j^V(V^0), \qquad (j = small, medium, big),\\
f_E^0 &= \frac{\sum_{j=small}^{big} h_j^V f_{E(j)}(V^0)}{\sum_{j=small}^{big} h_j^V}.
\end{aligned} \tag{20}$$

where, $A$ is fuzzy set of the antecedent variable $V$, $h_j$ is the conformity degree of the antecedent. The static $P-V$ curve is drawn by equation (8).

In this way, problems of a calculation error and the serious consideration degree are solved by combining the data of two phases of breath.

## IV. EXAMPLES OF ESTIMATION

In this section, some experimental results are shown to estimate the static $P-V$ curve with real clinical data of artificial respiration. Here, ranges of the fuzzy variables ($small$, $medium$ and $big$) are set by manual operation and each of the serious consideration degree is assumed as degree of 0.5. Fig.2 and Fig.4 are membership functions of the fuzzy variables in each experiment. Fig.3 and Fig.5 are experimental results that estimate static $P-V$ curve with real clinical data of artificial respiration. The circle points and the square points are quasi-static elastance $P-V$ points for inspection. It is shown that the estimated static $P-V$ curve passes near of inspection data similar the true static $P-V$ curve. Table.1 is mean-square error of estimated static $P-V$ curve by each method to inspection data. Table.1 shows that estimated static $P-V$ curve by new method passes near of inspectional data points.

It is important that we must set suitable range of the fuzzy variables. In these experiments, we were able to find suitable values by a hand-operated design. However, the estimated precision is not satisfactory when fuzzy variables are not suitable values. Considering of the problems mentioned above and pulmonary characteristics that differ in patients, it is necessary to devise optimization algorithm for range setting of the fuzzy variables. Finally, we aim at the automation of estimation of pulmonary elastance and the improvement of estimated precision in the future.

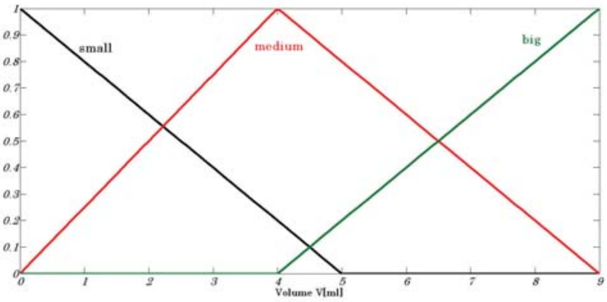


Fig.2. Data1: Membership functions of fuzzy variables.

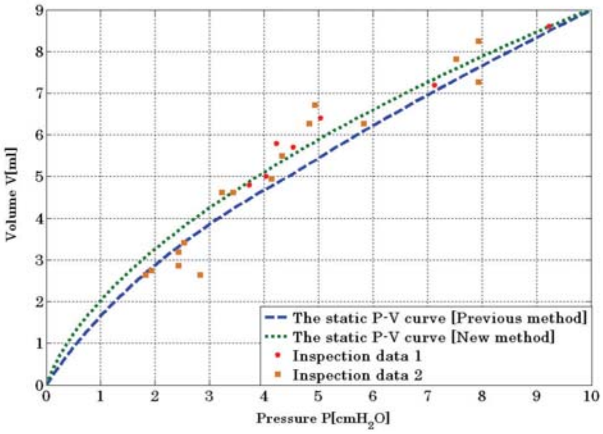


Fig.3. Data1: Estimated static $P - V$ curve and experimental static elastic recoil pressure-volume points(◯, □).

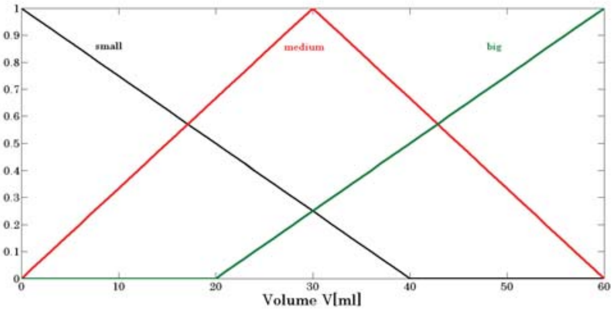


Fig.4. Data2: Membership functions of fuzzy variables.

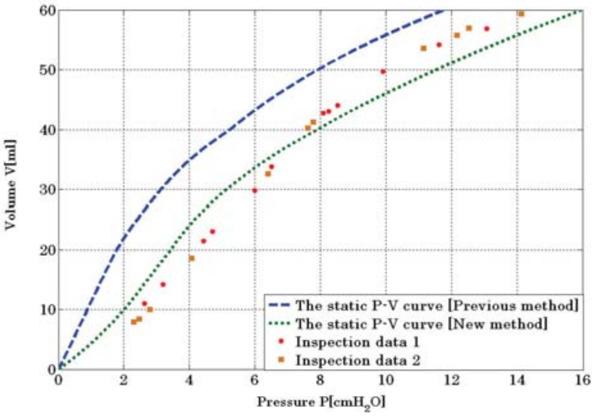


Fig.5. Data2: Estimated static $P - V$ curve and experimental static elastic recoil pressure-volume points(◯, □).

## V. CONCLUSIONS

The pulmonary elastance is one of the important features of lung, and it is a basis for deciding the air-way pressure limit value. In this study, an estimation method of pulmonary elastance based on fuzzy logic is proposed. Using technique to combine the data of two phase of breath, one fuzzy rule that merged the expiratory fuzzy rule with the inspiratory fuzzy rule is made. The reasoning result of fuzzy rule is calculated by the Center of Gravity Method.

TABLE I
MEAN-SQUARE ERROR OF ESTIMATED STATIC $P - V$ CURVE BY EACH METHOD TO INSPECTION DATA.

| | Previous [cmH$_2$O] | New [cmH$_2$O] |
|---|---|---|
| Data 1 | 0.5658 | 0.3525 |
| Data 2 | 5.4845 | 1.2985 |

As a result of estimated pulmonary elastance with real clinical data of artificial respiration, we can draw the static $P-V$ curve that passes near of inspectional data points. However, it is necessary to set suitable range of the fuzzy variables of the antecedent part. Therefore, by devising optimization algorithm for range setting, we aim at the automation of estimation of pulmonary elastance and the improvement of estimated precision in the future.

## REFERENCES

[1] S. Kanae: The setting of the ventilation condition of the artificial respiration and the modeling of respiratory system, Measurement and Control Vol.49, No.7, pp.485-488, in Japanese, 2010.

[2] K. Muramatsu, K. Yukitake, M. Nakamura, I. Matsumoto, Y. Motohiro: Monitoring of Nonlinear Respiratory Elastance Using a Multiple Linear Regression Analysis, European Respiratory Journal 17, pp.1158-1166, 2001.

[3] S. Kanae, K. Muramatsu, Z.J. Yang, K.Wada: Modeling of respiration and estimation of pulmonary elastance, 2004 ASCC, pp.648-651, 2004.

[4] M. Sugeno: Fuzzy control, Daily publication industry newspaper publishing company, in Japanese, 1990.

[5] H. Seki, H. Ishii, M. Mizumoto: On the Nonlinear Identification by Functional Type SIRMs Connected Type Fuzzy Reasoning Method (Theory of Modeling and Optimization), Record of institute for mathematics analysis 2006, Vol.1526, pp.173-180, in Japanese, 2006.

[6] M. Nakamichi, S. Kanae, Z.J. Yang, K. Wada: Estimation of pulmonary elastance by functional type SIRMs fuzzy reasoning method, The 30th CCC, pp.6105-6108, 2011.

[7] S. Sagara, Z. Zhao: Numerical Integration Approach to On-Line Identification of Continuous-Time Systems, Automatica 26, pp.63-74, 1990.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Developing Real-Time System Identification for UAVs

Pierre-Daniel Jameson *, Alastair Cooke †
School of Engineering
Department of Aircraft Enginering
Dynamics, Simulation, and Control Group
Cranfield University, Bedfordshire
United Kingdom
Email: p.d.jameson@cranfield.ac.uk*, a.k.cooke@cranfield.ac.uk†

*Abstract*—**This work is based on adapting suitable techniques in real-time aircraft system identification for the use in rigid body dynamic modelling of Unmanned Aerial Vehicles (UAVs). In general due to their reduced size UAVs lend themselves well to the recovery of flight data in full scale atmospheric trials. Timely recovery of the true model parameters is necessary to minimise flight tests for rapid prototype development. Using the Cranfield University Jetstream-31 Flying Laboratory (G-NFLA) as a test bed an investigation into suitable parameter update methods applicable for UAV purposes is being performed; and builds on previous work which considered the constraints pertinent to small UAVs, such as the absence of air flow vanes.This paper outlines an approach to achieve post-manoeuvre parameter estimation by applying the equation-error method in the frequency domain, and an example using flight data for the Jestream-31 flying laboratory is presented.**

## I. Introduction

### A. Background

With the evolution of aircraft design and manufacture, flight testing still remains an integral aspect of the process; however, commercial demands require that a flight campaign achieve its goals in the minimal time possible. Advances in simulation and wind-tunnel modelling enable us to better predict aircraft dynamics which are necessary for the development of flight control laws, and provide the starting point for full-scale trials. Numerous examples of aircraft system identification (SysID) to determine suitable dynamic models for manned aircraft are widely reported [1], [2]. Despite the rapid development of UAV platforms widespread application of this technique has yet to occur in the unmanned field.

Real-time SysID was primarily driven by the need to monitor highly unstable aircraft behaviour in non-linear flight regimes, while expanding the operational flight envelope [3]. Recent development has focused on creating self-healing control systems, such as adaptive re-configurable control laws to provide robustness against airframe damage or control surface failures [4], [5]. In the case of UAVs real-time identification, would facilitate rapid prototyping especially in low-cost projects with their constrained development time. Development of SysID for a small UAV scenario could lead to flight trials focused towards dynamic model validation, with the prior verification step done using the simulation environment.

Therefore, current research in the department is primarily concerned with developing post-manoeuver estimation of the aerodynamic derivatives, and builds on the work undertaken by Carnduff [6] which outlined SysID methods for UAVs. The ability to check the estimated derivatives while the aircraft is flying would enable detection of poor data readings due to deficient excitation manoeuvres or atmospheric turbulence. Subsequently, appropriate action could then be taken while all the equipment and personnel are in place.

### B. Aircraft system identification

A brief explanation of key points relating to the process of aircraft SysID will now follow, an in-depth explanation can be found in [7]. The conventional SysID method can be seen in figure 1, the process depends on the *a priori* knowledge about the aircraft and is then followed by five distinct steps: experiment design, Data Compatibility Check (DCC), model structure determination, parameter estimation and model validation.

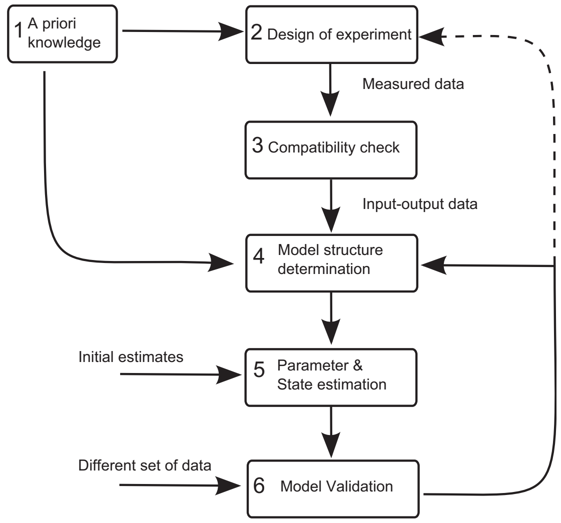


Fig. 1. An overview of the SysID process, [8].

The experiment design relates to the type of input used to perturb the aircraft from an initial state of balanced forces

Fig. 2. The Cranfield Jetstream-31 flying laboratory

and moments defined as *trim*, chosen to be a steady level flight condition for this study. The DCC step processes the data for kinematic consistency between the inertial and air data, thus allowing sensor errors or data drop-outs to be dealt with. In certain cases the model structure for the aircraft is then determined from the data, the model is then populated with the estimated model parameters and states. Finally, the representative model is validated by using a separate data set not previously used during the modelling process. Should a validation fail, alternative model structures or estimation techniques should be trialled before making the expensive decison to use a different excitation input (dotted line in figure 1). From *a priori* knowledge of the aircraft: the data was known to be compatible therefore the DCC step was neglected, and reduced order models were postulated enabeling the SysID process to proceeded directly to the Parameter Estimation (PE) step.

### C. Jetstream Aircraft

G-NFLA shown in figure 2 operates as the United Kingdom's National Flying Laboratory Aircraft and is primarily used to demonstrate flight dynamics principles to undergraduate students. As a result the types of excitation inputs used are limited to impulses and doublets. The aircraft is instrumented with an Inertial Reference System (IRS), GPS, and air data probes. Displays in-front of every seat (similar to an in-flight entertainment display) are driven by an on-board computer which transforms the measured data into strip-charts and guages depending on the topic being covered. It is anticipated that this computer will be used to run the parameter estimation algorithms; therefore an emphasis has been placed on incurring minimal additional computational burden. Current practice is to retrieve the recorded flight data from the computer hard drive for post-flight analysis using a USB key.

### D. Motivation

The aim of the current work is to develop a flexible tool box for post-manoeuver parameter estimation for UAVs. Using G-NFLA as a flying test-bed, suitable algorithms have been tested with postulated models for the reduced order SPPO, and Dutch Roll modes. As flight test data is primarily available on an opportunistic basis, the ability to reduce processing time would maximise such opportunities.

## II. Method

### A. Models

From the *a priori* knowledge about the aircraft, the standard rigid body equations of motion can be used to formulate a model which accurately describes the aircraft dynamics. The assumption when using such models is that the longitudinal and lateral dynamics are suitably distinct in order to be decoupled. The following body axes models are taken from [9], where the SPPO model is:

$$\begin{bmatrix} \dot{\boldsymbol{W}} \\ \dot{\boldsymbol{q}} \end{bmatrix} = \begin{bmatrix} Z_w & Z_q \\ M_w & M_q \end{bmatrix} \begin{bmatrix} \boldsymbol{W} \\ \boldsymbol{q} \end{bmatrix} + \begin{bmatrix} Z_\eta \\ M_\eta \end{bmatrix} [\boldsymbol{\eta}] \tag{1}$$

and the Dutch Roll model is:

$$\begin{bmatrix} \dot{\boldsymbol{V}} \\ \dot{\boldsymbol{r}} \end{bmatrix} = \begin{bmatrix} Y_v & Y_r \\ N_v & N_r \end{bmatrix} \begin{bmatrix} \boldsymbol{V} \\ \boldsymbol{r} \end{bmatrix} + \begin{bmatrix} Y_\zeta \\ N_\zeta \end{bmatrix} [\boldsymbol{\zeta}] \tag{2}$$

Here it must be noted that the above stability and control derivatives are in concise form, and therefore require manipulation in order to relate them to the respective non-dimensional values (see appendix of [9]).

### B. Body Translational Rates

Typically, large aircraft can be readily equipped to measure air data using $\alpha$- or $\beta$-vanes; these measurements enable the body translational rates: $u_b$, $v_b$ and $w_b$ to be determined and form part of the dependent variables shown in equations 1 and 2. In smaller UAVs access to such air data measurements can prove unfeasible. In the absence of these vanes, measurements from an IRS can be used to calculate the translational rates (see [10], [11]). These rates can be determined using 'track-fixed' accelerations; neglecting turbulence the IRS has more reliable sensors with respect to the analog vane type instruments.

### C. Frequency Domain Transformation

With reference to Klein [12] and more recently Morelli [13] the Equation-Error method in the frequency domain can be formulated. Analysis in the frequency domain using the finite Fourier Transform method is better suited to the linear model identification being considered for two key points. Firstly, when the Fourier transform is applied the bias and drift in the measured data is removed. Secondly, using *a priori* knowledge of the expected frequency range the data to be analysed can be easily reduced. Further to the above benefits, differentiation and convolution in the frequency domain simplifies to multiplication by $j\omega$, allowing terms such as $\dot{p}$, $\dot{q}$, $\dot{r}$ and $\dot{u}_b$, $\dot{v}_b$, $\dot{w}_b$ to be calculated. Considering the aircraft model in state space form:

$$\dot{x}(t) = \mathbf{A}x(t) + \mathbf{B}u(t), \quad x(0) = 0 \tag{3}$$

$$y(t) = \mathbf{C}x(t) + \mathbf{D}u(t) \tag{4}$$

$$z(t) = y(t) + \nu(t) \tag{5}$$

where *x(t)*, *u(t)*, and *y(t)* are the state, input and output vectors, *z(t)* is the measurement equation with error $\nu(t)$. Unknowns within **A**, **B**, **C**, and **D** are assigned to the unknown parameter matrix, $\boldsymbol{\theta}$.

Equation 6 is the Fourier transform associated with *x(t)* for the finite time interval [0, *T*], a simple Euler approximation can be implemented resulting in equation 7.

$$\begin{aligned} \tilde{x}(\omega) &= \int_0^T x(t)e^{-j\omega}dt & (6) \\ \tilde{x}(\omega) &\approx \Delta t \sum_{i=0}^{N-1} x(i)e^{-j\omega\Delta t} & (7) \end{aligned}$$

where the complex number $j = \sqrt{-1}$, $\omega =$ is the angular frequency, *i* is the discrete time index, $\Delta t$ is the sampling interval, and $N$ + 1 is the total number of data points. The summation on the right in equation 7 is known as the Discrete Fourier Transform (DFT), $\tilde{X}(\omega)$:

$$\tilde{X}(\omega) = \sum_{i=0}^{N-1} x(i)e^{-j\omega\Delta t} \qquad (8)$$

thus equation 7 can be written as:

$$\tilde{x}(\omega) = \tilde{X}(\omega)\Delta t \qquad (9)$$

As equation 9 is effectively a first order Euler approximation [14] of equation 6 corrections such as those outlined by Morelli [15] can be made to account for the inaccuracies. However, by selecting a sampling rate much greater than the frequencies of interest (a small $\Delta t$) these corrections can be suitably neglected. Therefore, the Fourier transform for the linear model given in equations 3, and 4 can be expressed:

$$\begin{aligned} j\boldsymbol{\omega}\tilde{\boldsymbol{x}}(\boldsymbol{\omega}) &= \mathbf{A}\tilde{\boldsymbol{x}}(\boldsymbol{\omega}) + \mathbf{B}\tilde{\boldsymbol{u}}(\boldsymbol{\omega}) & (10) \\ \tilde{\boldsymbol{y}}(\boldsymbol{\omega}) &= \mathbf{C}\tilde{\boldsymbol{x}}(\boldsymbol{\omega}) + \mathbf{D}\tilde{\boldsymbol{u}}(\boldsymbol{\omega}) & (11) \end{aligned}$$

Provided that the state, outputs and inputs variables are measured the individual state or output (equation 10 and 11 respectively) can be formulated using the equation error method to estimate the stability and control derivatives in $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$, and $\boldsymbol{D}$.

### D. Equation Error

It can be seen that the least squares cost function in the frequency domain for the $k^{th}$ state vector of equation 10 is:

$$J_k = \frac{1}{2}\sum_{n-1}^{m} \left| j\omega_n \tilde{x}_k(n) - \boldsymbol{A}_k\tilde{\boldsymbol{x}}(n) - \boldsymbol{B}_k\tilde{\boldsymbol{u}}(n) \right|^2 \qquad (12)$$

where $\tilde{\boldsymbol{x}}(n)$ and $\tilde{\boldsymbol{u}}(n)$ denote the Fourier transform of the state and control vectors for frequency $\omega_n$. Equation 11 can also be formulated as above for the relavent output equations. As rigid body aircraft dynamics typically lie below 2Hz, the *m* frequencies of interest in the summation of equation 12 were performed at 0.04Hz intervals from 0.01 - 2Hz, [16]. By neglecting the zero frequency which corresponds to the *trim* and measurement biases, the respective initial conditions no longer need to be estimated for the cost function. Now grouping the unknowns from $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ into $\boldsymbol{\theta}$, the standard least squares formulation for the complex data is:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\epsilon} \qquad (13)$$

expanding the vectors:

$$\begin{aligned} \boldsymbol{Y}^T &\equiv [j\omega_1\tilde{x}_k(1), j\omega_2\tilde{x}_k(2), \ldots, j\omega_m\tilde{x}_k(m)] & (14) \\ \boldsymbol{X}^T &\equiv [\tilde{\boldsymbol{x}}(1)\ \tilde{\boldsymbol{u}}(1), \tilde{\boldsymbol{x}}(2)\ \tilde{\boldsymbol{u}}(2), \ldots, \tilde{\boldsymbol{x}}(m)\ \tilde{\boldsymbol{u}}(m)] & (15) \end{aligned}$$

and $\boldsymbol{\epsilon}$ is the complex equation error in the frequency domain; the cost function to be minimised is identical to that in equation 12 where "†" denotes the complex conjugate:

$$\begin{aligned} J &= \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X})^\dagger(\boldsymbol{Y} - \boldsymbol{X}) & (16) \\ \hat{\boldsymbol{\theta}} &= \left[Re(\boldsymbol{X}^\dagger\boldsymbol{X})\right]^{-1} Re(\boldsymbol{X}^\dagger\boldsymbol{Y}) & (17) \end{aligned}$$

and the estimated parameter covariance matrix is:

$$cov(\hat{\boldsymbol{\theta}}) \equiv E\left\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T\right\} = \boldsymbol{\sigma}^2\left[Re(\boldsymbol{X}^\dagger\boldsymbol{X})\right]^{-1} \quad (18)$$

the equation error variance, $\boldsymbol{\sigma}^2$ is estimated from the residuals:

$$\boldsymbol{\sigma}^2 = \frac{1}{(m - n_p)}\left[(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})^\dagger(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})\right] \qquad (19)$$

where $n_p$ is the number of parameter elements in $\boldsymbol{\theta}$, and the parameter standard errors can be calculated by using equation 19 and then square rooting the diagonal elements of matrix $cov(\hat{\boldsymbol{\theta}})$ in equation 18.

### E. Real-time parameter estimation in frequency domain

Current real-time methods include: Recursive Least Squares (RLS), Extended Kalman Filter (EKF), and batch estimation methods (seen as sequential least squares) [17]. Identification in the cases of fault-detection and reconfigurable control requires immediate results (use RLS and EKF). In comparison for dynamic modelling, a near real-time capability is acceptable (use batch estimation method). Batch methods use strips of data at defined time intervals to approximate the time variation in the parameters and enable principally off-line methods such as the least squares technique to be used in a real-time setting. In order to gain the benefits of working in the frequency domain, a recursive finite Fourier transform (RFT) can be coupled with the sequential least squares.

With reference to the discrete Fourier transform (equation 8) the RFT for a specific frequency of interest $\boldsymbol{\omega}$, at sample time $i\Delta t$ can be related to the result at sample time ($i$ - 1)$\Delta t$ as follows:

$$X_i(\boldsymbol{\omega}) = X_{i-1}(\boldsymbol{\omega}) + x_i e^{-j\boldsymbol{\omega}i\Delta t} \qquad (20)$$

with

$$e^{-j\boldsymbol{\omega}\Delta t} = e^{-j\boldsymbol{\omega}t}e^{-j\boldsymbol{\omega}(i-1)\Delta t} \qquad (21)$$

Note that $e^{-j\boldsymbol{\omega}\Delta t}$ is constant for a given frequency $\omega$ and sampling interval of $\Delta t$. Therefore, equations 20 and 21 enable the data to be transformed for a given frequency with the use of one addition and two multiplications respectively. The time-domain data can be discarded as the RFT behaves as a memory, as the recursion proceeds results for each new sample are added to the overall information held in the constant $e^{-j\boldsymbol{\omega}\Delta t}$ term.

In order to improve the response of the PE to the most recent conditions a selective amnesia, $\lambda$, can be applied to the RFT as in equation 22 to remove past data. The effect of varying $\lambda$ from 0.95 - 1 can be seen as a type of windowing on the data, this is illustrated in figure 3, when $\lambda$ = 1, this results in the case of the general RFT where each data point is given equal weighting.

$$X_i(\boldsymbol{\omega}) = \lambda X_{i-1}(\boldsymbol{\omega}) + x_i e^{-j\boldsymbol{\omega} i\Delta t} \tag{22}$$

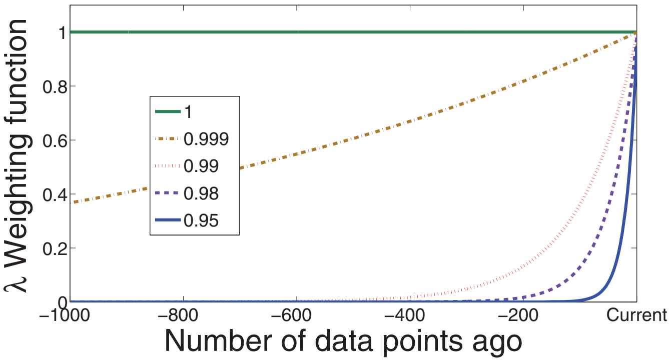


Fig. 3. Effective window of forgetting factor

For the implementation, data is recorded at 50 Hz and saved in csv format onboard the aircraft. Using Matlab the relevant state measurements for the desired mode are selected and then passed through the RFT of equations 22 and 21 at 1Hz. Subsequently the least-squares estimation is performed at 25Hz using equations 17 - 19. In order to prevent ill-conditioning when implementing equation 17 the first estimation uses 2 seconds of data to allow sufficient information to be gathered.

## III. Analysis

The flight trials were conducted at: 6045*ft* and 170*kt* (1843*m*, 87.5*m*/*s*) with a center of gravity at 30 % of the Mean Aerodynamic Chord. Two recordings were used, the first set for parameter estimation and the second set as a means of validation. Using *a priori* knowledge of the aircraft empirical estimates[1] of the concise derivatives were calculated for use as reference values. Tests were conducted with the forgetting factor varied from 0.95 - 1, the full results for $\lambda$ = 1 and an indicative set of results for $\lambda$ = 0.98 are presented below.

### A. Longitudinal

Trace plots for the reduced order longitudinal dynamics are shown in Figures 4 and 5, where the empirical estimates are plotted alongside the varying parameter estimates. The longitudinal dynamics were excited by closely coupling two elevator ($\eta$) impulses in a positive and negative direction, these are shown in the bottom subplot of each figure. The empirical (Emp) and final estimated (Est) values are listed in table I respectively; the Est values correspond to the finally estimated value at the end of each parameter trace plot and are indicated by a dot. The estimation results for the moment derivatives when a forgetting factor is used are presented in figure 6. Finally, figure 10 shows the prediction response of both sets of data from table I for the pitch rate response for the validation data set at the same flight condition.

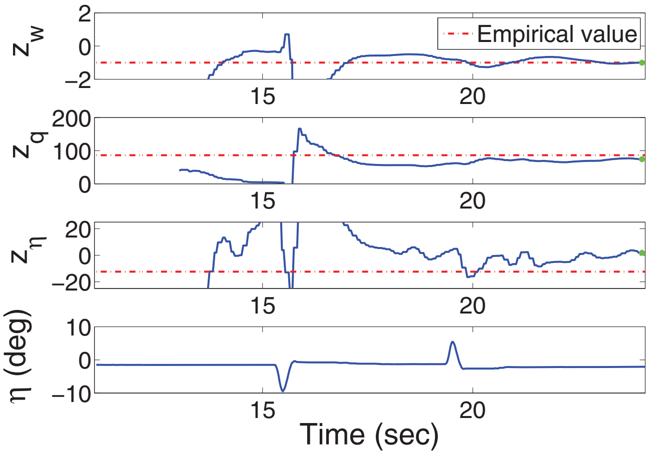


Fig. 4. SPPO: force derivatives

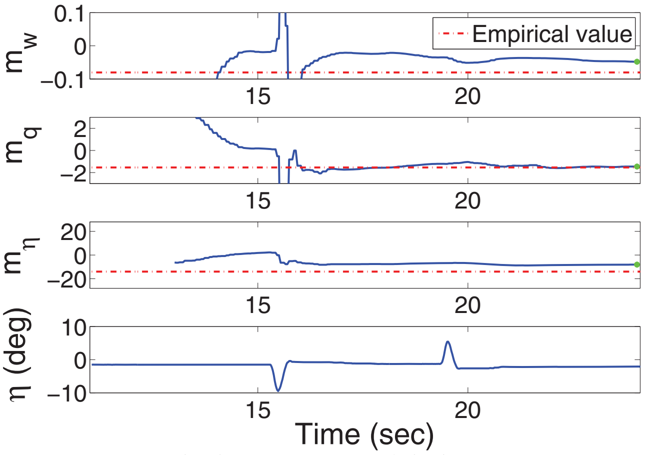


Fig. 5. SPPO: moment derivatives

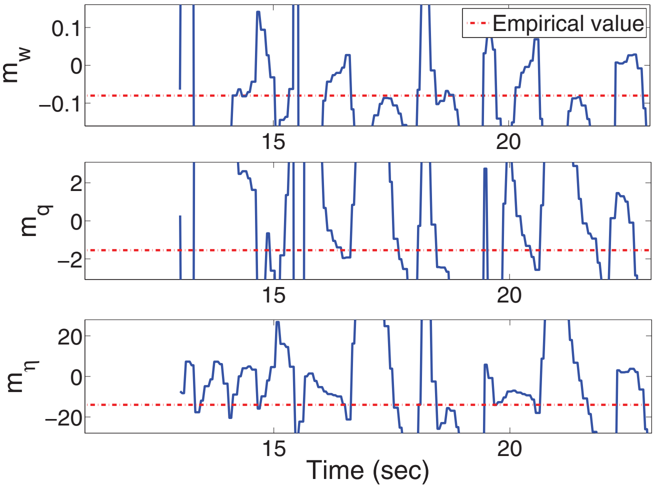


Fig. 6. SPPO: moment derivatives with $\lambda$ = 0.98

### B. Lateral

The lateral dynamic results are presented in a similar manner as above. In order to excite the lateral dynamics the yaw-damper needs to be disabled, then the pilot drives the mode by applying rudder ($\zeta$) doublets on the pedals as shown in the final subplot of figures 7 and 8.

[1] Using ESDU, a statistically significant data set and engineering judgment.

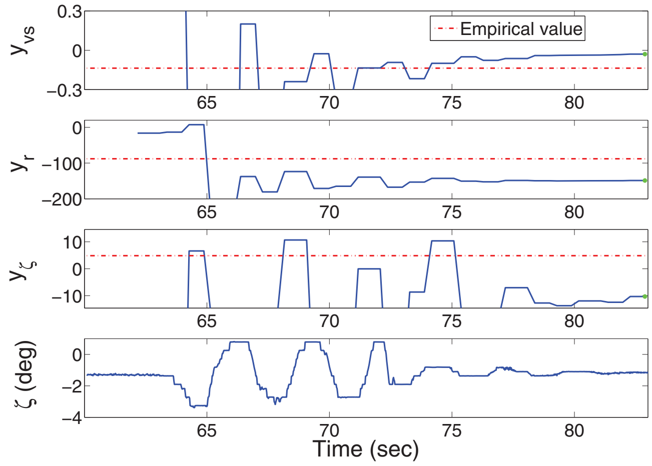


Fig. 7. DR: force derivatives

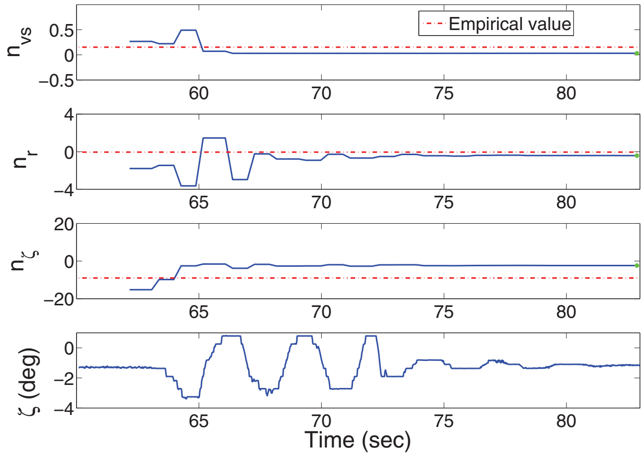


Fig. 8. DR: moment derivatives

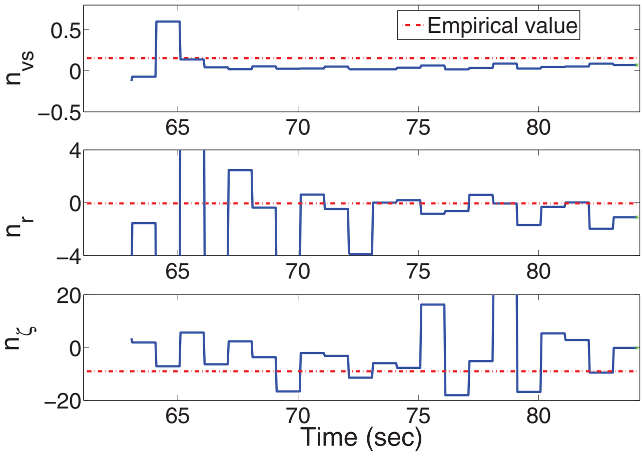


Fig. 9. DR: moment derivatives with $\lambda$ = 0.98

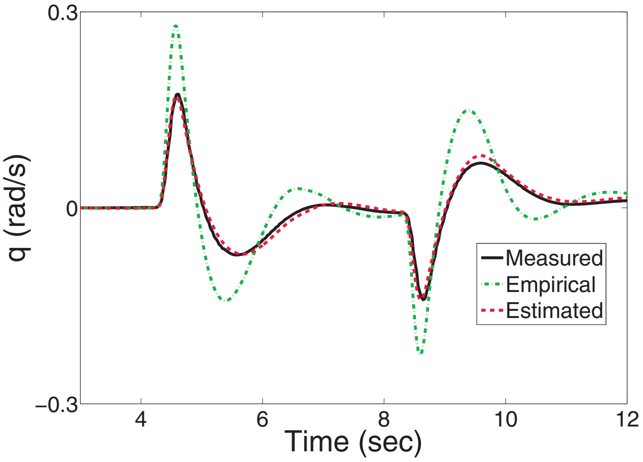


Fig. 10. SPPO: pitch rate validation plot

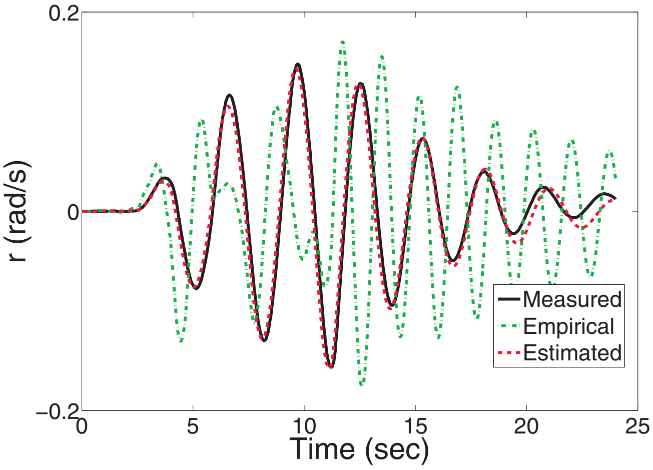


Fig. 11. DR: yaw rate validation plot

## IV. Discussion

Both sets of longitudinal derivatives established constant estimate values by the end of the 25 second data segment. In the case of the force derivatives (figure 4) all three parameters showed an improvement after the initial elevator deflection, and a small variability following the second deflection. Following the first input the moment derivatives (figure 5) were close to their steady state values with little variation following the second deflection. By contrast, the derivative estimates using the forgetting factor (figure 6), $\lambda$ = 0.98 showed no signs of convergence.

The pitch rate validation plot figure 10 demonstrates that the basic longitudinal dynamics have been captured with the reduced order model, and that the empirical estimates produce an under damped response as could be expected due to the informed approximations used. Referring to table I the majority of the empirical and estimated derivatives compare well. The exception is $z_\eta$ which has an incorrect sign and a reduced magnitude, however this is also highlighted as having the largest standard error of $\pm 19.599$. Nevertheless, it is important to note that the pitch damping term, $m_q$ (which is a key term for the SPPO mode) closely matches the empirical estimate and has a low standard error, $\pm 0.554$.

For the lateral case the force derivatives in figure 7 displayed a gradual improvement as the excitation manoeuvre progressed. In comparison the moment derivatives converged more rapidly to their respective final values (figure 8). Again

TABLE I
SPPO DERIVATIVES AND STANDARD ERRORS

| $\hat{\theta}$ | Emp | Est | $s$ |
|---|---|---|---|
| $z_w$ | -1.002 | -1.005 | 0.263 |
| $z_q$ | 86.170 | 74.362 | 13.125 |
| $z_\eta$ | -12.377 | 1.868 | 19.599 |
| $m_w$ | -0.080 | -0.048 | 0.011 |
| $m_q$ | -1.547 | -1.459 | 0.554 |
| $m_\eta$ | -14.002 | -8.078 | 0.826 |

TABLE II
DR DERIVATIVES AND STANDARD ERRORS

| $\hat{\theta}$ | Emp | Est | $s$ |
|---|---|---|---|
| $y_v$ | -0.137 | -0.029 | 0.051 |
| $y_r$ | -88.004 | -148.73 | 4.829 |
| $y_\zeta$ | 4.865 | -10.308 | 18.140 |
| $n_v$ | 0.153 | 0.031 | 0.001 |
| $n_r$ | -0.049 | -0.415 | 0.057 |
| $n_\zeta$ | -8.980 | -2.389 | 0.213 |

for the forgetting factor case, the moment derivatives are seen in figure 9 to exhibit a greater variation. For the validation data set the RFT estimated parameters produced a satisfactory yaw rate prediction in figure 11, with the empirical estimates resulting in an undamped response. In addition, a slightly under damped response can also be seen for the estimated derivatives after 20 seconds, this can possibly be attributed to unmodeled roll effects as the reduced order DR model is predominantly associated with yaw, but by definition DR is a coupled yaw-roll mode.

Comparing the DR derivatives in table II the majority of derivatives do not agree. Considering the two derivatives with the greatest standard errors ($s$), firstly, the estimated side force due to yaw rate derivative, $y_r$ is $\approx 1.7$ times greater than the empirical value. Secondly, the estimated side force due to rudder, $y_\zeta$ is of a different sign and order of magnitude in comparison to the empirical value. However, referring to the validation plot the DR mode dynamics would seem to have been captured, and therefore further investigation needs to be undertaken to explain the discrepancies that has been found.

Having analysed results for both the SPPO and DR modes the combined least squares in frequency domain RFT method is a promising solution for the online estimation problem. In the present work incorporating a forgetting factor with the RFT was found to be unsuitable particularly with the limited aircraft excitation manoeuvers available, as the removal of information content proved to be too sensitive. Finally, the ability to restrict the frequency range provides several benefits, firstly it significantly reduces the number of computations for the RFT and secondly, it acts as a filter to reduce the effects of the higher aero-elastic frequencies present (associated with the flexible airframe).

## V. CONCLUSION AND FUTURE WORK

A suitable methodology to address on-line estimation for a small UAV in the frequency domain has been outlined, in addition the results of trials using the Cranfield Jetstream-31 have been presented. Areas for further development lie in automatically stopping the batch estimation once the parameter estimates have converged to within a specified accuracy.

Furthermore, there exists the opportunity to enhance the way in which the flight data results are recorded, as the UAV flight envelope is opened. Results for runs in repeat conditions can be incorporated to the data set by taking advantage of the constant $e^{-j\omega\delta t}$ in the RFT and therefore continuing the recursion.

## ACKNOWLEDGEMENTS

This research was jointly funded as a CASE award by EPSRC and BAE Systems, and with the help of the Eric Beverley Bursary from The Worshipful Company of Coachmakers and Coach Harness Makers of London. The first author would like to thank Mr Chris Fielding, Dr Stephen Carnduff, and Dr Eugene Morelli for their time and advice.

## REFERENCES

[1] P. G. Hamel and R. V. Jategaonkar, “Evolution of flight vehicle system identification,” *Journal of Aircraft*, vol. 33 (1), pp. 9–28, January - February 1996.

[2] K. W. Iliff, “Parameter estimation for flight vehicles,” *Journal of Guidance, Control, and Dynamics*, vol. 12 (5), pp. 609–622, September - October 1989.

[3] K. Basappa and R. Jategaonkar, “Evaluation of recursive methods for aircraft parameter estimation,” in *AIAA Atmospheric Flight Mechanics Conference and Exhibit*. Providence, Rhode Island: AIAA-2004-5063, 16th-19th August 2004.

[4] D. G. Ward, J. Monaco, and M. Bodson, “Development and flight testing of a parameter identification algorithm for reconfigurable control,” *Journal of Guidance, Control and Dynamics*, vol. 21 (6), pp. 948–956, November - December 1998.

[5] G. Chowdhary, W. M. DeBusk, and E. N. Johnson, “Real-time system identification of a small multi-engine aircraft with structural damage,” in *AIAA Infotec@Aerospace*. Atlanta, Georgia: AIAA-2010-3472, 20th-22nd April 2010.

[6] S. D. Carnduff, “System identification of Unmanned Aerial Vehicles,” PhD Thesis, Cranfield University, Cranfield, Bedfordshire, 2008.

[7] V. Klein, “Estimation of aircraft aerodynamic parameters from flight data,” *Progress in Aerospace Sciences*, vol. 26 (1), pp. 1–77, 1989.

[8] ——, “A review of system identification methods applied to aircraft,” The George Washington University, Tech. Rep. Joint Institute for Acoustics and Flight Sciences Report: N83 33901, 1983.

[9] M. V. Cook, *“Flight Dynamic Principles : A linear systems approach to aircraft stability and control”*. Amsterdam: Elsevier, 2007.

[10] P.-D. Jameson and A. Cooke, “Developing system identification for UAVs,” in *25th Bristol International UAV Systems Conference*, Bristol, United Kingdom, 12th - 14th April 2010.

[11] E. A. Morelli, “Real-time aerodynamic parameteric estimation without air flow angle measurements,” in *AIAA Guidance, Navigation, and Control Conference and Exhibit*. Toronto, Ontario: AIAA-2010-7951, 2nd-5th August 2010.

[12] V. Klein, “Aircraft parameter estimation in frequency domain,” in *AIAA Atmospheric Flight Mechanics Conference and Exhibit*. Palo Alto, California: AIAA-1978-1344, 7th - 9th August 1978.

[13] E. A. Morelli, “Real-time parameter estimation in the frequency domain,” in *AIAA Guidance, Navigation and Control Conference and Exhibit*. Portland, Oregon: AIAA-99-4043, 9th-11th August 1999.

[14] R. Jategaonkar, *“Flight Vehicle System Identification: A Time Domain Methodology”*. Reston, Virginia: AIAA, 2006.

[15] E. A. Morelli and V. Klein, “Accuracy of aerodynamic model parameters estimated from flight test data,” *Journal of Guidance, Control, and Dynamics*, vol. 20 (1), pp. 74–80, January - February 1997.

[16] E. A. Morelli, “Flight-test experiment design for characterizing stability and control of hypersonics vehicles,” *Journal of Guidance, Control and Dynamics*, vol. 32 (3), pp. 949–959, May-June 2009.

[17] V. Klein and E. A. Morelli, *“Aircraft System Identification: Theory and Practice”*. Reston, Virginia: AIAA, 2006.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Determination of Dynamic Flexure Model Parameters for Ship Angular Deformation Measurement

Wei Wu[a,b], Sheng Chen[b,c] and Shiqiao Qin[d]

[a]School of Opto-Electronic Science and Engineering, National University of Defense Technology,
Changsha 410073, China

[b]Electronics and Computer Science, Faculty of Physical and Applied Sciences,
University of Southampton, Southampton SO17 1BJ, UK
E-mails: ww6g11@ecs.soton.ac.uk sqc@ecs.soton.ac.uk

[c]Faculty of Engineering, King Abdulaziz University, Jeddah 21589, Saudi Arabia

[d]School of Science, National University of Defense Technology, Changsha 410073, China
E-mail: sqqin8@nudt.edu.cn

*Abstract*— **In ship angular deformation measurement, Kalman filter used to estimate the deformation angle requires accurate dynamic flexure parameters. Traditionally, these dynamic flexure parameters are empirically set according to previous experience or determined from previously collected experimental data. Inevitably, the Kalman filter will perform poorly when the current application environment is differ with those used in the filter design. To overcome this problem, we propose an alternative online approach to estimate the dynamic flexure parameters based on the attitude difference measured by two laser gyro units. Specifically, the Tufts-Kumaresan (T-K) method is introduced to solve the unknown parameters of the dynamic flexure model from the computed attitude difference. Simulation results show that the proposed method can estimate the dynamic flexure parameters in real-time with a high degree of accuracy even in serious noise polluted conditions. A further advantage of the proposed approach is that it does not require *a priori* knowledge of the dynamic flexure characteristics.**

## I. Introduction

Ship angular deformation refers to the angular displacement existed between the shipboard sensors, such as the radar, optoelectronic detectors or missile coordinate frames, with the reference system of the master inertial navigation system (MINS) coordinate frame. The ship angular deformation consists of the static deformation component and the dynamic flexure, according to its time characteristics [1]. The static deformation is a time-invariant angular displacement due to installation errors. By contrast, the dynamic flexure is the structure flexure caused by the sea wave or wind induced loads, and it behaves like a random process. For high accurate tactical shipboard weapons or sensors, it is required to measure and compensate the angular deformation with respect to the MINS [2], [3]. Extensive works have focused on the ship angular deformation measurement problem in the past decades [4]–[7]. In particular, the inertial ship angular deformation measurement method is recognised as an efficient means to solve this problem, and it has attracted great attention in the recent years [8], [9].

The common procedure in inertial ship angular deformation measurement methods is to successively compute the attitude difference measured by two laser gyro units (LGUs) and to resolve the deformation angle through a real-time Kalman filter. In this procedure, the second-order Gauss-Markov process representing the dynamic flexure characteristics is adopted in Kalman filter design. Since the Kalman filter acts like an observer, the measurement accuracy is closely related to the accuracy of dynamic flexure model parameters, such as magnitude, frequency and damping ratio, employed in Kalman filters [6]. In general, the more accurate the dynamic flexure model parameters used, the more accurate the final deformation measurement is achieved [10]. Two existing approaches are usually adopted to determine the unknown parameters of the second-order Gauss-Markov process based dynamic flexure model: empirical method and statistical method. In an empirical method, as discussed in the works of [4], [8], the unknown parameters are simplified as the functions of the correlation time and variance of the dynamic flexure signal. By contrast, in a statistical method, as adopted by the reference [11], the unknown parameters are determined from the previously recorded dynamic flexure measurement data based on statistical estimation algorithms. Obviously, these two existing approaches do not meet the on-line requirements and operating environments, since the exact ship structure and actual work conditions will generally differ from those used in estimating the dynamic flexure model parameters.

It is highly advisable to adapt the dynamic flexure model parameters on-line so that the dynamic flexure parameters employed by the real-time Kalman filter match the specific working condition and environment. To achieve this goal, we propose an on-line parameter estimation method by utilising the attitude difference measured by two LGUs in order to estimate the dynamic flexure model parameters more accurately. More specifically, we assume that the dynamic flexure angles can be depicted by using the second-order Markov process, and the Tufts-Kumaresan (T-K) method [12] is then applied to solve the unknown parameters from the correlation function of the LGUs' attitude difference. Our simulation results obtained demonstrate that the parameters identified

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

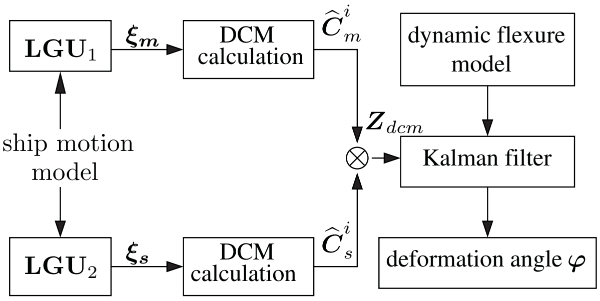

LGU$_1$
$\boldsymbol{\xi}_{\boldsymbol{m}}$
DCM calculation
$\widehat{\boldsymbol{C}}_m^i$
dynamic flexure model
$\boldsymbol{Z}_{dcm}$
ship motion model
Kalman filter
LGU$_2$
$\boldsymbol{\xi}_{\boldsymbol{s}}$
DCM calculation
$\widehat{\boldsymbol{C}}_s^i$
deformation angle $\varphi$

Fig. 1. Schematic diagram of ship angular deformation measurement.

using the T-K method have high accuracy. The proposed parameter estimation method requires no *a priori* knowledge of the dynamic flexure characteristics and it works well even under low signal to noise ratio (SNR) conditions.

## II. Ship Angular Deformation Measurement

Fig. 1 shows the schematic diagram of the ship angular deformation measurement system based on the attitude matching method [13]. Two LGUs, LGU$_1$ and LGU$_2$, are installed adjacent to the MINS and the shipboard sensor, respectively. Assume that LGU$_1$'s coordinates have been aligned with the the MINS frame ($m$-frame), while LGU$_2$'s coordinates have been aligned with sensor's frame ($s$-frame). As illustrated in Fig. 1, LGU$_1$ measures the Euler angle of the ship body motion $\boldsymbol{\xi}_m\big(\xi_{mx}, \xi_{my}, \xi_{mz}\big)$, and $\widehat{\boldsymbol{C}}_m^i$ is the corresponding direction cosine matrix (DCM) rotation from the $m$-frame to the inertial frame ($i$-frame), which is expressed in Eq. (1) at the bottom of this page. Similarly, $\boldsymbol{\xi}_s\big(\xi_{sx}, \xi_{sy}, \xi_{sz}\big)$ is the Euler angle measured by LGU$_2$, and $\widehat{\boldsymbol{C}}_s^i$ denotes the corresponding DCM rotation from the $s$-frame to the $i$-frame.

### A. Measurement Model

The measurement model is required to calculate the DCMs of $\widehat{\boldsymbol{C}}_m^i$ and $\widehat{\boldsymbol{C}}_s^i$ based on the angular information measured by LGU$_1$ and LGU$_2$. Let the calculated DCMs of $\widehat{\boldsymbol{C}}_m^i$ and $\widehat{\boldsymbol{C}}_s^i$ be

$$\widehat{\boldsymbol{C}}_m^i = \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{bmatrix}, \widehat{\boldsymbol{C}}_s^i = \begin{bmatrix} C'_{11} & C'_{12} & C'_{13} \\ C'_{21} & C'_{22} & C'_{23} \\ C'_{31} & C'_{32} & C'_{33} \end{bmatrix}, \quad (2)$$

where $C_{ij}$ and $C'_{ij}$ denote the element of $\widehat{\boldsymbol{C}}_m^i$ and $\widehat{\boldsymbol{C}}_s^i$ at the $i$th row and $j$th column, respectively.

Then the attitude matching function for the ship angular deformation measurement is given by [13]

$$\boldsymbol{Z}_{dcm} = \boldsymbol{B}\boldsymbol{\varphi} - \boldsymbol{A}\boldsymbol{\phi}_0 + \boldsymbol{B}\big(\widehat{\boldsymbol{C}}_i^m \boldsymbol{\Psi}_m - \widehat{\boldsymbol{C}}_i^s \boldsymbol{\Psi}_s\big), \quad (3)$$

where $\boldsymbol{\varphi}$ is the total deformation angle between the LGU$_1$ and LGU$_2$, which includes a static component $\boldsymbol{\phi}_0$ and a dynamic component $\boldsymbol{\theta}$, and has the relation of $\boldsymbol{\varphi} = \boldsymbol{\phi}_0 + \boldsymbol{\theta}$. The measurement vector $\boldsymbol{Z}_{dcm}$ given in Eq. (3) is expressed by

$$\boldsymbol{Z}_{dcm} = \begin{bmatrix} C_{13}C'_{12} + C_{23}C'_{22} + C_{33}C'_{32} \\ C_{13}C'_{11} + C_{23}C'_{21} + C_{33}C'_{31} \\ C_{11}C'_{12} + C_{21}C'_{22} + C_{31}C'_{32} \end{bmatrix}, \quad (4)$$

and the coefficient matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are given by

$$\boldsymbol{A} = \begin{bmatrix} C_{33}C'_{22} - C_{23}C'_{32} & C_{13}C'_{32} - C_{33}C'_{12} & C_{23}C'_{12} - C_{13}C'_{22} \\ C_{33}C'_{21} - C_{23}C'_{31} & C_{13}C'_{31} - C_{33}C'_{11} & C_{23}C'_{11} - C_{13}C'_{21} \\ C_{31}C'_{22} - C_{21}C'_{32} & C_{11}C'_{32} - C_{31}C'_{12} & C_{21}C'_{12} - C_{11}C'_{22} \end{bmatrix} \text{ and } \boldsymbol{B} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}, \quad (5)$$

respectively, while $\boldsymbol{\Psi}_m$ anf $\boldsymbol{\Psi}_s$ are the attitude measurement errors induced by the gyro bias $\boldsymbol{\varepsilon}_0$ and the random walk noise $\boldsymbol{\varepsilon}_r$ in the LGU$_1$ and LGU$_2$, respectively, which can be determined by the following differential equations

$$\begin{cases} \dot{\boldsymbol{\Psi}}_m = -\widehat{\boldsymbol{C}}_m^i\big(\boldsymbol{\varepsilon}_{m0} + \boldsymbol{\varepsilon}_{mr}\big), \\ \dot{\boldsymbol{\Psi}}_s = -\widehat{\boldsymbol{C}}_s^i\big(\boldsymbol{\varepsilon}_{s0} + \boldsymbol{\varepsilon}_{sr}\big), \end{cases} \quad (6)$$

in which $\boldsymbol{\varepsilon}_{m0}$ and $\boldsymbol{\varepsilon}_{mr}$ denote the gyro bias and random walk noise of LGU$_1$, while $\boldsymbol{\varepsilon}_{s0}$ and $\boldsymbol{\varepsilon}_{sr}$ are the gyro bias and random walk noise of LGU$_2$.

As can be seen from Eq. (3), the measurement vector $\boldsymbol{Z}_{dcm}$ is linearised with respect to the ship angular deformation angle $\boldsymbol{\varphi}$. Furthermore, by using the dynamic flexure model in Kalman filter design, the ship angular deformation angle can be optimally estimated.

### B. Kalman Filter Design

It is well known that the Kalman filter offers the optimal result only when the system noise is white. However, in Eq. (3), the dynamic flexure angle is caused by the ship motion and wave loads vibration, whose frequency is closely correlated with the ship angular motion. If we use a white noise to depict the dynamic flexure in Kalman filter design, the result will be poor. The second-order Gauss-Markov process has been used extensively to model the dynamic flexure in many successful applications [4], [9], [13]. We therefore use this a second-order Gauss-Markov process for modelling the the dynamic flexure. The correlation function of the second-order Gauss-Markov process takes the form

$$R_{\theta_i}(\tau) = \sigma_i^2 \exp\big(-\alpha_i|\tau|\big)\big(\cos\beta_i\tau + \frac{\alpha_i}{\beta_i}\sin\beta_i|\tau|\big), \quad (7)$$

where the index $i$ indicates the $x$, $y$ or $z$ coordinate, $\tau$ is the time lag, and $\sigma_i^2$ is the variance of the $i$-coordinate component of the dynamic flexure, while $\alpha_i$ is the damping factor and $\beta_i$ is the circular frequency.

$$\widehat{\boldsymbol{C}}_m^i\big(\boldsymbol{\xi}_m\big) = \begin{bmatrix} \cos\xi_{my}\cos\xi_{mz} - \sin\xi_{mx}\sin\xi_{my}\sin\xi_{mz} & -\cos\xi_{mx}\sin\xi_{mz} & \sin\xi_{my}\cos\xi_{mz} + \sin\xi_{mx}\sin\xi_{mz}\cos\xi_{my} \\ \cos\xi_{my}\sin\xi_{mz} + \sin\xi_{mx}\sin\xi_{my}\cos\xi_{mz} & \cos\xi_{mx}\cos\xi_{mz} & \sin\xi_{my}\sin\xi_{mz} - \cos\xi_{my}\sin\xi_{mx}\cos\xi_{mz} \\ -\sin\xi_{my}\cos\xi_{mx} & \sin\xi_{mx} & \cos\xi_{mx}\cos\xi_{my} \end{bmatrix}. \quad (1)$$

The differential equation representing the ship dynamic flexure having the correlation function given in Eq. (7) can be written as

$$\ddot{\theta}_i + 2\alpha_i\dot{\theta}_i + b_i^2\theta_i = 2b_i\sigma_i\sqrt{\alpha_i}e_i(t), \tag{8}$$

where $b_i^2 = \alpha_i^2 + \beta_i^2$ is the square of the prevailing variation frequency and $e_i(t)$ is a Gaussian white noise with unit variance.

From Eqs. (3) to (8), we can derive the state function for the Kalman filter as follows

$$\dot{\boldsymbol{X}} = \boldsymbol{F}\boldsymbol{X} + \boldsymbol{w}, \tag{9}$$

where $\boldsymbol{X} \in \mathbb{R}^{21\times 1}$ is the state vector given by

$$\boldsymbol{X} = \begin{bmatrix}\boldsymbol{\phi}_0^{\mathrm{T}} & \boldsymbol{\theta}^{\mathrm{T}} & \dot{\boldsymbol{\theta}}^{\mathrm{T}} & \boldsymbol{\Psi}_m^{\mathrm{T}} & \boldsymbol{\Psi}_s^{\mathrm{T}} & \tilde{\boldsymbol{\varepsilon}}_m^{\mathrm{T}} & \tilde{\boldsymbol{\varepsilon}}_s^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}, \tag{10}$$

with $^{\mathrm{T}}$ denoting the vector or matrix transpose operator and the two gyro error vectors expressed by $\tilde{\boldsymbol{\varepsilon}}_m = \boldsymbol{\varepsilon}_{m0} + \boldsymbol{\varepsilon}_{mr}$ and $\tilde{\boldsymbol{\varepsilon}}_s = \boldsymbol{\varepsilon}_{s0} + \boldsymbol{\varepsilon}_{sr}$, respectively, while the state transition matrix $\boldsymbol{F} \in \mathbb{R}^{21\times 21}$ is given by

$$\boldsymbol{F} = \begin{bmatrix}\boldsymbol{O}_{3\times 3} & & \\ & \boldsymbol{F}_{6\times 6}^1 & \\ & & \boldsymbol{F}_{12\times 12}^2\end{bmatrix} \tag{11}$$

with $\boldsymbol{O}_{l\times m}$ denoting the $l \times m$ zero matrix,

$$\boldsymbol{F}_{6\times 6}^1 = \begin{bmatrix} & \boldsymbol{O}_{3\times 3} & & & \boldsymbol{I}_3 & \\ -b_x^2 & 0 & 0 & -2\mu_x & 0 & 0 \\ 0 & -b_y^2 & 0 & 0 & -2\mu_y & 0 \\ 0 & 0 & -b_z^2 & 0 & 0 & -2\mu_z \end{bmatrix} \tag{12}$$

and

$$\boldsymbol{F}_{12\times 12}^2 = \begin{bmatrix}\boldsymbol{O}_{3\times 6} & -\widehat{\boldsymbol{C}}_m^i & \boldsymbol{O}_{3\times 3} \\ \boldsymbol{O}_{3\times 6} & \boldsymbol{O}_{3\times 3} & -\widehat{\boldsymbol{C}}_s^i \\ & \boldsymbol{O}_{6\times 12} & \end{bmatrix}, \tag{13}$$

in which $\boldsymbol{I}_3$ is the $3 \times 3$ identity matrix. The system noise vector $\boldsymbol{w} \in \mathbb{R}^{21\times 1}$ has the covariance matrix

$$\begin{aligned} E\left[\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\right] = \mathrm{diag}\Big\{&\boldsymbol{O}_{1\times 3}, 4b_x^2\sigma_x^2\alpha_x, 4b_y^2\sigma_y^2\alpha_y, 4b_z^2\sigma_z^2\alpha_z, \\ &\boldsymbol{O}_{1\times 9}, \left(\boldsymbol{\sigma}_{mr}^2\right)^{\mathrm{T}}, \left(\boldsymbol{\sigma}_{sr}^2\right)^{\mathrm{T}}\Big\}, \end{aligned} \tag{14}$$

where $E[\bullet]$ denotes the expectation operator and diag$\{\bullet\}$ the diagonal matrix, while $\boldsymbol{\sigma}_{mr}^2 \in \mathbb{R}^{3\times 1}$ whose elements are the three variances of the LGU$_1$'s gyro random walk noise vector $\boldsymbol{\varepsilon}_{mr}$ and $\boldsymbol{\sigma}_{sr}^2 \in \mathbb{R}^{3\times 1}$ contains the three variances of the LGU$_2$'s gyro random walk noise vector $\boldsymbol{\varepsilon}_{sr}$, respectively.

According to Eq. (3), the observation function is given by

$$\boldsymbol{Z}_{dcm} = \boldsymbol{H}\boldsymbol{X} + \boldsymbol{v}, \tag{15}$$

where $\boldsymbol{Z}_{dcm} \in \mathbb{R}^{3\times 1}$ is the observation vector, $\boldsymbol{v} \in \mathbb{R}^{3\times 1}$ is the measurement noise vector, and $\boldsymbol{H} \in \mathbb{R}^{3\times 21}$ is the observation matrix, which can be written as

$$\boldsymbol{H} = \begin{bmatrix}\boldsymbol{B} - \boldsymbol{A} & \boldsymbol{B} & \boldsymbol{O}_{3\times 3} & \boldsymbol{B}\widehat{\boldsymbol{C}}_i^m & -\boldsymbol{B}\widehat{\boldsymbol{C}}_i^s & \boldsymbol{O}_{3\times 6}\end{bmatrix}. \tag{16}$$

With the aid of the dynamic flexure model, the Kalman filter can accurately estimate the total deformation angle between the LGU$_1$ and LGU$_2$ frames. Let $\boldsymbol{\varphi}$ be the true deformation angle between LGU$_1$ and LGU$_2$, and $\widehat{\boldsymbol{\varphi}}$ be its estimate provided by the Kalman filter. The deformation estimate accuracy is determined by the error vector $\Delta\boldsymbol{\varphi}$ according to

$$\left[\Delta\boldsymbol{\varphi}\right]_{\mathrm{s-s}} = \boldsymbol{I}_3 - \boldsymbol{C}(\boldsymbol{\varphi})\boldsymbol{C}^{\mathrm{T}}(\widehat{\boldsymbol{\varphi}}), \tag{17}$$

where $\left[\Delta\boldsymbol{\varphi}\right]_{\mathrm{s-s}} \in \mathbb{R}^{3\times 3}$ denotes the skew-symmetric matrix of $\Delta\boldsymbol{\varphi}$ . From the above analysis, we can see that the deformation measuring accuracy depends on the accuracy of the dynamic flexure model parameters, $\alpha_i$, $\beta_i$ and $\sigma_i$, used in Eq. (12). Next, we present an on-line parameters estimation method based on the observation function of Eq. (3).

## III. Estimation of Dynamic Flexure Parameters

### A. Parameters Estimation Functions

Noting the relation $\boldsymbol{\varphi} = \boldsymbol{\phi}_0 + \boldsymbol{\theta}$, Eq. (3) can be rewritten as

$$\boldsymbol{Z}_{dcm} = \boldsymbol{Z}_{\phi_0} + \boldsymbol{Z}_{\theta} + \boldsymbol{Z}_{\psi}, \tag{18}$$

in which

$$\boldsymbol{Z}_{\phi_0} = (\boldsymbol{B} - \boldsymbol{A})\boldsymbol{\phi}_0, \tag{19}$$

$$\boldsymbol{Z}_{\theta} = \boldsymbol{B}\boldsymbol{\theta}, \tag{20}$$

$$\boldsymbol{Z}_{\psi} = \boldsymbol{B}\big(\widehat{\boldsymbol{C}}_i^m\boldsymbol{\Psi}_m - \widehat{\boldsymbol{C}}_i^s\boldsymbol{\Psi}_s\big). \tag{21}$$

The validity of Eq. (18) rests on the assumption that $\boldsymbol{\varphi}$ is small, which is generally true in reality.

The static component $\boldsymbol{\phi}_0$ in practice is compensated to within several milliradians using the course estimate results. In addition, $\big(\boldsymbol{B} - \boldsymbol{A}\big) = \boldsymbol{I}_3 - \widehat{\boldsymbol{C}}_i^m\widehat{\boldsymbol{C}}_s^i$ is very small. Taken into account these conditions, $\boldsymbol{Z}_{\phi_0}$ in Eq. (18) may be removed.

The attitude error term $\boldsymbol{Z}_{\psi}$ is induced by the gyro errors, including gyro bias and gyro random walk noise. According to the reference [14], the typical frequency of dynamic flexure ranges from 0.1 Hz to 0.25 Hz, while the frequency of the gyro noise induced attitude error is less than 0.01 Hz [15], which is far lower than the dynamic flexure frequency. Therefore, we can remove $\boldsymbol{Z}_{\psi}$ from Eq. (18) via a High-pass filter. The filtering process can be expressed by

$$\tilde{\boldsymbol{Z}}_{dcm} \approx \boldsymbol{F}^{-1}\big[\boldsymbol{H}_h(\omega)\boldsymbol{Z}_{dcm}(\omega)\big] \approx \boldsymbol{B}\boldsymbol{\theta}, \tag{22}$$

where $\boldsymbol{F}^{-1}[\bullet]$ denotes the inverse Fourier transform, $\boldsymbol{H}_h(\omega)$ is the transfer function of the High-pass filter, and $\boldsymbol{Z}_{dcm}(\omega)$ is the frequency-domain transformation of $\boldsymbol{Z}_{dcm}$. The correlation function of $\tilde{\boldsymbol{Z}}_{dcm}$ is given by

$$\begin{aligned} \boldsymbol{R}_Z(\tau) &= \langle\tilde{\boldsymbol{Z}}_{dcm}(t), \tilde{\boldsymbol{Z}}_{dcm}(t+\tau)\rangle \\ &= \langle\boldsymbol{B}\boldsymbol{\theta}(t), \boldsymbol{B}\boldsymbol{\theta}(t+\tau)\rangle = \langle\boldsymbol{\theta}(t), \boldsymbol{\theta}(t+\tau)\rangle, \end{aligned} \tag{23}$$

with $\langle\boldsymbol{\theta}(t), \boldsymbol{\theta}(t+\tau)\rangle$ denoting the correlation function of $\boldsymbol{\theta}(t)$.

Eq. (23) establishes an approximated relation of the dynamic flexure with the attitude measurement difference. By Comparing Eq. (23) with Eq. (7), the parameters of the dynamic flexure model, $\alpha_i$, $\beta_i$ and $\sigma_i$, can be determined.

### *B. Tufts-Kumaresan Method*

We apply the T-K method [12] to estimate the unknown parameters in Eq. (7) based on the measurement values of Eq. (23). The T-K method is a common choice to resolve closely spaced sinusoids, particularly when the data length is short and the SNR value is small. The T-K algorithm [12] is briefly summarised below.

Give the $N$ samples of a sequence $y(n)$, which consists of a sum of the $M$ exponentially damped sinusoidal signals

$$y(n) = \sum_{l=1}^{M} a_l \exp(s_l n) + q_n,\ n = 1, 2, \cdots, N, \tag{24}$$

where $a_l$ is the complex amplitude of the damped mode $\exp(s_l) = \exp\big(-\alpha_l + \mathrm{j}\beta_l\big)$ with $\alpha_l > 0$, and $q_n$ is the unknown white noise with variance $\sigma_q^2$. Using the complex conjugate data to set up the backward prediction function

$$\boldsymbol{A}\boldsymbol{b} = \boldsymbol{h}, \tag{25}$$

with

$$\boldsymbol{A} = \begin{bmatrix} y^*(2) & y^*(3) & \cdots & y^*(L+1) \\ y^*(3) & y^*(4) & \cdots & y^*(L+2) \\ \vdots & \vdots & \vdots & \vdots \\ y^*(N-L+1) & y^*(N-L+2) & \cdots & y^*(N) \end{bmatrix}, \tag{26}$$

$$\boldsymbol{b} = \begin{bmatrix} b_1 & b_2 & \cdots & b_L \end{bmatrix}^{\mathrm{T}}, \tag{27}$$

$$\boldsymbol{h} = \begin{bmatrix} y^*(1) & y^*(2) & \cdots & y^*(N-L) \end{bmatrix}^{\mathrm{T}}. \tag{28}$$

The prediction error filter polynomial

$$B(z) = 1 + b_1 z^{-1} + b_2 z^{-2} + \cdots + b_L z^{-L}, \tag{29}$$

has the zeros at $\exp\big(-s_l^*\big)$, $1 \le l \le M$, if $L$ is chosen to satisfy the inequality $M \leqslant L \leqslant N - M$. The roots of the polynomial of Eq. (29) yields the set of $M$ zeros, from which the $M$ damped modes $\exp\big(s_l\big)$, $1 \le l \le M$, can be determined. Then the amplitudes $a_l$ can easily be estimated based on the data set $\{y(n)\}$ described by Eq. (24) according to the least squares principle.

By substituting the discrete correlation results of Eq. (23) into Eq. (24), the parameters of the dynamic flexure model, $\alpha_i$, $\beta_i$ and $\sigma_i$, can be directly estimated by applying the T-K algorithm. Note that, the solution of Eq. (24) for real signal will give pairs of roots, $\exp\big(-s_l^*\big)$ and $\exp\big(-s_l\big)$, $1 \le l \le M$, and we simply use one root from each pair to calculate the dynamic flexure parameters.

## IV. Simulation Results

The schematic diagram of the gyro signal sample generation, parameter estimation and ship deformation angle estimation is shown in Fig. 2.

### *A. Simulation System Setup*

According to our experimental experience, the ship angular motion can be simplified as a random process, which can also be depicted as a second-order Gauss-Markov process

$$R_{\xi_i}(\tau) = \sigma_{\xi_i}^2 \exp\big(-\alpha_{\xi_i}|\tau|\big)\Big(\cos\beta_{\xi_i}\tau + \frac{\alpha_{\xi_i}}{\beta_{\xi_i}}\sin\beta_{\xi_i}|\tau|\Big), \tag{30}$$

where $i$ again denotes the $x$, $y$ or $z$ coordinate, while $\sigma_{\xi_i}^2$, $\alpha_{\xi_i}$ and $\beta_{\xi_i}$ are the variance, damping factor and circular frequency, respectively, of the $i$ coordinate of the ship attitude angle. We assume that during the simulation the parameters $\sigma_{\xi_i}^2$, $\alpha_{\xi_i}$ and $\beta_{\xi_i}$ are time invariant. Table I lists the simulation parameters of the ship attitude angles used. The values of $\beta_{\xi_i}$ and $\alpha_{\xi_i}$ are taken from our previous real-data identification results, while the values of $\sigma_{\xi_i}$ are set according to our experimental experience.

TABLE I

Ship attitude parameters used in the simulation system.

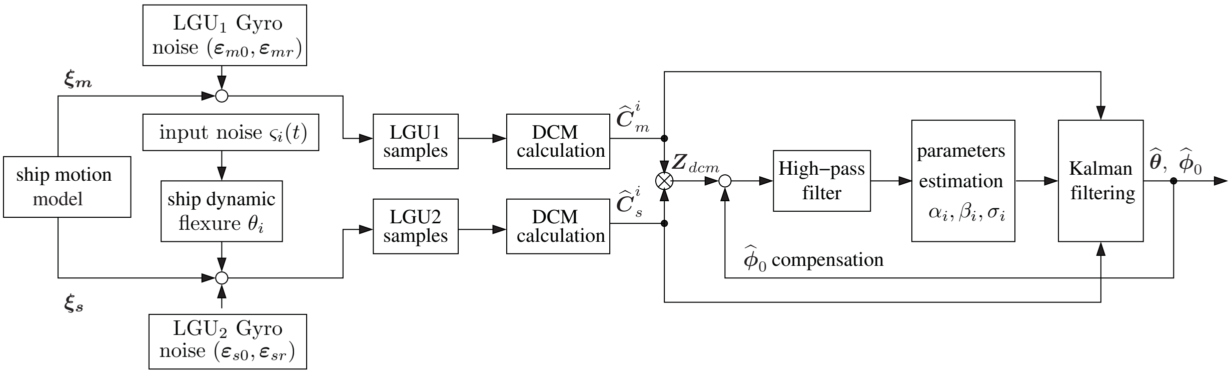
| | Magnitude<br>$\sigma_{\xi_i}$ (deg) | Frequency<br>$\beta_{\xi_i}/2\pi$ (Hz) | Damping factor<br>$\alpha_{\xi_i}$ (s$^{-1}$) |
|---|---|---|---|
| Pitch | 2.20 | 0.18 | 0.10 |
| Roll | 3.40 | 0.07 | 0.06 |
| Yaw | 0.80 | 0.05 | 0.12 |

TABLE II

Dynamic flexure parameters used in the simulation system.

| | Magnitude<br>$\sigma_i$ (mrad) | Frequency<br>$\beta_i/2\pi$ (Hz) | Damping factor<br>$\alpha_i$ (s$^{-1}$) |
|---|---|---|---|
| Pitch | 0.40 | 0.19 | 0.13 |
| Roll | 0.68 | 0.17 | 0.11 |
| Yaw | 0.50 | 0.18 | 0.10 |

The dynamic flexure angle is simulated by using three independent second-order Markov process, and we assume that



LGU$_1$ Gyro noise $(\boldsymbol{\varepsilon}_{m0}, \boldsymbol{\varepsilon}_{mr})$
$\boldsymbol{\xi}_m$
input noise $\varsigma_i(t)$
ship motion model
ship dynamic flexure $\theta_i$
$\boldsymbol{\xi}_s$
LGU$_2$ Gyro noise $(\boldsymbol{\varepsilon}_{s0}, \boldsymbol{\varepsilon}_{sr})$
LGU1 samples
LGU2 samples
DCM calculation
DCM calculation
$\widehat{\boldsymbol{C}}_m^i$
$\widehat{\boldsymbol{C}}_s^i$
$\boldsymbol{Z}_{dcm}$
High–pass filter
parameters estimation $\alpha_i, \beta_i, \sigma_i$
Kalman filtering
$\widehat{\boldsymbol{\theta}}$, $\widehat{\phi}_0$
$\widehat{\phi}_0$ compensation

Fig. 2. Schematic diagram of gyro signal sample generation and dynamic flexure parameters estimation.

the variances $\sigma_i$, damping factors $\alpha_i$ and frequencies $\beta_i$ used to simulate the dynamic flexure are time invariant, whose values are list in Table II. The damping factors and frequencies are identified from our real measurement data in sea trials, while the values of the variances are set according to our empirical experience to reflect certain level of real sea condition. Additionally, in order to reflect the real measurement environment, we add a Gaussian white noise $\varsigma_i(t)$ of variance $\sigma^2_{\varsigma_i}$ to the dynamic flexure signal $\theta_i$. The SNR of the dynamic flexure signal is then defined by

$$\mathrm{SNR}_i = 10 \log_{10} \frac{\sigma_i^2}{\sigma^2_{\varsigma_i}}, \tag{31}$$

where again $i$ indicates the $x$, $y$ or $z$ coordinate. We assume that the static deformation angle after the initial compensation is $[3.5 \text{ mrad } 3.5 \text{ mrad } 3.5 \text{ mrad}]$.

The parameters of the bias error vector $\boldsymbol{\varepsilon}_0$ and the random walk noise vector $\boldsymbol{\varepsilon}_r$ for the gyros are list in Table III. The simulated LGU$_1$ and LGU$_1$ gyro outputs are processed to derive the attitude values of $\widehat{\boldsymbol{C}}^i_m$ and $\widehat{\boldsymbol{C}}^i_s$, which are then used to estimate the dynamic flexure parameters as well as used by a Kalman filter to compute the deformation angle $\boldsymbol{\varphi}$.

TABLE III

PARAMETERS OF THE GYRO NOISES USED IN THE SIMULATION SYSTEM, WHERE WN STD DENOTES THE STANDARD DEVIATION OF THE WHITE NOISE COMPONENT.

| | LGU$_1$ | | LGU$_2$ | |
|---|---|---|---|---|
| | Bias<br>$\boldsymbol{\varepsilon}_{m0}$<br>(deg/hr) | Random walk<br>$\boldsymbol{\varepsilon}_{mr}$'s WN STD<br>$(\deg/\sqrt{\text{hr}})$ | Bias<br>$\boldsymbol{\varepsilon}_{s0}$<br>(deg/hr) | Random walk<br>$\boldsymbol{\varepsilon}_{sr}$'s WN STD<br>$(\deg/\sqrt{\text{hr}})$ |
| X | 0.005 | 0.002 | 0.02 | 0.005 |
| Y | 0.005 | 0.002 | 0.02 | 0.005 |
| Z | 0.005 | 0.002 | 0.02 | 0.005 |

## B. Results and Analysis

The total data length for the ship deformation measurement was $T = 600$ s with the sample frequency of 20 Hz. The cut-off frequency for the High-pass filter used was set to 0.05 Hz. To better compensate the approximation error induced by the static deformation angle $\boldsymbol{\phi}_0$, we ran the T-K algorithm twice in parameters estimation. Specifically, in the first run, we obtained an improved estimate $\widehat{\boldsymbol{\phi}}_0$, and then fed back this value to compensate the error term $\boldsymbol{Z}_{\phi_0}$. The second run's result was accepted as our estimate. All the results presented were obtained by averaging over 100 independent trials in the presence of the randomly generated noise for simulating the ship attitude, dynamic flexure and gyro noise signals.

TABLE IV

MEANS AND STANDARD DEVIATIONS OF THE ESTIMATED DYNAMIC FLEXURE PARAMETERS OBTAINED UNDER THE CONDITION OF $\sigma^2_{\varsigma_i} = 0$, GIVEN $T = 600$ S, $N = 20$ S, $L = 6$ S AND $M = 2$.

| | Magnitude<br>$\sigma_i$ (mrad) | Frequency<br>$\beta_i/2\pi$ (Hz) | Damping factor<br>$\alpha_i$ (s$^{-1}$) |
|---|---|---|---|
| Pitch | 0.3950 (0.0064) | 0.1892 (0.0041) | 0.1272 (0.0252) |
| Roll | 0.6722 (0.0089) | 0.1685 (0.0039) | 0.1054 (0.0188) |
| Yaw | 0.4961 (0.0075) | 0.1798 (0.0034) | 0.1013 (0.0183) |

The performance of the T-K method depends on its algorithmic parameters, $M$, $N$ and $L$ [12]. We used $M = 2$ for Eq. (25), which yielded a pair of roots for every damping mode. Appropriate values for $N$ and $L$ were determined by extensive experiments, and they were found to be $L = 6$ s or 120 samples, and $N = 20$ s or 400 samples.

Table IV lists the means and standard deviations (in bracket) of the estimation results for the parameters $\alpha_i$, $\beta_i$ and $\sigma_i$ at the condition of $\sigma^2_{\varsigma_i} = 0$. Comparing with the true values given in Table II, it can be seen that the parameter estimates are very accurate. The ship angular deformation measurement results based on the identified dynamic flexure model are shown in Table V, where it can be observed that a high accurate measurement is achieved.

TABLE V

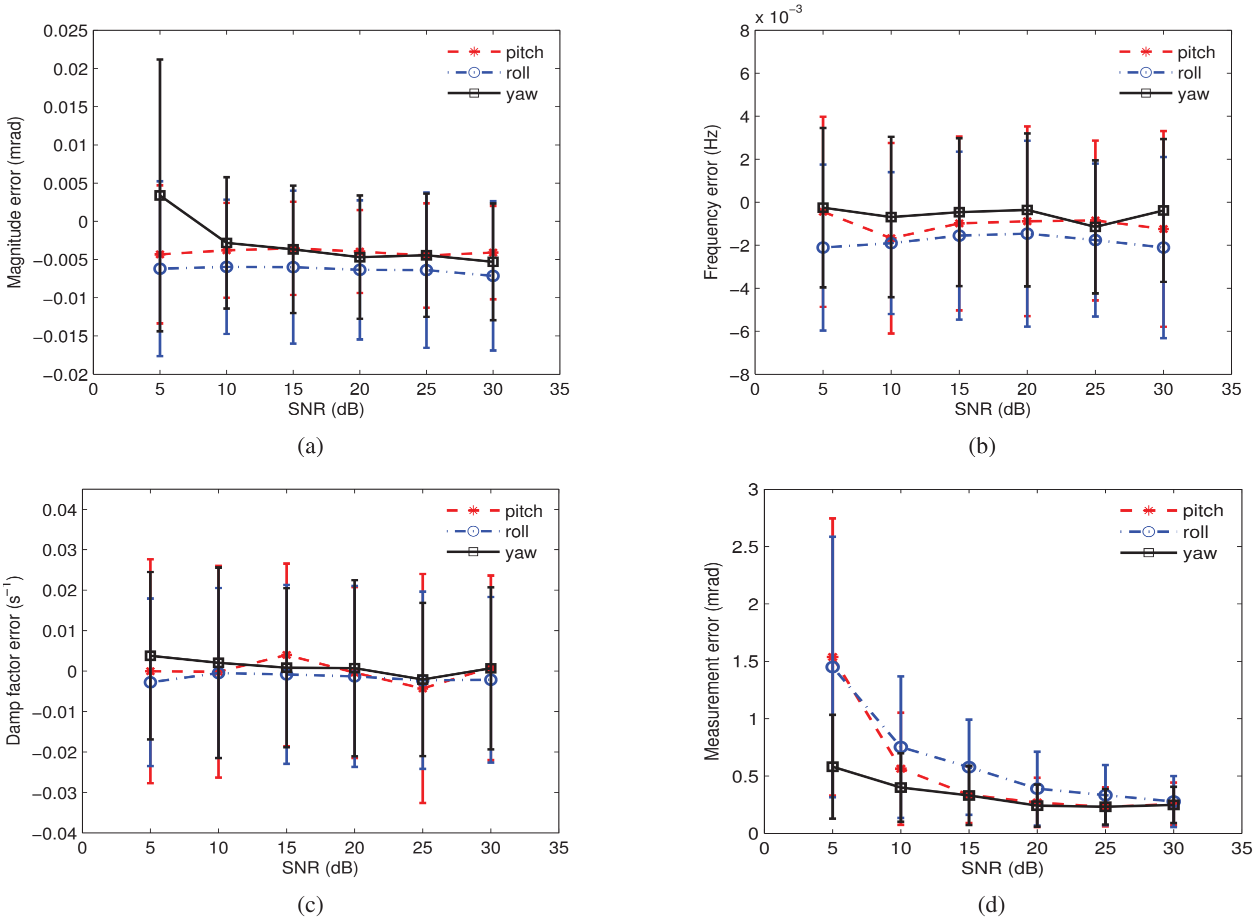
PERFORMANCE OF THE KALMAN FILTER BASED SHIP ANGULAR DEFORMATION MEASUREMENT OBTAINED BASED ON THE DYNAMIC FLEXURE MODEL IDENTIFIED UNDER THE CONDITION OF $\sigma^2_{\varsigma_i} = 0$.

| | Mean and standard deviation of true deformation angle (mrad) | Mean and standard deviation of KF estimated deformation angle (mrad) | Mean and standard deviation of KF based measurement error (mrad) |
|---|---|---|---|
| Pitch | 3.4981 (0.4267) | 3.5179 (0.5525) | 0.2626 (0.2005) |
| Roll | 3.4792 (0.7209) | 3.5131 (0.7776) | 0.3259 (0.2369) |
| Yaw | 3.5027 (0.4840) | 3.5483 (0.5626) | 0.1944 (0.1382) |

We next investigated the accuracy of the dynamic flexure parameter estimate and the Kalman filter performance under different SNR conditions. Fig. 3 shows the mean parameters estimation errors as well the average deformation measurement errors given different SNR values. As can be seen from Fig. 3 (a) to (c), the parameters, $\alpha_i$, $\beta_i$ and $\sigma_i$, can be estimated to a high degree of accuracy while the average parameter estimation errors and their error bars were similar across the range of the SNR values tested, except for the yaw magnitude error at the SNR value of 5 dB. This demonstrates the robustness of the algorithm under the noise polluted shipboard environment. However, the estimates of the magnitudes $\sigma_i$ and frequencies $\beta_i$ were slightly biased. This bias may be caused by removing the term $\boldsymbol{Z}_{\phi_0}$ as well as using high-pass filtering to remove the attitude error term $\boldsymbol{Z}_{\psi}$ in Eq. (18). By using the estimated dynamic flexure parameters, the average deformation measurement errors and their corresponding standard deviations are depicted in Fig. 3 (d). The results demonstrate that the Kalman filter is capable of achieving high accurate ship angular deformation measurement under serious noise polluted environments.

# V. CONCLUSIONS

An efficient on-line dynamic flexure parameter estimation method has been proposed for the ship angular deformation measurement system based on the attitude difference provided by two LGUs measures. The relationship between the attitude difference correlation function and that of the second-order Gauss-Markov process representing the dynamic flexure model has been presented, and the Tufts-Kumaresan method has been applied to identify the unknown dynamic flexure model



(a)

(b)

(c)

(d)

Fig. 3. Mean parameter estimation errors and average measurement error obtained under different SNR values, given $N = 20$ s, $L = 6$ s and $M = 2$: (a) error for magnitude $\sigma_i$, (b) error for frequency $\beta_i/2\pi$, (c) error for damping factor $\alpha_i$, and (d) average deformation measurement error, where vertical lines indicate the corresponding standard deviations or error bars.

parameters. From the extensive simulation results, it has been shown that the proposed model parameter estimation method is capable of obtaining accurate estimates of the unknown dynamic flexure parameters for the application of accurate ship deformation measurement. Our approach does not require *a priori* knowledge of the dynamic flexure characteristics, and it equips the ship angular deformation measurement system with the ability to adapt to various work conditions. Our future work will study how to reduce or remove the parameter estimate bias in the T-K based model parameters estimation algorithm.

## References

[1] D. L. Day and J. Arrud, “Impact of structural flexure on precision tracking,” *Naval Engineers J.*, vol. 111, no. 3, pp. 133–138, May 1999.

[2] M. G. Petovello, K. O’Keefe, G. Lachapelle, and M. E. Cannon, “Measuring aircraft carrier flexure in support of autonomous aircraft landings,” *IEEE Trans. Aerospace and Electronic Systems*, vol. 45, no. 2, pp. 523–535, April 2009.

[3] C. Bacchus, I. Barford, D. Bedford, J. Chung, P. Dailey, et al, “Digital array radar for ballistic missile defense and counter-stealth systems analysis and parameter tradeoff study,” *Report NPS-SE-06-001*, Naval Postgraduate School, Monterey, USA, Sept. 2006.

[4] A. M. Schnider, “Kalman filter formulations for transfer alignment of strapdown inertial units,” *Navigation*, vol. 30, no. 1, pp. 72–89, 1983.

[5] G. Wang, K. Pran, G. Sagvolden, G. B. Havsgard, et al, “Ship hull structure monitoring using fibre optic sensors,” *Smart Materials and Structures*, vol. 10, no. 3, pp. 472–478, June 2001.

[6] A. V. Mochalov and A. V. Kazantasev, “Use of ring laser units for measurement of moving object deformation,” in *Proc. SPIE 4680*, Feb. 5, 2002, pp. 85–92.

[7] Q. Yu, G. Jiang, S. Fu, Z. Chao, Y. Shang, and X. Sun, “Fold-ray videometrics method for the deformation measurement of nonintervisible large structures,” *Applied Optics*, vol. 48, no. 24, pp. 4683–4687, Aug. 2009.

[8] F. Sun, C. J. Guo, W. Gao, and B. Li, “A new inertial measurement method of ship dynamic deformation,” in *Proc. 2007 Int. Conf. Mechatronics and Automation* (Harbin, China), Aug. 5-8, 2007, pp. 3407–3412.

[9] L. Joon and Y.-C. Lim, “Transfer alignment considering measurement time delay and ship body flexure,” *J. Mechanical Science and Technology*, vol. 23, no. 1, pp. 195–203, 2009.

[10] P. D. Groves, “Optimising the transfer alignment of weapon INS,” *J. Navigation*, vol. 56, no. 2, pp. 323–335, May 2003.

[11] J. E. Kain and J. R. Cloutier, “Rapid transfer alignment for tactical weapon applications,” in *Proc. AIAA Conf. Guidance, Navigation and Control* (Boston, USA), Aug. 14-16, 1989, pp. 1290–1300.

[12] R. Kumaresan and D. Tufts, “Estimating the parameters of exponentially damped sinusoids and pole-zero modeling in noise,” *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 30, no. 6, pp. 833–840, Dec. 1982.

[13] J.-X. Zheng, S.-Q. Qin, X.-S. Wang, and Z.-S. Huang, “Attitude matching method for ship deformation measurement,” *J. Chinese Inertial Technology*, vol. 18, no. 2, pp. 175–180, 2010.

[14] P. G. Shoals and D. E. Brunner, “Dynamic ship flexure measurement program,” *Report A047040*, Naval Ship Weapon Systems Engineering Station Port Hueneme, CA, Aug. 24, 1973.

[15] J.-X. Zheng, S.-Q. Qin, X.-S. Wang, and Z.-S. Huang, “Influences of Gyro biases on ship angular flexure measurement,” in *Proc. 2011 Symp. Photonics and Optoelectronics* (Wuhan, China), May 16-18, 2011, pp. 1–4.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Singularly Impulsive Dynamical Systems with time delay: Mathematical Model and Stability

Nataša A. Kablar, Vlada Kvrgić,
*Lola Institute, Kneza Viseslava 70a, Belgrade 11000, Serbia*
nkablar.ae01@gtalumni.org, vlada.kvrgic@li.rs

**Abstract** - *In this paper we introduce new class of system, so called singularly impulsive or generalized impulsive dynamical systems with time delay. Dynamics of this system is characterized by the set of differential and difference equations with time delay, and algebraic equations. They represent the class of hybrid systems, where algebraic equations represent constraints that differential and difference equations with time delay need to satisfy. In this paper we present model, assumptions on the model, two classes of singularly impulsive dynamical systems with delay - time dependenet and state dependent. Further, we present Lyapunov and asymptotic stability theorems for nonlinear time-dependent and state-dependent singularly impulsive dynamical systems with time delay.*

## I. Introduction

Modern complex engineering systems as well as biological and physiological systems typically possess a multi-echelon hierarchical hybrid architecture characterized by continuous-time dynamics at the lower levels of hierarchy and discrete-time dynamics at the higher levels of the hierarchy. Hence, it is not surprising that hybrid systems have been the subject of intensive research over the past recent years (see Branicky et al. (1998), Ye et al. (1998 b), Haddad, Chellaboina and Kablar (2001a-b)). Such systems include dynamical switching systems Branicky (1998), Leonessa et al. (2000), nonsmooth impact and constrained mechanical systems, Back et al. (1993), Brogliato (1996), Brogliato et al. (1997), biological systems Lakshmikantham et al. (1989), demographic systems Liu (1994), sampled-data systems Hagiwara and Araki (1988), discrete-event systems Passino et al. (1994), intelligent vehicle/highway systems Lygeros et al. (1998) and flight control systems, etc. The mathematical descriptions of many of these systems can be characterized by impulsive differential equations, Simeonov and Bainov (1985), Liu (1988), Lakshmikantham et al. (1989, 1994), Bainov and Simeonov (1989, 1995), Kulev and Bainov (1989), Lakshmikantham and Liu (1989), Hu et al. (1989), Samoilenko and Perestyuk (1995), Haddad, Chellaboina and Kablar (2001a-b). Impulsive dynamical systems can be viewed as a subclass of hybrid systems.

Motivated by the results on impulsive dynamical systems presented in Haddad, Chellaboina, and Kablar (2001, 2005), the authors previous work on singular or generalized systems, and results on singularly impulsive dynamical systems published in Kablar(2003, 2010) we presented new class of *singularly impulsive* or *generalized impulsive dynamical systems with time delay*. It presents novel class of hybrid systems and generalization of impulsive dynamical systems to incorporate singular nature of the systems and time delays. Extensive applications of this class of systems can be found in contact problems and in hybrid systems.

We present mathematical model of the singularly impulsive dynamical systems with time delay. We show how it can be viewed as general systems from which impulsive dynamical systems with time delay, singular continuous-time systems with time delay and singular dicrete-time systems with time delay, as well as without time delay,follow. Then we present Assumptions needed for the model and the division of this class of systems to time-dependent and state-dependent singularly impulsive dynamical systems with time delay with respect to the resetting set. Finally, we draw some conclusions and define future work.

In this paper for the class of nonlinear singularly impulsive dynamical systems with time delay we develop Lyapunov and asymptotic stability results. Results are further specialized to linear case. Note that for addressing the stability of the zero solution of a singularly impulsive dynamical system the usual stability definitions are valid. Then we draw some conclusions and define future work.

At first, we establish definitions and notations. Let $\mathbb{R}$ denote the set of real numbers, let $\mathbb{R}^n$ denote the set of $n \times 1$ real column vectors, let $\mathcal{N}$ denote the set of nonnegative integers, and let $I_n$ or $I$ denote the $n \times n$ identity matrix. Furthermore, let $\partial\mathcal{S}, \dot{\mathcal{S}}, \bar{\mathcal{S}}$ denote the boundary, the interior, and a closure of the subset $\mathcal{S} \subset \mathbb{R}^n$, respectively. Finally, let $\mathrm{C}^0$ denote the set of continuous functions and $\mathrm{C}^{\mathrm{r}}$ denote the set of functions with $r$ continuous derivatives.

## II. Mathematical Model of Singularly Impulsive Dynamical Systems with Time Delay

A singularly impulsive dynamical system with delay consists of three elements:

1. A possibly singular continuous-time dynamical equation with time delay, which governs the motion of the system between resetting events;
2. A possibly singular difference equation with time delay, which governs the way the states are instantaneously changed when a resetting occurs; and

3. A criterion for determining when the states of the system are to be reset.

Mathematical model of these systems is described with

$$E_{\rm c}\dot{x}(t) = f_{\rm c}(x(t,\tau)) + G_{\rm c}(x(t,\tau))u_{\rm c}(t), \quad (t, x(t,\tau), u_{\rm c}(t)) \notin \mathcal{S}, \tag{II.1}$$

$$E_{\rm d}\triangle x(t) = f_{\rm d}(x(t,\tau)) + G_{\rm d}(x(t,\tau))u_{\rm d}(t), \quad (t, x(t,\tau), u_{\rm c}(t)) \in \mathcal{S}, \tag{II.2}$$

$$y_{\rm c}(t) = h_{\rm c}(x(t,\tau)) + J_{\rm c}(x(t,\tau))u_{\rm c}(t), \quad (t, x(t,\tau), u_{\rm c}(t)) \notin \mathcal{S}, \tag{II.3}$$

$$y_d(t) = h_{\rm d}(x(t,\tau) + J_{\rm d}(x(t,\tau))u_{\rm d}(t), \quad (t, x(t,\tau), u_{\rm c}(t)) \in \mathcal{S}, \tag{II.4}$$

where $t \geq 0$, $\tau > 0$, $x(0) = x_0$, $x(t,\tau) \in \mathcal{D} \subset \mathbb{R}^{\ltimes} \times \mathbb{N}$, $\mathcal{D}$ is an open set with $0 \in \mathcal{D}$, $u_{\rm c} \in \mathcal{U}_{\rm c} \subset \mathbb{R}^{\gtrdot_{\rm c}}$, $u_{\rm d}(t_k) \in \mathcal{U}_{\rm d} \subset \mathbb{R}^{\gtrdot_{\rm d}}$, $t_k$ denotes $k^{\rm th}$ instant of time at which $(t, x(t,\tau), u_{\rm c}(t))$ intersects $\mathcal{S}$ for a particular trajectory $x(t,\tau)$ and input $u_{\rm c}(t), y_{\rm c}(t) \in \mathbb{R}^{\lessdot_{\rm c}}, y_{\rm d}(t_k) \in \mathbb{R}^{\lessdot_{\rm d}}$, $f_{\rm c} : \mathcal{D} \to \mathbb{R}^{\ltimes}$ is Lipschitz continuous and satisfies $f_{\rm c}(0) = 0$, $G_{\rm c} : \mathcal{D} \to \ltimes \times \gtrdot_{\rm c}$, $f_{\rm d} : \mathcal{D} \to \mathbb{R}^{\ltimes}$ is continuous and satisfies $f_{\rm d}(0) = 0$, $G_{\rm d} : \mathcal{D} \to \mathbb{R}^{n \times m_{\rm d}}$, $h_{\rm c} : \mathcal{D} \to \mathbb{R}^{l_{\rm c}}$ and satisfies $h_{\rm c}(0) = 0$, $J_{\rm c} : \mathcal{D} \to \mathbb{R}^{l_{\rm c} \times m_{\rm c}}$, $h_{\rm d} : \mathcal{D} \to \mathbb{R}^{l_{\rm d}}$ and satisfies $h_{\rm d}(0) = 0$, $J_{\rm d} : \mathcal{D} \to \mathbb{R}^{l_{\rm d} \times m_{\rm d}}$, and $\mathcal{S} \subset [0,\infty) \times \mathbb{R}^n \times \mathcal{U}_{\rm c}$ is the *resetting set*. Here, as in Haddad, Chellaboina, and Kablar (2001a) we assume that $u_{\rm c}(\cdot)$ and $u_{\rm d}(\cdot)$ are restricted to the class of *admissible* inputs consisting of measurable functions $(u_{\rm c}(t), u_{\rm d}(t)) \in \mathcal{U}_{\rm c} \times \mathcal{U}_{\rm d}$ for all $t \geq 0$ and $k \in \mathcal{N}_{[0,t)} \equiv k : 0 \leq t_k < t$, where the constraint set $\mathcal{U}_{\rm c} \times \mathcal{U}_{\rm d}$ is given with $(0,0) \in \mathcal{U}_{\rm c} \times \mathcal{U}_{\rm d}$. We refer to the differential equation (II.1) as the *continuous-time dynamics with time delay*, and we refer to the difference equation (II.2) as the *resetting law*.

Matrices $E_{\rm c}$, $E_{\rm d}$ may be singular matrices. In case $E_{\rm c} = I$, $E_{\rm d} = I$, and $\tau = 0$ (II.1)–(II.4) represent standard impulsive dynamical systems described in Haddad, Chellaboina, and Kablar (2001a), and Haddad, Kablar, and Chellaboina (2000, 2005), where stability, dissipativity, feedback interconnections, optimality, robustness, and disturbance rejection has been analyzed. In absence of discrete dynamics they specialize to singular continuous-time systems, with further specialization $E_{\rm c} = I$ to standard continuous-time systems. If only discrete dynamics is present they specialize to singular discrete-time systems, with further specialization $E_{\rm d} = I$ to standard discrete-time systems.

In case $E_{\rm c} = I$, $E_{\rm d} = I$, and $\tau \neq 0$, (II.1)–(II.4) represent standard impulsive dynamical systems with time delay. In absence of discrete dynamics they specialize to singular continuous-time systems with time delay, with further specialization $E_{\rm c} = I$ to standard continuous-time systems with time delay. If only discrete dynamics is present they specialize to singular discrete-time systems with time delay, with further specialization $E_{\rm d} = I$ to standard discrete-time systems with time delay.

Therefore, theory of the singularly impulsive or generalized impulsive dynamical systems with time delay once developed, can be viewed as a generalization of the singular and impulsive dynamical system with time delay theory, unifying them into more general new system theory.

In what follows is given basic setting and division of this class of systems with respect to the definition of the resetting sets, accompanied with adequate assumptions needed for the model.

We make the following additional assumptions:

A1. $(0, x_0, u_{\rm c0}) \notin \mathcal{S}$, where $x(0) = x_0$ and $u_{\rm c}(0) = u_{\rm c0}$, that is, the initial condition is not in $\mathcal{S}$.

A2. If $(t, x(t,\tau), u_{\rm c}(t)) \in \bar{\mathcal{S}} \backslash \mathcal{S}$ then there exists $\epsilon > 0$ such that, for all $0 < \delta < \epsilon$, $s(t+\delta; t, x(t,\tau), u_{\rm c}(t+\delta)) \notin \mathcal{S}$.

A3. If $(t_k, x(t_k), u_{\rm c}(t_k)) \in \partial S \cap \mathcal{S}$ then there exists $\epsilon > 0$ such that, for all $0 < \delta < \epsilon$ and $u_{\rm d}(t_k) \in \mathcal{U}_{\rm d}$, $s(t_k + \delta; t_k, E_{\rm d}x(t_k) + f_{\rm d}(x(t_k)) + G_{\rm d}(x(t_k))u_{\rm d}(t_k), u_{\rm c}(t_k+\delta)) \notin \mathcal{S}$.

A4. We assume consistent initial conditions (and prior and after every resetting).

Assumption A1 ensures that the initial condition for the resetting differential equation (II.1), (II.2) is not a point of discontinuity, and this assumption is made for convenience. If $(0, x_0, u_{\rm c0}) \in \mathcal{S}$, then the system initially resets to $E_{\rm d}x_0^+ = E_{\rm d}x_0 + f_{\rm d}(x_0) + G_{\rm d}(x_0)u_{\rm d}(0)$ which serves as the initial condition for the continuous dynamics (II.1). It follows from A3 that the trajectory then leaves $\mathcal{S}$. We assume in A2 that if a trajectory reaches the closure of $\mathcal{S}$ at a point that does not belong to $\mathcal{S}$, then the trajectory must be directed away from $\mathcal{S}$, that is, a trajectory cannot enter $\mathcal{S}$ through a point that belongs to the closure of $\mathcal{S}$ but not to $\mathcal{S}$. Finally, A3 ensures that when a trajectory intersects the resetting set $\mathcal{S}$, it instantaneously exits $\mathcal{S}$, see Figure 1. We make the following remarks.

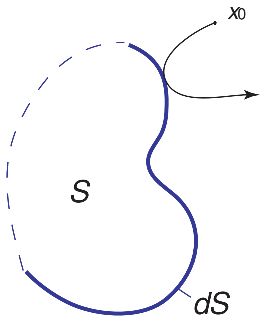


**Figure 1.** Resetting Set.

**Remark II.1.** *It follows from A3 that resetting removes the pair $(t_k, x_k, u_{\rm c}(t_k))$ from the resetting set $\mathcal{S}$. Thus, immediately after resetting occurs, the continuous-time dynamics (II.1), and not the resetting law (II.2), becomes the active element of the singularly impulsive dynamical system.*

**Remark II.2.** *It follows from A1-A3 that no trajectory can intersect the interior of $\mathcal{S}$. According to A1, the trajectory $x(t)$ begins outside the set $\mathcal{S}$. Furthermore, it follows from A2 that a trajectory can only reach $\mathcal{S}$ through a point belonging to both $\mathcal{S}$ and its boundary. Finally, from A3, it follows that if a trajectory reaches a point $\mathcal{S}$ that is on the boundary of $\mathcal{S}$, then the trajectory is instantaneously removed from $\mathcal{S}$. Since* *a continuous trajectory starting outside of $\mathcal{S}$ and intersecting the interior of $\mathcal{S}$ must first intersect the boundary of $\mathcal{S}$, it follows that no trajectory can reach the interior of $\mathcal{S}$.*

**Remark II.3.** *It follows from A1-A3 and Remark 1.2 that $\partial\mathcal{S} \cup \mathcal{S}$ is closed and hence the resetting times $t_k$ are well defined and distinct.*

**Remark II.4.** *Since the resetting times are well defined and distinct, and since the solutions to (II.1) exist and are unique, it follows that the solutions of the singularly impulsive dynamical system (II.1), (II.2) also exist and are unique over a forward time interval.*

In Haddad, Chellaboina and Kablar (2001a), the resetting set $\mathcal{S}$ is defined in terms of a countable number of functions $n_k : \mathbb{R}^n \to (0, \infty)$, and is given by

$$\mathcal{S} = \cup_k \{(n_k(x), x, u_{\mathrm{c}}(n_k(x)) : x \in \mathbb{R}^n\}. \qquad \text{(II.5)}$$

The analysis of singularly impulsive dynamical systems with time delay and with a resetting set of the form (II.5) can be quite involved. In particular, such systems exhibit Zenoness, beating, as well as confluence phenomena wherein solutions exhibit infinitely many transitions in a finite times, and coincide after a given point of time, Haddad, Chellaboina and Kablar (2001a). In this paper we assume that existence and uniqueness properties of a given singularly impulsive dynamical system with time delay are satisfied in forward time. Furthermore, since singularly impulsive dynamical systems of the form (II.1)-(II.4) involve impulses at variable times they are time-varying systems.

Here we will consider singularly impulsive dynamical systems involving two distinct forms of the resetting set $\mathcal{S}$. In the first case, the resetting set is defined by a prescribed sequence of times which are independent of state $x$. These equations are thus called *time-dependent singularly impulsive dynamical systems with time delay*. In the second case, the resetting set is defined by a region in the state space that is independent of time. These equations are called *state-dependent singularly impulsive dynamical systems with time delay*.

### A. Time-Dependent Singularly Impulsive Dynamical Systems with Time Delay

Time-dependent singularly impulsive dynamical systems with time delay can be written as (II.1)–(II.4) with $\mathcal{S}$ defined as

$$\mathcal{S} = n \times \mathbb{R}^n \times \mathcal{U}_{\mathrm{c}}, \qquad \text{(II.6)}$$

where

$$n = t_1, t_2, \ldots \qquad \text{(II.7)}$$

and $0 < t_1 < t_2 < \ldots$ are prescribed resetting times. When an infinite number of resetting times are used and $t_k \to \infty$ as $k \to \infty$, then $\mathcal{S}$ is closed. Now (II.1)–(II.4) can be rewritten in the form of the *time-dependent singularly impulsive dynamical system with time delay*

$$E_{\mathrm{c}}\dot{x}(t) = f_{\mathrm{c}}(x(t,\tau)) + G_{\mathrm{c}}(x(t,\tau))u_{\mathrm{c}}(t),\ t \neq t_k, \qquad \text{(II.8)}$$
$$E_{\mathrm{d}}\triangle x(t) = f_{\mathrm{d}}(x(t,\tau)_{+}G_{\mathrm{d}}(x(t,\tau))u_{\mathrm{d}}(t),\ t = t_k, \qquad \text{(II.9)}$$
$$y_c(t) = h_{\mathrm{c}}(x(t,\tau)) + J_{\mathrm{c}}(x(t,\tau))u_{\mathrm{c}}(t),\ t \neq t_k, \qquad \text{(II.10)}$$
$$y_d(t) = h_{\mathrm{d}}(x(t,\tau)) + J_{\mathrm{d}}(x(t,\tau))u_{\mathrm{d}}(t),\ t = t_k. \qquad \text{(II.11)}$$

Since $0 \notin \tau$ and $t_k < t_{k+1}$, $\tau > 0$, it follows that the assumptions A1–A3 are satisfied. Since time-dependent singularly impulsive dynamical systems with time delay involve impulses at a fixed sequence of times, they are time-varying systems.

**Remark II.5.** *The time-dependent singularly impulsive dynamical system with time delay (II.8)–(II.11), with $E_{\mathrm{c}} = I$ and $E_{\mathrm{d}} = I$ includes as a special case the impulsive control problem addressed in the literature wherein at least one of the state variables of the continuous-time plant can be changed instantaneously to any given value given by an impulsive control at a set of control instants $\tau$, Haddad, Chellaboina and Kablar (2001a).*

### B. State-Dependent Singularly Impulsive Dynamical Systems with Time Delay

State-dependent singularly impulsive dynamical systems with time delay can be written as (II.1)–(II.4) with $\mathcal{S}$ defined as

$$\mathcal{S} = [0, \infty) \times \mathcal{Z}, \qquad \text{(II.12)}$$

where $\mathcal{Z} = \mathcal{Z}_x \times \mathcal{U}_{\mathrm{c}}$ and $\mathcal{Z}_x \subset \mathbb{R}^n$. Therefore, (II.1)–(II.4) can be rewritten in the form of the *state-dependent singularly impulsive dynamical system with time delay*

$$E_{\mathrm{c}}\dot{x}(t) = f_{\mathrm{c}}(x(t,\tau)) + G_{\mathrm{c}}(x(t,\tau))u_{\mathrm{c}}(t),$$
$$(x(t,\tau), u_{\mathrm{c}}(t)) \notin \mathcal{Z}, \qquad \text{(II.13)}$$
$$E_{\mathrm{d}}\triangle x(t) = f_{\mathrm{d}}(x(t,\tau)) + G_{\mathrm{d}}(x(t,\tau))u_{\mathrm{d}}(t),$$
$$(x(t,\tau), u_{\mathrm{c}}(t)) \in \mathcal{Z}, \qquad \text{(II.14)}$$
$$y_{\mathrm{c}}(t) = h_{\mathrm{c}}(x(t,\tau)) + J_{\mathrm{c}}(x(t,\tau))u_{\mathrm{c}}(t),$$
$$(x(t,\tau), u_{\mathrm{c}}(t)) \notin \mathcal{Z}, \qquad \text{(II.15)}$$
$$y_d(t) = h_{\mathrm{d}}(x(t,\tau)) + J_{\mathrm{d}}(x(t,\tau))u_{\mathrm{d}}(t),$$
$$(x(t,\tau), u_{\mathrm{c}}(t)) \in \mathcal{Z}. \qquad \text{(II.16)}$$

We assume that $(x_0, u_{\mathrm{c}0}) \notin \mathcal{Z}$, $\tau > 0$, $(0, 0) \notin \mathcal{Z}$, and that the resetting action removes the pair $(x, u_{\mathrm{c}})$ from the set $\mathcal{Z}$; that is, if $(x, u_{\mathrm{c}}) \in \mathcal{Z}$ then $(E_{\mathrm{d}}x + f_{\mathrm{d}}(x) + G_{\mathrm{d}}(x)u_{\mathrm{d}}, u_{\mathrm{c}}) \notin \mathcal{Z}$, $u_{\mathrm{d}} \in \mathcal{U}_{\mathrm{d}}$. In addition, we assume that if at time $t$ the trajectory $(x(t,\tau), u_{\mathrm{c}}(t)) \in \bar{\mathcal{Z}} \backslash \mathcal{Z}$, then there exists $\epsilon > 0$ such that for $0 < \delta < \epsilon$, $(x(t + \tau + \delta), u_{\mathrm{c}}(t + \delta)) \notin \mathcal{Z}$.

These assumptions represent the specialization of A1–A3 for the particular resetting set (II.12). It follows from these assumptions that for a particular initial condition, the resetting times $\tau_k(x_0)$ are distinct and well defined. Since the resetting set $\mathcal{Z}$ is a subset of the state space and is independent of time, state-dependent singularly impulsive dynamical systems with time delay are time-invariant systems. Finally, in the case where $\mathcal{S} \equiv [0, \infty) \times \mathbb{R}^n \times \mathcal{Z}_{u_{\mathrm{c}}}$, where $\mathcal{Z}_{u_{\mathrm{c}}} \subset \mathcal{U}_{\mathrm{c}}$ we refer to (II.13)–(II.16) as an input-dependent singularly impulsive dynamical system with time delay. Both these cases represent a generalization to the impulsive control problem considered in the literature.

## III. Lyapunov and Asymptotic Stability of Singularly Impulsive Dynamical Systems with Time Delay

In this section we present Lyapunov and asymptotic stability results of singularly impulsive dynamical systems with time delay.

**Theorem III.1.** *Suppose there exists a continuously differentiable function $V : \mathcal{D} \to [0, \infty)$ satisfying $V(0) = 0$, $V(E_{\mathrm{c/d}}x) \geq 0$, $x \neq 0$, and*

$$V'(E_{\mathrm{c}}x)f_{\mathrm{c}}(x) \leq 0, \qquad x \in \mathcal{D}, \tag{III.17}$$

$$V(E_{\mathrm{d}}x + f_{\mathrm{d}}(x) \leq V(x), \qquad x \in \mathcal{D}. \tag{III.18}$$

*Then the zero solution $x(t,\tau) \equiv 0$ of the undisturbed $((u_{\mathrm{c}}(t), u_{\mathrm{d}}(t_k)) \equiv (0,0))$ time-dependent singularly impulsive dynamical system with time delay (II.8),(II.9) is Lyapunov stable. Furthermore, if the inequality (III.17) is strict for all $x \neq 0$, then the zero solution $x(t,\tau) \equiv 0$ of the undisturbed $((u_{\mathrm{c}}(t), u_{\mathrm{d}}(t_k)) \equiv (0,0))$ time-dependent singularly impulsive dynamical system with time delay (II.8), (II.9) is asymptotically stable. If, in addition, $\mathcal{D} = \mathbb{R}^n$ and*

$$V(E_{\mathrm{c/d}}x) \to \infty \text{ as } \|x\| \to \infty, \tag{III.19}$$

*then the zero solution $x(t,\tau) \equiv 0$ of the undisturbed $((u_{\mathrm{c}}(t), u_{\mathrm{d}}(t_k)) \equiv (0,0))$ time-dependent singularly impulsive dynamical system with time delay (II.8), (II.9) is globally asymptotically stable, Haddad, Chellaboina, and Kablar (2001), Kablar (2003b).*

*Proof:* Prior to the first resetting time, we can determine the value of $V(x(t,\tau))$ as

$$V(E_{\mathrm{c}}x(t,\tau)) \;=\; V(E_{\mathrm{c}}x(0)) + \int_0^t V'(x(E_{\mathrm{c}}))f_{\mathrm{c}}(x(s,\tau)\mathrm{d}s, \qquad t \in [0, t_1]. \tag{III.20}$$

Between consecutive resetting times $t_k$ and $t_{k+1}$, we can determine the value of $V(x(t,\tau))$ as its initial value plus the integral of its rate of change along the trajectory $x(t,\tau)$, that is,

$$\begin{aligned} V(E_{\mathrm{c/d}}x(t,\tau) \;&=\; V(E_{\mathrm{d}}x(t_k) + f_{\mathrm{d}}(x(t_k)) \\ &\quad + \int_{t_k}^t V'(x(E_{\mathrm{c}}))f_{\mathrm{c}}(x(s,\tau))\mathrm{d}s, \\ &\qquad t \in (t_k, t_{k+1}], \end{aligned} \tag{III.21}$$

for $k = 1, 2, \ldots$. Adding and subtracting $V(x(E_{\mathrm{d}}t_k))$ to and from the right hand side of the (III.21) yields

$$\begin{aligned} V(E_{\mathrm{c/d}}x(t,\tau)) = {} & V(E_{\mathrm{c}}x(t_k)) + [V(E_{\mathrm{d}}x(t_k) \\ & + f_{\mathrm{d}}(x(t_k)) - V(E_{\mathrm{d}}x(t_k))] \\ & + \int_{t_k}^t V'(E_{\mathrm{c}}x(s,\tau))f_{\mathrm{c}}(x(s,\tau))\mathrm{d}s, \quad t \in (t_k, t_{k+1}], \end{aligned} \tag{III.22}$$

and in particular at time $t_{k+1}$,

$$\begin{aligned} V(E_{\mathrm{d}}x(t_{k+1})) = {} & V(E_{\mathrm{d}}x(t_k)) + [V(E_{\mathrm{d}}x(t_k) \\ & + f_{\mathrm{d}}(x(t_k))) - V(E_{\mathrm{d}}x(t_k))] \\ & + \int_{t_k}^{t_{k+1}} V'(x(s,\tau))f_{\mathrm{c}}(x(s,\tau))\mathrm{d}s. \end{aligned} \tag{III.23}$$

By recursively substituting (III.23) into (III.22) and ultimately into (III.20), we obtain

$$V(E_{\mathrm{c}}x(t,\tau)) = V(E_{\mathrm{c}}x(0)) + \int_0^t V'(E_{\mathrm{c}}x(s,\tau))f_{\mathrm{c}}(x(s,\tau))\mathrm{d}s$$

$$+ \sum_{i=1}^{k}[V(E_{\mathrm{d}}x(t_i) + f_{\mathrm{d}}(x(t_i))) - V(E_{\mathrm{d}}x(t_i))]]. \tag{III.24}$$

If we allow $t_0 = 0$, and $\sum_{i=1}^{0} = 0$, then (III.24) is valid for $k \in \mathcal{N}$. From (III.24) and (III.18) we obtain

$$\begin{aligned} & V(E_{\mathrm{c}}x(t,\tau)) \;\leq \\ & V(E_{\mathrm{c}}x(0)) + \int_0^t V'(E_{\mathrm{c}}x(s,\tau))f_{\mathrm{c}}(x(s,\tau))\mathrm{d}s, \\ & \qquad t \geq 0. \end{aligned} \tag{III.25}$$

Furthermore, it follows from (III.17) that

$$V(E_{\mathrm{c}}x(t,\tau)) \leq V(E_{\mathrm{c}}x(0)), \qquad t \geq 0, \tag{III.26}$$

so that Lyapunov stability follows from standard arguments.

Next, it follows from (III.18) and (III.24) that

$$V(E_{\mathrm{c}}x(t,\tau)) - V(E_{\mathrm{c}}x(s,\tau)) \;\leq\; \int_s^t V'(x(E_{\mathrm{c}}s,\tau))f_{\mathrm{c}}(x(s,\tau))\mathrm{d}s, \qquad t > s, \tag{III.27}$$

and, assuming strict inequality in (III.17), we obtain

$$V(E_{\mathrm{c}}x(t,\tau)) < V(E_{\mathrm{c}}x(s,\tau)), \qquad t > s, \tag{III.28}$$

provided $x(s,\tau) \neq 0$. Asymptotic stability, and, with (III.19), global asymptotic stability, then follow from standard arguments. ■

**Remark III.1.** *If in Theorem III.1 the inequality (III.18) is strict for all $x \neq 0$ as opposed to the inequality (III.17), and an infinite number of resetting times are used, that is, the set $\tau = \{t_1, t_2, \ldots\}$ is infinitely countable, then the zero solution $x(t,\tau) \equiv 0$ of the undisturbed time-dependent singularly impulsive dynamical system with time delay (II.8), (II.9) is also asymptotically stable. A similar remark holds for Theorem 2.2.2.*

**Remark III.2.** *In the proof of Theorem III.1, we note that assuming strict inequality in (III.17), the inequality (III.28) is obtained* provided $x(s,\tau) \neq 0$. *This proviso is necessary since it may be possible to reset the states to the origin, in which case $x(s,\tau) = 0$ for a finite value of $s$. In this case, for $t > s$, we have $V(E_{\mathrm{c}}x(t,\tau)) = V(E_{\mathrm{c}}x(s,\tau)) = V(0) = 0$. This situation does not present a problem, however, since reaching the origin in finite time is a stronger condition than reaching the origin as $t \to \infty$.*

**Remark III.3.** *If, additionally, in Theorem III.1 there exist scalars $\alpha, \beta, \epsilon > 0$, and $p \geq 1$, such that $\alpha\|x\|^p \leq V(E_{\rm c}x) \leq \beta\|x\|^p$, $x \in \mathcal{D}$, and (III.17) is replaced by $V'(E_{\rm c}x)f_{\rm c}(x) \leq -\epsilon V(E_{\rm c}x)$, $x \in \mathcal{D}$, then the zero solution $x(t,\tau) \equiv 0$ of the undisturbed $((u_{\rm c}(t), u_{\rm d}(t_k)) \equiv (0,0))$ time-dependent singularly impulsive dynamical system with time delay (II.8), (II.9) is exponentially stable. A similar remark holds for Theorem 2.2.2.*

**Remark III.4.** *Theorem III.1 presents sufficient conditions for time-dependent singularly impulsive dynamical systems with time delay in terms of Lyapunov functions that do not depend explicitly on time. Since time-dependent singularly impulsive dynamical systems are time-varying, Lyapunov functions that explicitly depend on time can also be considered. However, in this case the conditions on the Lyapunov functions required to guarantee stability are significantly harder to verify. For further details see Bainov and Simeonov (1989), Samoilenko and Perestyuk (1995), Ye, Michael, and Hou (1998).*

Next, we state a stability theorem for nonlinear state-dependent singularly impulsive dynamical systems with time delay.

**Theorem III.2.** *Suppose there exists a continuously differentiable function $V : \mathcal{D} \to [0,\infty)$ satisfying $V(0) = 0$, $V(E_{\rm c}x) \geq 0$, $x \neq 0$, and*

$$V'(E_{\rm c}x)f_{\rm c}(x) \leq 0, \qquad x \notin \mathcal{Z}_x, \qquad \text{(III.29)}$$

$$V(E_{\rm d}x + f_{\rm d}(x)) \leq V(E_{\rm c}x), \qquad x \in \mathcal{Z}_x. \qquad \text{(III.30)}$$

*Then the zero solution $x(t,\tau) \equiv 0$ of the undisturbed $((u_{\rm c}(t), u_{\rm d}(t_k)) \equiv (0,0))$ state-dependent singularly impulsive dynamical system with time delay (II.13), (II.14) is Lyapunov stable. Furthermore, if the inequality (III.29) is strict for all $x \neq 0$, then the zero solution $x(t,\tau) \equiv 0$ of the undisturbed $((u_{\rm c}(t), u_{\rm d}(t_k)) \equiv (0,0)$ state-dependent singularly impulsive dynamical system with time delay (II.13), (II.14) is asymptotically stable. If, in addition, $\mathcal{D} = \mathbb{R}^n$ and (III.19) is satisfied, then the zero solution $x(t,\tau) \equiv 0$ of the undisturbed $((u_{\rm c}(t), u_{\rm d}(t)k)) \equiv (0,0))$ state-dependent singularly impulsive dynamical system with time delay (II.13), (II.14) is globally asymptotically stable, Haddad, Chellaboina, and Kablar (2001), Kablar (2003b).*

*Proof:* For $\mathcal{S} = [0,\infty) \times \mathcal{Z}_x$ it follows from Assumptions A1–A3 that the resetting times $n_k(x_0)$ are well defined and distinct for every trajectory of (II.13), (II.14) with $(u_{\rm c}(t), u_{\rm d}(t_k)) \equiv (0,0)$. Now, the proof follows as in the proof of Theorem III.1 with $t_k$ replaced by $n_k(x_0)$. ■

**Remark III.5.** *To examine the stability of linear state-dependent singularly impulsive dynamical systems with time delay set $f_{\rm c}(x) = A_{\rm c}x$, and $f_{\rm d}(x) = (A_{\rm d} - E_{\rm d})x$ in Theorem III.2. Considering the quadratic Lyapunov function candidate $V(E_{\rm c/d}x) = x^{\rm T}E_{\rm c/d}^{\rm T}PE_{\rm c/d}x$, for the argument $E_{\rm c}x$ and $E_{\rm d}x$, respectively where $P > 0$, it follows from Theorem III.2 that the conditions*

$$x^{\rm T}(A_{\rm c}^{\rm T}PE_{\rm c} + E_{\rm c}^{\rm T}PA_{\rm c})x < 0, \qquad x \notin \mathcal{Z}_x, \qquad \text{(III.31)}$$

$$x^{\rm T}(A_{\rm d}^{\rm T}PA_{\rm d} - E_{\rm d}^{\rm T}PE_{\rm d})x \leq 0, \qquad x \in \mathcal{Z}_x, \qquad \text{(III.32)}$$

*establish asymptotic stability for linear state-dependent singularly impulsive dynamical systems with time delay. These conditions are implied by $P > 0$, $A_{\rm c}^{\rm T}PE_{\rm c} + E_{\rm c}^{\rm T}PA_{\rm c} < 0$, and $A_{\rm d}^{\rm T}PA_{\rm d} - E_{\rm d}^{\rm T}PE_{\rm d} \leq 0$ which can be solved using Linear Matrix Inequality (LMI) feasibility problem Boyd et al. (1994). See also Haddad, Chellaboina, and Kablar (2001a).*

## IV. Conclusion

In this paper we presented new class of singularly impulsive or generalized impulsive dynamical systems with delay. We gave assupmtions needed for the model and basic division of singularly impulsive dynamical systems into twio classes: time dependenet and state dependent. Next, we developed Lyapunov and asymptotic stability results.

## V. Future Work

It is left to develop invariant set theorem for singularly impulsive dynamical systems. Next, further work will concentrate to specializing this results and developing to time-delay systems. The last is motivated by recognized need in biological applications.

On the other hand finite-time and practical stability results will be developed for the class of impulsive and singularly impulsive dynamical systems with delay.

## VI. Acknowledgment

The author is grateful to all that were supportive and creative in achieving these results.

## VII. References

[1] Bainov D.D. and P.S. Simeonov, *Systems with Impulse Effect: Stability, Theory and Applications*. England, Ellis Horwood Limited, 1989.

[2] Boyd S., L.E. Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory.* In: SIAM studies in applied mathematics, 1994.

[3] Back A., J. Guckenheimer, and M. Myers, ”A dynamical simulation facility for hybrid systems,” In R. Grossman, A. Nerode, A. Ravn and H. Rischel (Eds), *Hybrid Systems*, New York: Springer-Verlag, pp. 255–267, 1993.

[4] Branicky M. S., ”Multiple-Lyapunov functions and other analysis tools for switched and hybrid systems,” *IEEE Transactions on Automatic Control*, Vol. 43, pp. 475–482, 1998.

[5] Branicky M. S., V. S. Borkar, and S. K. Mitter, ”A unified framework for hybrid control: model and optimal control theory,” *IEEE Transactions on Automatic Control*, Vol. 43, pp. 31–45, 1998.

[6] Brogliato B., *Non-smooth Impact Mechanics: Models, Dynamics and Control*, London: Springer-Verlag, 1996.

[7] Brogliato B., S. I. Niculescu, and P. Orhant, ”On the control of finite-dimensional mechanical systems with unilateral

constraints," *IEEE Transactions on Automatic Control*, Vol. 42, pp. 200–215, 1997.

[8] Guan Z-H., C.W.Chan, A. Y.T.Leung, and G. Chen, "Robust Stabilization of Singular-Impulsive-Delayed Systems with Nonlinear Perturbations," *IEEE Trans. On Circ. And Sys. - I: Fundamental Theory and Applications*, vol. 48, No. 3, 2001.

[9] Haddad W.M., V.Chellaboina, N.A. Kablar, "Nonlinear Impulsive Dynamical Systems: Stability and Dissipativity," *Proc. IEEE Conf. Dec. Contr.*, pp. 5158-5163, Phoenix, AZ, 1999a. Also in: *Int. J. Contr.*, vol. 74, pp. 1631-1658, 2001a.

[10] Haddad W.M., V.Chellaboina, N.A. Kablar, "Nonlinear Impulsive Dynamical Systems: Feedback Interconnections and Optimality," *Proc. IEEE Conf. Dec. Contr.*, Phoenix, AZ, 1999b. Also in: *Int. J. Contr.*, vol. 74, pp. 1659-1677, 2001b.

[11] Haddad W.M., N.A. Kablar, V.Chellaboina, "Robustness of Uncertain Nonlinear Impulsive Dynamical Systems," *Proc. IEEE Conf. Dec. Contr.*, Sidney, pp. 2959-2964, Australia, 2000. Also in: *Nonlinear Anal.*, submitted.

[12] Haddad W.M., N.A. Kablar, V.Chellaboina, "Optimal Disturbance Rejection of Nonlinear Impulsive Dynamical Systems," *Nonlinear Anal.*, published, 2005.

[13] Kablar N.A., "Singularly Impulsive or Generalized Impulsive Dynamical Systems," *Proc. Amer. Contr. Conf.*, Denver, CO, 2003a.

[14] Kablar N.A., "Lyapunov and Asymptotic Stability of Singularly Impulsive Dynamical Systems," *Proc. IEEE Conf. Dec. Contr.*, USA, 2003b.

[15] Kablar N.A., "Finite-Time Stability of Singularly Impulsive Dynamical Systems," *IEEE Conf. Decision and Control*, Atlanta, USA, 2010.

[16] Kablar N.A., "Robust Stability Analyse of Singularly Impulsive Dynamical Systems," *Proc. Amer. Contr. Conf.*, USA, 2006.

[17] Lakshmikantham V., D. D. Bainov, and P. S. Simeonov, *Theory of Impulsive Differential Equations*, Singapore: World Scientic, 1989.

[18] Lakshmikantham V. and X. Liu, "On quasi stability for impulsive differential systems," *Non. Anal. Theory, Methods and Applications*, Vol. 13, pp. 819–828, 1989.

[19] Lakshmikantham V., S. Leela, and S. Kaul, "Comparison principle for impulsive differential equations with variable times and stability theory," *Non. Anal. Theory, Methods and Applications*, Vol. 22, pp. 499–503, 1994.

[20] Leonessa A., W. M. Haddad, and V. Chellaboina, *Hierarchical Nonlinear Switching Control Design with Applications to Propulsion Systems*, London: Springer- Verlag, 2000.

[21] Liu X., "Stability results for impulsive differential systems with applications to population growth models," *Dynamic Stability Systems*, Vol. 9, pp. 163–174, 1994.

[22] Lygeros J., D. N. Godbole, and S. Sastry, "Verified hybrid controllers for automated vehicles," *IEEE Transactions on Automatic Control*, Vol. 43, pp. 522–539, 1998.

[23] Passino K.M., A. N. Michel, and P. J. Antsaklis, "Lyapunov stability of a class of discrete event systems," *IEEE Transactions on Automatic Control*, Vol. 39, pp. 269–279, 1994.

[24] Raibert M. H, *Legged Robots that Balance*, MIT Press, Cambridge, MA, 1986.

[25] Samoilenko A. M. and N.A. Perestyuk, *Impulsive Differential Equations*. World Scientific, 1995.

[26] Ye H., A.N. Michel, and L. Hou, "Stability Analysis of Systems with Impulsive Effects," *IEEE Trans. Autom. Contr.*, vol. 43, pp. 1719–1723, 1998.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Fault Detection and Isolation for PEMFC Systems under Closed-loop Control

Mahanijah Md Kamal, Student Member, IEEE
Control System Research Group, School of Engineering
Liverpool John Moores University, Byrom Street,
Liverpool L3 3AF, United Kingdom
M.B.Md-Kamal@2010.ljmu.ac.uk

Dingli Yu
Control System Research Group, School of Engineering
Liverpool John Moores University, Byrom Street,
Liverpool L3 3AF, United Kingdom
d.yu@ljmu.ac.uk

*Abstract*— **In this work, a model-based fault detection and isolation (FDI) is developed for proton exchange membrane (PEM) fuel cell (FC) stack that is under feed-forward plus feedback control. The fault detection is achieved using an independent radial basis function (RBF) network model, whilst the fault isolation is based on the RBF classification. The novelty is that the RBF model of independent mode is used to predict the future outputs of the FC stack and a RBF classifier is used to classify five types of fault introduced to the PEMFC systems. To validate the method, a benchmark model developed by Michigan University is used in the simulation to analyze the effectiveness of the method for actuator, component and three sensor faults. The FDI results corresponding to those scenarios show that the simulated different types of fault are successfully detected and isolated.**

***Keywords-Proton exchange membrane fuel cell; feed-forward; feedback; fault detection; fault isolation; radial basis function; independent model***

## I. INTRODUCTION

Process faults, if undetected, have a serious impact on process economy, product quality, safety, pollution level and productivity. In order to detect, diagnose and correct these abnormal process behaviors, efficient and advanced automated diagnostic systems are of great importance to modern industries [1]. Once a fault has been detected and its evolution is monitored, the severity of that fault can be evaluated and a decision can be made on the course of action to take. Monitoring creates the opportunity to strategically plan and schedule outrages and to manage equipment utilization and availability [2]. Fault detection, isolation and reconfiguration (FDIR) is an important and challenging task in many engineering applications and continues to be an active area of research in the control environment [3]. In some cases, if a fault can be quickly detected and identified, appropriate reconfiguration control actions may be taken. FDIR is a control methodology which ensures continual safe and acceptable operation of the system when a fault occurs through fault detection and isolation (FDI). Many devices depend on automatic control for satisfactory operation, and while assuring stability and performance with all components functioning properly. If the control system's structure or parameter can be altered in response to system failure, it is said to be reconfigurable [4].

There are a large number of publications on the fuel cell (FC) studies, but studies on FDI are still a few. Model-based FDI methods for PEMFC become more and more important because it involved the comparison between the observed behavior of the process with a reference model. Model-based approach gives the insight analysis of the subsystem interactions and also provides guidelines during the conduction of the experiment. The system behavior can been analyzed in depth understanding and later this information can be used for future design and development. For fault detection problem, the most effective way is by using the model-based approach based on a residual generation. Here, the difference between the actual process and estimated output of the process is used as a residual vector. In [5] presented and tested a model-based fault diagnosis methodology based on the relative residual fault sensitivity. In this method, it checks the consistency of observed behavior and then isolate the component that is in fault in different sensitivities. While, a robust fault detection based on the use of LPV observer using output zonotopes was proposed by [6]. Here, fault isolation is based on set of structured residuals that are analyzed using a relative fault sensitivity approach.

Neural networks have been proposed as an alternative method for fault diagnosis by many authors especially to tackle the nonlinear behavior. Reference [7] used a Bayesian network as an early alert to diagnose faults in the air reaction fan, faults in the cooling system, growth of the fuel crossover and internal loss current and faults in the hydrogen feed line. Alternatively, to improve reliability and durability of PEMFC systems, [8] presents a flooding diagnosis based on black-box model of elman neural network (ENN). Here, ENN is used to do a comparison between measured and calculated pressure drops. The model-based on ENN is trained with data recorded in flooding-free condition and the difference between calculated and experimental pressure drop is used as the residual. Also concern about this problem in FC, [9] presents an electric equivalent model for FC system diagnosis emphasis on FC flooding detection induced by temperature. In this paper, to tackle the efficiency of the overall PEMFC systems, a model-based FDI based on residual generation is used to implement the fault detection whilst for fault isolation a RBF networks is

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

used as a classifier. Therefore, to make the FDI monitoring system more efficient and robust to the faults in the PEMFC systems, an independent RBF network is used for fault identification, detection and isolation. The aim of this work is to develop a FDI scheme under closed-loop system for PEMFC using an independent RBF network model which can detect five types of faults in the FC systems accordingly but also can also isolate them accurately.

## II. PEMFC DYNAMICS

The proton exchange membrane fuel cell (PEMFC) systems offer high efficiency and low emissions and has been become popular as an alternate power source for various application such as transportation, telecommunication, portable utilities, stationary and power generation. A typical PEMFC system normally consist of four subsystems, which include the reactant flow subsystem, the heat and temperature subsystem, the water management subsystem and the power management subsystem. The PEMFC stack is made up of 381 cells with an active area of 280cm$^2$ and the stack operating temperature is at 80°C developed by University Michigan is used as a test bench.

### A. Compressor Model

The flow and temperature out of the compressor ($W_{cp}$ and $T_{cp}$) depend on the compressor rotational speed $\omega_{cp}$. A lumped rotational model is used to represent the dynamic behaviour of the compressor [10]:

$$J_{cp}\frac{d\omega_{cp}}{dt} = \tau_{cm} - \tau_{cp} \tag{1}$$

where $\tau_{cm}(v_{cm},\omega_{cp})$ is the compressor motor (CM) torque and $\tau_{cp}$ is the load torque. The compressor motor torque is calculated using a static motor equation:

$$\tau_{cm} = \eta_{cm}\frac{k_t}{R_{cm}}\left(V_{cm} - k_v\omega_{cp}\right) \tag{2}$$

where $k_t$, $R_{cm}$ and $k_v$ are motor constants and $\eta_{cm}$ is the motor mechanical efficiency. The torque required to drive the compressor is calculated using the thermodynamic equation:

$$\tau_{cp} = \frac{c_p T_{atm}}{\omega_{cp}\eta_{cp}}\left[\left(\frac{p_{sm}}{p_{atm}}\right)^{(\gamma-1)/\gamma} - 1\right]W_{cp} \tag{3}$$

where $\gamma$ is the ratio of the specific heats of air (=1.4), $c_p$ is the constant pressure specific heat capacity of air (=1004 J.kg$^{-1}$.K$^{-1}$), $\eta_{cp}$ is the motor compressor efficiency, $p_{sm}$ is the pressure inside the supply manifold and $p_{atm}$ and $T_{atm}$ are the atmospheric pressure and temperature, respectively.

### B. Supply Manifold Model

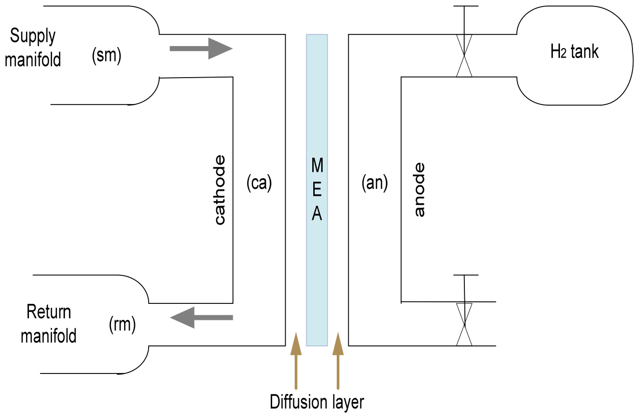


Figure 1. The fuel cell reactant supply system

The cathode supply manifold (sm) includes pipe and stack manifold volumes between the compressor and the fuel cells as shown in Fig. 1. The supply manifold pressure, *psm*, is governed by mass continuity and energy conservation equations [11]:

$$\frac{dm_{sm}}{dt} = W_{cp} - W_{sm,out} \tag{4}$$

$$\frac{dp_{sm}}{dt} = \frac{\gamma R_a}{V_{sm}}\left(W_{cp}T_{cp} - W_{sm,out}T_{sm}\right) \tag{5}$$

where R is the universal gas constant and $M_a^{atm}$ is the molar mass atmospheric air at $\Phi_{atm}$, $V_{sm}$ is the manifold volume and $T_{sm} = \frac{p_{sm}v_{sm}M_a^{atm}}{Rm_{sm}}$ is the supply manifold gas temperature.

## III. FDI METHOD WITH INDEPENDENT RBF MODEL

The basic structure of an independent radial basis function (RBF) model for PEMFC dynamic systems proposed in this work can be referred to Fig. 2. Here, two inputs and three outputs of the process with their delayed values form the 8 inputs of the RBF model, while the three process outputs are the model outputs. The chosen input output orders are according to the training experience and checking the process dynamics. The model prediction errors are generated for residual generation.

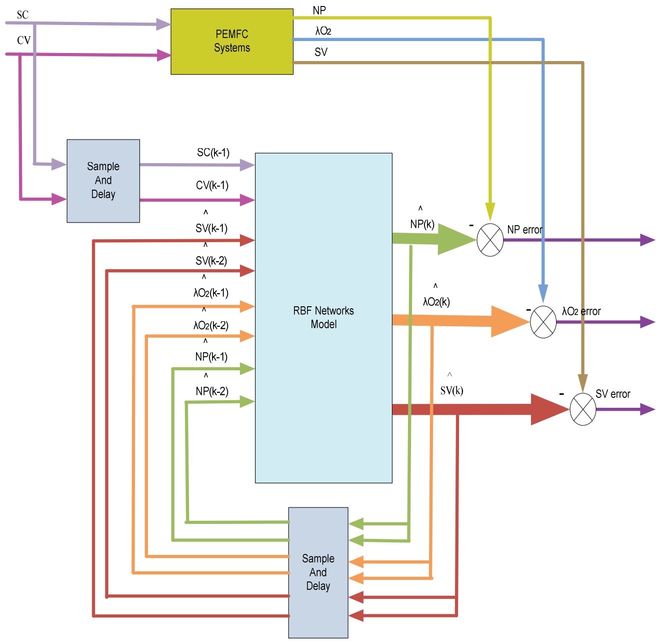


Figure 2. The structure of an independent RBF network

## IV. CONTROLLER DESIGN

For air supply in the PEMFC systems, the required air flow is indicated by the desired oxygen excess ratio, $\lambda O_2$ = 2. Generating rapid increase in air flow, however, requires a large amount of power drawn by the compressor motor and a affecting the system net power [12]. The combination of feed-forward and feedback control design objective is to manipulate the compressor motor input voltage, $V_{cm}$, in order to maintain $\lambda O_2 = 2$.

### A. Feed-forward Controller

A feed-forward control is used to control $V_{cm}$ based on the current drawn from the FC stack. In this work, look-up table act as feed-forward control as presented in Table I with respect to the signal range of stack current ranging from 100 to 300 amperes. To design the feed-forward controller, the stack current signal is adjust at the value illustrated in Table I and fed to the FC stack while tuning the $V_{cm}$ until $\lambda O_2 = 2$.

TABLE I. THE DESIGN OF FEED-FORWARD CONTROLLER

| **Stack Current (Ampere)** | **Compressor Voltage (Volt)** | **Gain = Output / Input** | **Stack Current (Ampere)** | **Compressor Voltage (Volt)** | **Gain = Output / Input** |
|---|---|---|---|---|---|
| 105 | 102.33 | 0.9746 | 190 | 163 | 0.8579 |
| 110 | 105.98 | 0.9635 | 195 | 166.45 | 0.8536 |
| 115 | 109.98 | 0.9563 | 200 | 169.85 | 0.8493 |
| 120 | 113.15 | 0.9429 | 205 | 173.25 | 0.8452 |
| 125 | 116.77 | 0.9342 | 210 | 176.65 | 0.8412 |
| 130 | 120.43 | 0.9264 | 215 | 179.85 | 0.8365 |
| 135 | 124.05 | 0.9189 | 220 | 183.25 | 0.8330 |
| 140 | 127.68 | 0.9120 | 225 | 186.45 | 0.8287 |
| 145 | 131.29 | 0.9054 | 230 | 189.65 | 0.8246 |
| 150 | 134.89 | 0.8993 | 235 | 192.75 | 0.8202 |
| 155 | 138.48 | 0.8934 | 240 | 195.95 | 0.8165 |
| 160 | 142.05 | 0.8878 | 245 | 199.05 | 0.8124 |
| 165 | 145.58 | 0.8823 | 250 | 202.05 | 0.8082 |
| 170 | 149.15 | 0.8774 | 255 | 205 | 0.8040 |
| 175 | 152.65 | 0.8723 | 260 | 208 | 0.8000 |
| 180 | 156.10 | 0.8672 | 265 | 211 | 0.7962 |
| 185 | 159.60 | 0.8627 | 270 | 214.10 | 0.7930 |

### B. Proportional-Integral-Derivative Controller

A proportional-integral-derivative (PID) controller is used to reduce the effects the disturbances that can be measured and also to improve the response to reference signals. The PID controller equation is given by:

$$PID_{controller} = 200\left(1 + \frac{1}{0.6153\,s} + 0.05\,s\right) \qquad (6)$$

Fig. 3 shows the overall control systems of feed-forward and a closed-loop control implemented in this work. In the diagram, the stack current acts as a disturbance to the PEMFC systems with a reference input set at 2 ($\lambda O_2 = 2$). The output of $\lambda O_2$ need to maintain in order to avoid oxygen starvation from happening.

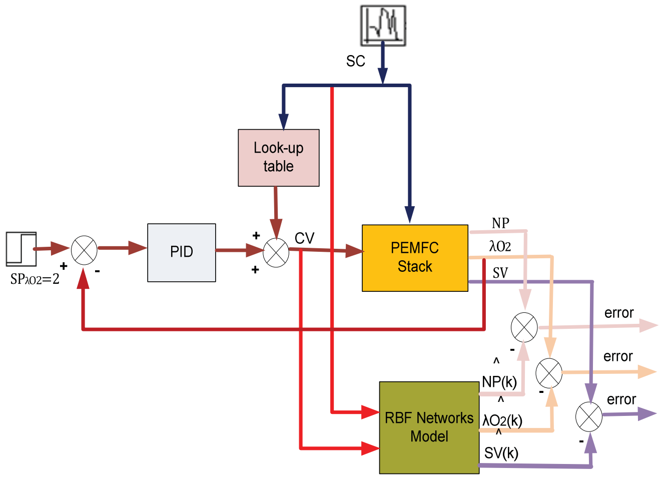


Figure 3. The overall system of FDI using feed-forward and feed-back controller

## V. SIMULATING FAULTS

In this study, five faults are introduced to a known test bench PEMFC based on the model developed in Michigan University. First one is an actuator fault, which is simulated by superimposing a -10% change of the compressor motor voltage measurement. The second is the air leak in the supply manifold which is a typical component fault. The third to fifth are three sensor faults for the three outputs, which are simulated by 10% deviation superimposed to the net power, $\lambda O_2$ and stack voltage output measurements. The PEMFC simulator was modified to include five possible fault scenarios which may occur during the normal operation of PEMFC systems. Fig. 4 shows the five faults introduced to the overall PEMFC systems.

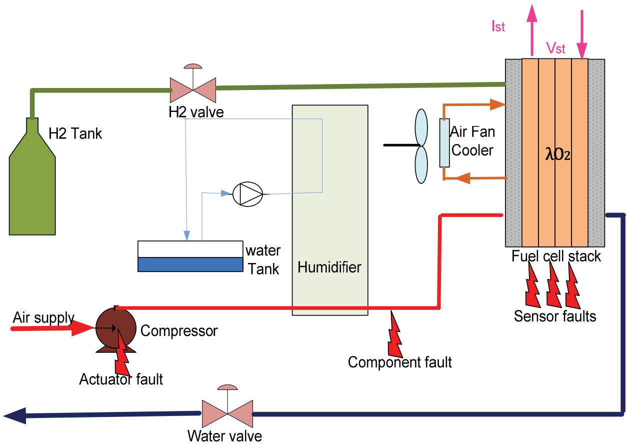


Figure 4. The schematic of PEMFC systems with five types of faults

*1) Actuator fault:* Mostly centrifugal compressor is used in FCs are susceptible to surge and choke that limit the efficiency and performance of the compressor. The compressor voltage will be changed if the compressor experience surge and choke and affected the air flow in the supply manifold. The compressor motor performance is reduced by -10% of the total compressor motor voltage from the sample intervals, k=2500-2550 to reflect the scenario of the fault which happens at the actuator part.

2) *Component fault:* Air leakage in the supply manifold makes the pressure in the cathode decrease. Therefore to collect the FC stack data subjected to the air leak fault, equation (5) is modified to:

$$\frac{dp_{sm}}{dt} = \frac{\gamma R_a}{V_{sm}}\left(W_{cp}T_{cp} - W_{sm,out}T_{sm} - \Delta l\right) \tag{7}$$

where Δl is used to simulate the leakage from the air manifold, which is subtracted to increase the air outflow from the supply manifold. Δl=0 represents that there is no air leakage in the supply manifold. The air leakage is simulated by -10% change of the pressure inside the supply manifold. The fault occurs at the sample intervals, *k* = 2000-2050.

3) *Sensor faults:* Net power, $\lambda O_2$ and stack voltage sensors are considered experiencing over-reading faults. The faulty sensor data used was the data from the collected data set, superimposed with a 10% change of the measured net power over the sample interval, k = 500-550, a 10% change of the measured $\lambda O_2$ over the sample intervals, k = 1000-1050 and a 10% change of the measured stack voltage over the sample intervals, k = 1500-1550.

## VI. FAULT DETECTION AND ISOLATION

### *A. Fault Detection*

Though the filtered squared model prediction error for each output could be used as fault detection signal, a residual signal is generated by combine these prediction errors, so that the sensitivity of the residual to each fault can be significantly enhanced, and consequently the false alarm rate would be reduced. The residual in this work is defined as in (8).

$$re = \sqrt{e_{NP}^2 + e_{\lambda O2}^2 + e_{SV}^2} \tag{8}$$

where $e_{NP}$ is the filtered modeling error of net power, $e_{\lambda O2}$ is the filtered modeling error of $\lambda O_2$ and $e_{SV}$ is the filtered modeling error of stack voltage. The signal with faults is clearly been identified and less influences by a noise signal.

### *B. Fault Isolation*

RBF classifier is a nonlinear static network. The network is trained with a set of data collected under each of the five faults and no-fault condition. The five outputs are arranged in this way: The target for any one output is arranged to be "1" when the corresponding single fault occurs, and to be "0" when this single fault does not occur. In this study, 3000 samples of data were collected with the first fault occurring during *k* = 500~550, the second fault occurring during *k* = 1000-1050, and etc. Then, the generated filtered and squared model prediction error vector from the fault detection part was used as the input data of the RBF classifier. Correspondingly, the target matrix $X_0$ has 3000 rows and 5 columns. The entries from the 500th row to the 550th row in the first column are "1", while the other entries are "0". The arrangement for the column 2 to 5 is done in the same way. This is shown as in Table II.

TABLE II. THE TARGET MATRIX IN TRAINING THE RBF CLASSIFIER

| **Rows** | **$X_0$** | | | | |
|---|---|---|---|---|---|
| 500~550 | [1 | 0 | 0 | 0 | 0] |
| 1000~1050 | [0 | 1 | 0 | 0 | 0] |
| 1500~1050 | [0 | 0 | 1 | 0 | 0] |
| 2000~2050 | [0 | 0 | 0 | 1 | 0] |
| 2500~2550 | [0 | 0 | 0 | 0 | 1] |

## VII. SIMULATION RESULTS

A random amplitude signals (RAS) of stack current used as disturbances to the PEMFC systems has been injected to the FC stack. At the same time, the constructed table described in the previous section act as a feed-forward control is the input to compressor motor. The RAS excitation signals of stack current are generated randomly to cover the whole range of frequencies and entire operating space of amplitude in the PEMFC systems.

Later, a data set with 3000 samples is acquired from the plant when the five faults are simulated to the plant as described in previous section. The simulation result of three PEMFC outputs and the corresponding five faults is shown in Fig. 5. It shows the squared filtered model prediction error for the three output variables. As can be seen, there are more than one faults occurred in these three outputs.

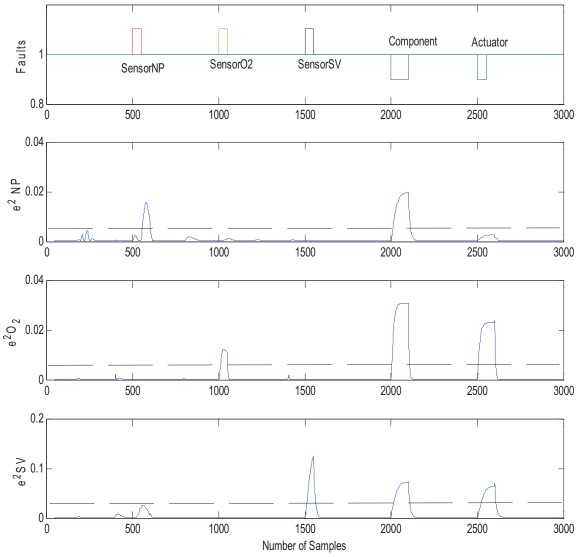


Figure 5. Filtered model predicted errors

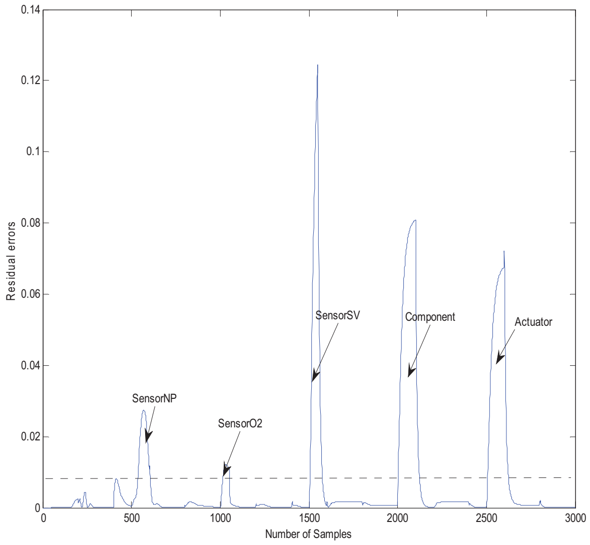


Figure 6. The fault classification of residual generator

In order to do fault classification, the residual generation as stated in equation (8) was applied. Here, the fault occurrence can clearly identified and detected with their respective threshold after the implementation. It is observed in Fig. 6 that all five faults of +10% for three sensors and -10% for component and actuator faults are clearly detected.

The target matrix in Table II was used in training of the RBF classifier. The centres and widths of the network were chosen using the K-means clustering algorithm and the p-nearest centre algorithm. The weights were trained with using the RLS algorithm with the following data, $\mu$= 0.99999, $w(0) = 1.0\times10^{-6}\times U_{(nh\times3)}$, $P(0)=1.0\times10^{8}\times I_{(nh)}$; where $I$ is an identity matrix and $U$ is an ones matrix. The RBF networks model only used the three rows of the PEMFC outputs matrix which contain the values of net power, $\lambda O_2$ and stack voltage. After training, a similar data set with also 3000 samples, with the same five faults simulated, was collected. These data was applied to the fault detection part and then to the isolation part with the trained RBF classifier. The five outputs of the classifier are displayed in Fig. 7.

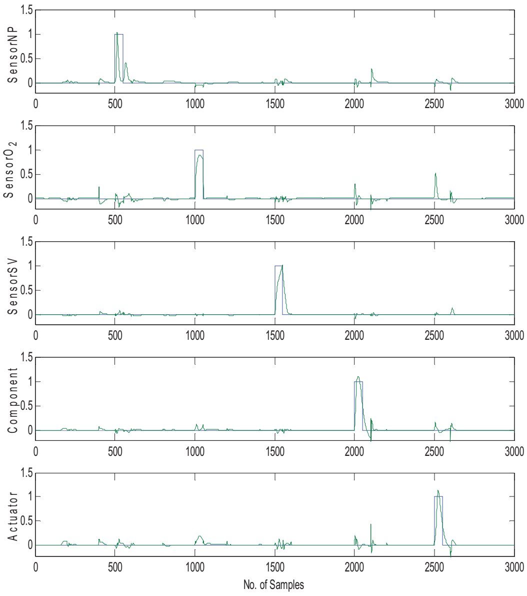


Figure 7. The fault isolation for five faults during training

From the fault isolation signals in Fig. 7 it is clearly observed that all considered five faults have been isolated. The RBF classifier successfully suppressed the corresponding output value for the no-fault-occurring period, while promoted the corresponding output value for the fault-occurring period. It is noticed in Fig. 7 that the fault isolation signals are very noisy and that would cause false alarm. Then, the RBF classifier outputs are filtered and the filtered signals are displayed in Fig. 8. It is obvious that the filtered fault isolation signals are much smoother and the robustness of the signal to modeling errors, interactions between variables and noise is greatly enhanced. It is important to isolate the malfunction devices in the systems for easy troubleshooting and maintenance purposes. By doing this step, the device can easily be replaced and any appropriate action can be taken quickly and therefore it can save time and increase productivity.

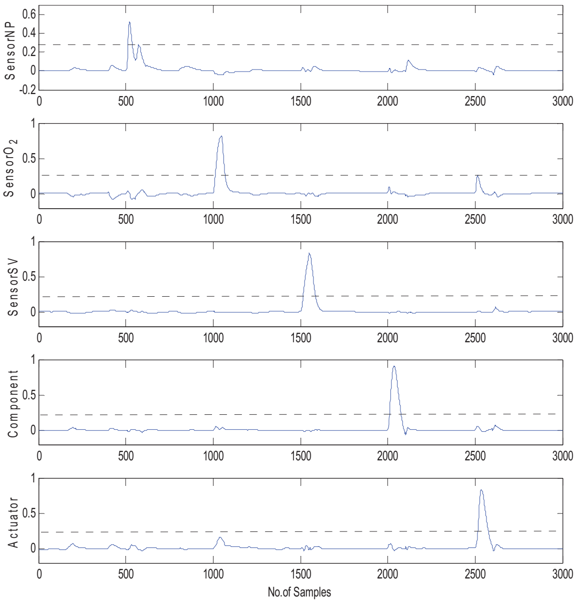


Figure 8. The location of five faults in the PEMFC systems

## VIII. CONCLUSIONS

A combination of feed-forward and feedback controller is designed to regulate the $\lambda O_2$ during the changes of stack current in the FC stack. FDI has been developed for the PEMFC under the closed-loop control. Five faults are simulated and diagnosed. The simulation results show that the new approach using the residual generation to do fault detection and the RBF classifier to apply fault isolation is successfully implemented. The 10% faults in the actuator, component and three sensors can be clearly detected and isolated. Here, the fault condition is considered occurred as a single fault at a time. But this result can be extended to the fault condition of multi-faults occurring simultaneously. The extension for fault detection part is straightforward, while for fault isolation needs more complex training of the fault classifier. The developed method has a big potential to be applied to real world dynamic systems. Also, the method is not limited to FC systems, and can be applied to other multivariable nonlinear dynamic systems with some modifications.

## REFERENCES

[1] A. Rosich, R. Sarrete, V. Puig and T. Escobet, "Efficient Optimal Sensor Placement for Model-based FDI using an Incremental Algorithm", in Proceedings of the 46th IEEE Conference on Decision and Control, pp. 2590-2595, 2007

[2] J.P. Gibeault and J.K. Kirkup, "Early detection and Continuous Monitoring of Dissolved Key Fault Gasses in Transformers and Shunt Reactors", Electrical Electronics Insulation Conference, pp. 285-293, 1995.

[3] I. Hwang, S. Kim, Y. Kim and C. E. Seah, "A Survey of Fault Detection, Isolation and Reconfiguration Methods", IEEE Transactions On Control Systems Technology, vol. 18, no. 3, pp. 636-653, 2010.

[4] R. F. Stengel, "Intelligent Failure-Tolerant Control", in 5th International Symposium on Intelligent Control, pp. 14-23, 1990.

[5] T. Escobet, D. Feroldi, S. De Lira, V. Puig, J. Quevedo, J. Riera and M. Serra, "Model-based fault diagnosis in PEM fuel cell systems", Journal of Power Sources 192, pp. 216-223, 2009.

[6] S. De Lira, V. Puig and J. Quevedo, "PEM Fuel Cells System Robust LPV model-based Fault Diagnosis", in 20th International Workshop on Principles of Diagnosis, pp. 91-98, 2009.

[7] L. A. M. Riascos, M. G. Simões and P. E. Miyagi, "Fault identification in Fuel Cells Based on Bayesian Network Diagnosis", in ABCM Symposium Series in Mechatronics - vol. 2, pp. 757-764, 2006.

[8] N. Yousfi Steiner, D. Candusso, D. Hissel and P. Mocoteguy, "Model-based diagnosis for proton exchange membrane fuel cells", Mathematics and Computer In Simulation, pp. 158-170, 2010.

[9] A. Hernandez, D. Hissel and R. Outbib,"Modeling and Fault Diagnosis on a Polymer Electrolyte Fuel Cell Using Electrical Equivalent Analysis", IEEE Transactions on Energy Conversion, vol. 25, no. 1, pp. 148-160, 2010.

[10] J. T. Pukrushpan, H. Peng and A. G. Stefanopoulou, "Control-Oriented Modeling and Analysis for Automotive Fuel Cell Systems", Journal of Dynamic Systems, Measurement and Control, vol. 126, pp. 14-25, 2004.

[11] J. T. Pukrushpan, A. G. Stefanopoulou and H. Peng, "Control of Fuel Cell Breathing", IEEE Control Systems Magazines, vol. 24, no. 2, pp. 30-46, 2004.

[12] J. T. Pukrushpan, A. G. Stefanopoulou and H. Peng, "Modeling and Control for PEM Fuel Cell Stack System", in Proceedings of the American Control Conference, pp. 3117-3122, 2002.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Determination of the domain of attraction and regions of guaranteed cost for robust model predictive controllers based on linear matrix inequalities

Fernanda Quelho Rossi, Ronaldo Waschburger, Roberto Kawakami Harrop Galvão
Electronics Engineering Department, Instituto Tecnológico de Aeronáutica (ITA)
São José dos Campos, SP, Brazil
Emails: fer.qrossi@gmail.com, ronaldow@ita.br, kawakami@ita.br

***Abstract*—The robust model-based predictive control (RMPC) formulation originally proposed in [1] ensures convergence of the state trajectory to the origin and satisfaction of operational constraints, provided that a given system of LMIs is feasible at the beginning of the control task. The largest domain of attraction of the origin under the resulting closed-loop control law can be defined as the set of all state values for which the LMIs are feasible. The present paper demonstrates that such a set is convex and symmetric about the origin, which allows the determination of extreme points through the solution of a modified version of the original RMPC optimization problem. An inner approximation of the largest domain of attraction can then be generated as the convex hull of these extreme points. The convexity and symmetry properties are also demonstrated for regions of guaranteed cost, defined as the set of initial states for which the resulting cost is upper-bounded by a given value. Inner approximations of such regions can also be obtained by solving a modified version of the RMPC optimization problem. For illustration, a numerical simulation model of an angular positioning system is employed, as in [1]. In this example, the proposed approximations were found to be in agreement with the feasibility and cost results obtained in a pointwise manner for a grid of initial conditions.**

***Index Terms*—Robust model predictive control, linear matrix inequalities, domain of attraction, convex optimization.**

## I. Introduction

The term Model-based Predictive Control (MPC) refers to a body of techniques that involve the solution of an optimal control problem within a receding horizon [2]. MPC has become widespread in several application areas, mainly due to the possibility of addressing operational constraints in an explicit manner [3]. Constraint satisfaction and closed-loop stability can be guaranteed by a proper formulation of the receding-horizon optimization problem. Usually, the adoption of target sets that are invariant under locally stabilizing control laws is employed for this purpose [4]. However, the design guarantees may be lost if the prediction model does not match the actual plant dynamics. Such a mismatch may arise due to modelling approximations (linearization and order reduction, for example), parameter uncertainties or variations in the plant behaviour due to faults or natural aging. This problem has motivated the development of robust MPC (RMPC) techniques.

Early formulations of RMPC involved the online solution of a min-max optimization problem, where the worst case value of the cost function was evaluated over a set of uncertain plants [5], [6], [7]. However, the computational cost of the resulting problem could become prohibitive for actual implementation. In this context, Kothare and collaborators [1] proposed an RMPC approach based on linear matrix inequalities (LMIs). By using the proposed approach, the optimization problem was cast into a semi-definite programming (SDP) form [8], which allowed the use of efficient numerical solvers to obtain the optimal control in polynomial time. Moreover, operational constraints could be easily introduced by augmenting the problem formulation with additional LMIs. This seminal work was later extended to encompass the use of output feedback [9], [10], [11], as well as the control of nonlinear systems [12] systems with uncertain time delay [13], [14], [15], and systems with asymmetric output constraints [16]. Modifications to the LMI formulation aimed at reducing conservatism have also been proposed [17], [18], [19], [20].

Within the framework developed in [1], the state trajectory is guaranteed to converge to the origin with satisfaction of the operational constraints, provided that the system of LMIs is feasible at the beginning of the control task. Therefore, the largest domain of attraction of the origin under the closed-loop control law can be defined as the set of all state values (i.e. initial conditions for the state trajectory) for which the LMIs are feasible. However, the analytical or numerical characterization of such a domain of attraction was not discussed in [1]. In fact, although the solution of the SDP problem for a given initial condition can be used to establish an asymptotically stable invariant ellipsoid [1], [21], such an ellipsoid is not necessarily the largest domain of attraction for the origin. It is possible to maximize the size of this invariant ellipsoid by using a determinant maximization procedure, as proposed in [12]. Yet, one may still argue that the largest domain of attraction is not necessarily of ellipsoidal shape.

The present paper establishes some properties of the largest domain of attraction $\mathcal{D}$ for the RMPC approach developed in [1]. More specifically, it is demonstrated that $\mathcal{D}$ is convex and symmetric about the origin. In view of such properties, extreme points of $\mathcal{D}$ can be found by solving a modified version of the original RMPC optimization problem. An inner approximation of $\mathcal{D}$ can then be generated as the convex hull of the extreme points thus obtained. The resulting approximation can be useful, for instance, to design schemes for the commutation between different RMPC controllers, as well as to choose an appropriate initial condition during the planning stage of a control manoeuvre.

It is worth noting that the initial feasibility of the RMPC optimization problem guarantees that the state trajectory will converge to the origin, but does not ensure that the performance will be acceptable. Therefore, it would also be of value to characterize a set $\mathcal{D}^{\bar{\gamma}}$ of initial conditions for which the resulting cost function value is smaller or equal to a given scalar $\bar{\gamma} > 0$. In this work, such a set will be termed "region of guaranteed cost". As an additional contribution of the present paper, the symmetry and convexity properties of $\mathcal{D}$ are also demonstrated for $\mathcal{D}^{\bar{\gamma}}$. Thus, an inner approximation of $\mathcal{D}^{\bar{\gamma}}$ can also be obtained by solving a modified version of the original RMPC optimization problem.

For illustration, a numerical simulation model of an angular positioning system is employed, as in [1]. The remaining of this paper is organized as follows. Section II describes the LMI-based RMPC formulation adopted in the present work. Section III demonstrates the symmetry and convexity properties of the largest domain of attraction $\mathcal{D}$ and proposes a procedure for obtaining extreme points of $\mathcal{D}$ through the solution of an SDP problem. The corresponding developments for regions of guaranteed cost $\mathcal{D}^{\bar{\gamma}}$ are derived in Section IV. The numerical example is discussed in Section V. Finally, concluding remarks are presented in Section VI.

## II. Robust model predictive control employing linear matrix inequalities

The RMPC approach under consideration is concerned with uncertain state-space models of the form:

$$x(k+1) = A(k)x(k) + B(k)u(k),\ [A(k),\, B(k)] \in \Omega \quad (1)$$

where $x(k) \in \mathbb{R}^{n_x}$, $u(k) \in \mathbb{R}^{n_u}$ are the state and input variables, respectively, and $\Omega$ is an uncertainty polytope with known vertices $A_i \in \mathbb{R}^{n_x \times n_x}$, $B_i \in \mathbb{R}^{n_x \times n_u}$, $i = 1, 2, \ldots, L$. It is assumed that component-wise amplitude constraints are to be imposed on the inputs $u(k)$, as well as on $n_y$ output variables defined as $y_l(k) = C_l x(k)$ $(l = 1, 2, \ldots, n_y)$, where $C_l \in \mathbb{R}^{1 \times n}$ are known matrices.

Let $J_\infty(k)$ denote the following infinite-horizon cost function:

$$J_\infty(k) = \sum_{j=0}^{\infty} \left[ ||x(k+j|k)||_S^2 + ||u(k+j|k)||_R^2 \right] \quad (2)$$

where $S \in \mathbb{R}^{n_x \times n_x}$ and $R \in \mathbb{R}^{n_u \times n_u}$ are positive-definite weight matrices and $(\bullet|k)$ denotes a predicted value, which is computed on the basis of the information available at time $k$. It is assumed that the system state is directly measured, so that $x(k|k) = x(k)$.

The optimization problem to be solved at time $k$ can be formulated as

$$\min_{u(k+j|k),\ i\geq 0} \quad \max_{[A(k),\, B(k)]\in\Omega} J_\infty(k) \quad (3)$$

subject to

$$|u_r(k+j|k)| \leq u_{r,max},\ r = 1, 2, \ldots, n_u,\ j \geq 0 \quad (4)$$

$$|y_l(k+j|k)| \leq y_{l,max},\ l = 1, 2, \ldots, n_y,\ j \geq 1 \quad (5)$$

where $u_{r,max}$, $y_{l,max}$ denote the bounds on the magnitude of the $r$th input and $l$th output variables, respectively.

As demonstrated in [1], an upper bound $\gamma$ on the cost $J_\infty(k)$ is minimized by solving the following semidefinite programming (SDP) problem:

$$\min_{\gamma, Q>0, Y, X, Z} \gamma \quad (6)$$

subject to[1]

$$\begin{bmatrix} Q & x(k) \\ * & 1 \end{bmatrix} \geq 0 \quad (7)$$

$$\begin{bmatrix} Q & 0 & 0 & A_iQ + B_iY \\ * & \gamma I & 0 & S^{1/2}Q \\ * & * & \gamma I & R^{1/2}Y \\ * & * & * & Q \end{bmatrix} \geq 0,\ i = 1, 2, \ldots, L \quad (8)$$

$$\begin{bmatrix} X & Y \\ * & Q \end{bmatrix} \geq 0 \quad (9)$$

$$X_{rr} \leq u_{r,max}^2,\ r = 1, 2, \ldots, n_u \quad (10)$$

$$\begin{bmatrix} Z & C_l(A_iQ + B_iY) \\ * & Q \end{bmatrix} \geq 0, \quad \begin{array}{l} l = 1, 2, \ldots, n_y, \\ i = 1, 2, \ldots, L \end{array} \quad (11)$$

$$Z_{ll} \leq y_{l,max}^2,\ l = 1, 2, \ldots, n_y, \quad i = 1, 2, \ldots, L \quad (12)$$

and then using a state feedback control law $u(k+j|k) = Fx(k+j|k)$ over the prediction horizon, with $F = YQ^{-1}$. It is worth noting that the solution of this optimization problem depends on the present state $x(k)$, which appears in the first LMI (7). Therefore, in what follows the SDP given by (6) - (12) will be termed $\mathbb{P}(x(k))$.

By applying the control law in a receding horizon manner, i.e. by solving $\mathbb{P}(x(k))$ in order to obtain a new gain matrix $F$ at each sampling time $k$, the closed-loop system can be shown to be robustly asymptotically stable, provided that $\mathbb{P}(x(k))$ is feasible at the initial time $k = 0$ [1]. Henceforth, with a slight abuse of language, an initial condition $x(0) = \xi \in \mathbb{R}^{n_x}$ will be termed "feasible" if the optimization problem $\mathbb{P}(\xi)$ is feasible, i.e. if there exists a feasible solution $(\gamma, Q > 0, Y, X, Z)$ to the system of LMIs (7) - (12) with $x(k)$ replaced with $\xi$.

## III. Domain of attraction

The RMPC controller described in the previous section is a regulator that steers the system state $x(k)$ to the origin, starting from a given initial condition $x(0)$. Asymptotic stability and constraint satisfaction are guaranteed, provided that the optimization problem is feasible at the initial time $k = 0$ [1]. Therefore, the largest domain of attraction $\mathcal{D}$ of the origin for the closed-loop system can be defined as the set of all feasible initial conditions $x(0) \in \mathbb{R}^{n_x}$. Henceforth, the term "largest"

[1] Symbol $*$ is used to represent the elements below the main diagonal of a symmetric matrix.

will be omitted for brevity. In what follows, some properties of $\mathcal{D}$ will be established.

**Proposition 1** (Symmetry of $\mathcal{D}$)**.** The domain of attraction $\mathcal{D}$ is symmetric about the origin, i.e. if $\xi \in \mathcal{D}$, then $-\xi \in \mathcal{D}$.

*Proof:* Initially, it should be noted that $x(k)$ only appears in the first LMI (7) in the definition of $\mathbb{P}(x(k))$. By using the Schur complement [8], [2], the LMI (7) with $Q > 0$ is seen to be equivalent to $1 - x^T(k)Q^{-1}x(k) \geq 0$. If this inequality is satisfied with $x(k) = \xi$, it is also satisfied with $x(k) = -\xi$. Now, assume that $\xi$ is an element of $\mathcal{D}$. By definition, there exists a feasible solution $(\gamma_\xi, Q_\xi, Y_\xi, X_\xi, Z_\xi)$ to $\mathbb{P}(\xi)$. It can then be seen that $(\gamma_\xi, Q_\xi, Y_\xi, X_\xi, Z_\xi)$ is also a feasible solution to $\mathbb{P}(-\xi)$, which shows that $-\xi$ is an element of $\mathcal{D}$. ■

**Proposition 2** (Convexity of $\mathcal{D}$)**.** The domain of attraction $\mathcal{D}$ is a convex set, i.e. if $\xi_1$, $\xi_2 \in \mathcal{D}$, then $\lambda\xi_1 + (1-\lambda)\xi_2 \in \mathcal{D}$ for any $\lambda \in [0, 1]$.

*Proof:* Let $\xi_1$, $\xi_2$ be two elements of $\mathcal{D}$. Then, by definition, there exist feasible solutions $(\gamma_1, Q_1 > 0, Y_1, X_1, Z_1)$ and $(\gamma_2, Q_2 > 0, Y_2, X_2, Z_2)$ to the system of LMIs (7) - (12) with $x(k)$ replaced with $\xi_1$ and $\xi_2$, respectively. Now, let $\xi_3 = \lambda\xi_1 + (1-\lambda)\xi_2$ and $(\gamma_3, Q_3, Y_3, X_3, Z_3) = \lambda(\gamma_1, Q_1, Y_1, X_1, Z_1) + (1-\lambda)(\gamma_2, Q_2, Y_2, X_2, Z_2)$, with $\lambda \in [0, 1]$. It follows that $Q_3 = \lambda Q_1 + (1-\lambda)Q_2 > 0$. Moreover:

$$\begin{bmatrix} Q_3 & \xi_3 \\ * & 1 \end{bmatrix} = \begin{bmatrix} \lambda Q_1 + (1-\lambda)Q_2 & \lambda\xi_1 + (1-\lambda)\xi_2 \\ * & 1 \end{bmatrix} = \lambda \underbrace{\begin{bmatrix} Q_1 & \xi_1 \\ * & 1 \end{bmatrix}}_{\geq 0} + (1-\lambda) \underbrace{\begin{bmatrix} Q_2 & \xi_2 \\ * & 1 \end{bmatrix}}_{\geq 0} \geq 0 \qquad (13)$$

$$\begin{bmatrix} Q_3 & 0 & 0 & A_iQ_3 + B_iY_3 \\ * & \gamma I & 0 & S^{1/2}Q_3 \\ * & * & \gamma I & R^{1/2}Y_3 \\ * & * & * & Q_3 \end{bmatrix} = \begin{bmatrix} \lambda Q_1 + (1-\lambda)Q_2 & 0 & 0 & A_i[\lambda Q_1 + (1-\lambda)Q_2] + B_i[\lambda Y_1 + (1-\lambda)Y_2] \\ * & \gamma I & 0 & S^{1/2}[\lambda Q_1 + (1-\lambda)Q_2] \\ * & * & \gamma I & R^{1/2}[\lambda Y_1 + (1-\lambda)Y_2] \\ * & * & * & \lambda Q_1 + (1-\lambda)Q_2 \end{bmatrix}$$

$$= \lambda \underbrace{\begin{bmatrix} Q_1 & 0 & 0 & A_iQ_1 + B_iY_1 \\ * & \gamma I & 0 & S^{1/2}Q_1 \\ * & * & \gamma I & R^{1/2}Y_1 \\ * & * & * & Q_1 \end{bmatrix}}_{\geq 0} + (1-\lambda) \underbrace{\begin{bmatrix} Q_2 & 0 & 0 & A_iQ_2 + B_iY_2 \\ * & \gamma I & 0 & S^{1/2}Q_2 \\ * & * & \gamma I & R^{1/2}Y_2 \\ * & * & * & Q_2 \end{bmatrix}}_{\geq 0} \geq 0,$$

$$i = 1, 2, \ldots, L \qquad (14)$$

$$\begin{bmatrix} X_3 & Y_3 \\ * & Q_3 \end{bmatrix} = \begin{bmatrix} \lambda X_1 + (1-\lambda)X_2 & \lambda Y_1 + (1-\lambda)Y_2 \\ * & \lambda Q_1 + (1-\lambda)Q_2 \end{bmatrix} = \lambda \underbrace{\begin{bmatrix} X_1 & Y_1 \\ * & Q_1 \end{bmatrix}}_{\geq 0} + (1-\lambda) \underbrace{\begin{bmatrix} X_1 & Y_1 \\ * & Q_1 \end{bmatrix}}_{\geq 0} \geq 0 \qquad (15)$$

$$X_{3,rr} = \lambda \underbrace{X_{1,rr}}_{\leq u^2_{r,max}} + (1-\lambda) \underbrace{X_{2,rr}}_{\leq u^2_{r,max}} \leq u^2_{r,max}, \quad r = 1, 2, \ldots, n_u \qquad (16)$$

$$\begin{bmatrix} Z_3 & C_l(A_iQ_3 + B_iY_3) \\ * & Q_3 \end{bmatrix} = \lambda \underbrace{\begin{bmatrix} Z_1 & C_l(A_iQ_1 + B_iY_1) \\ * & Q_1 \end{bmatrix}}_{\geq 0} + (1-\lambda) \underbrace{\begin{bmatrix} Z_2 & C_l(A_iQ_2 + B_iY_2) \\ * & Q_2 \end{bmatrix}}_{\geq 0} \geq 0 \qquad (17)$$

$$Z_{3,ll} = \lambda \underbrace{Z_{1,ll}}_{\leq y^2_{l,max}} + (1-\lambda) \underbrace{Z_{2,ll}}_{\leq y^2_{l,max}} \leq y^2_{l,max}, \quad l = 1, 2, \ldots, n_y, \quad i = 1, 2, \ldots, L \qquad (18)$$

Therefore, $(\gamma_3, Q_3 > 0, Y_3, X_3, Z_3)$ is a feasible solution to the system of LMIs (7) - (12) with $x(k)$ replaced with $\xi_3$, which shows that $\xi_3 \in \mathcal{D}$. ■

Given the properties of convexity and symmetry about the origin, extreme points of $\mathcal{D}$ can be obtained by solving an SDP problem of the form

$$\min_{\beta, \gamma, Q > 0, Y, X, Z} \beta \qquad (19)$$

subject to

$$\begin{bmatrix} Q & \beta\xi \\ * & 1 \end{bmatrix} \geq 0 \qquad (20)$$

and the remaining LMIs (8) – (12) of the original RMPC optimization problem. In (20), $\xi \in \mathbb{R}^{n_x}$ is a constant vector that defines the direction along which the extreme point is to be found. The extreme point will be given by $\beta^*\xi$, where $\beta^*$ is the minimal value of $\beta$ resulting from the optimization process, provided that $\mathcal{D}$ is bounded along the direction of $\xi$. It is worth noting that $-\beta^*\xi$ will also be an extreme point, due to the symmetry property. Fig. 1a illustrates the process of obtaining an extreme point of $\mathcal{D}$ in a two-dimensional case ($n_x = 2$).

By solving this SDP problem with different $\xi$ vectors, a number of extreme points of $\mathcal{D}$ can be obtained. An inner approximation of $\mathcal{D}$ can then be generated as the convex hull of those points, as illustrated in Fig. 1b.

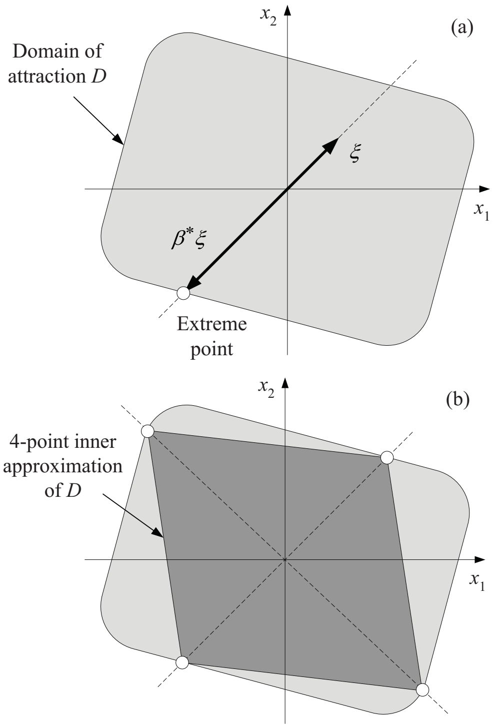


Fig. 1. (a) Determination of an extreme point of $\mathcal{D}$ through the minimization of $\beta$. (b) Inner approximation of $\mathcal{D}$ obtained as the convex hull of four extreme points.

## IV. Regions of Guaranteed Cost

The domain of attraction $\mathcal{D}$ involves only the feasibility of the RMPC optimization problem (6) – (12), regardless of the achievable value for the cost function $\gamma$. By including a constraint on $\gamma$, it is possible to characterize a region of initial conditions for which the cost is guaranteed to be smaller than a certain bound, as defined below.

**Definition 1** (Region of guaranteed cost)**.** Given a value of $\bar{\gamma} > 0$, the region $\mathcal{D}^{\bar{\gamma}}$ is defined as the set of initial conditions $x(0) \in \mathcal{D}$ for which the optimal solution $\gamma^*$ obtained by solving $\mathbb{P}(x(0))$ is smaller or equal to $\bar{\gamma}$.

The symmetry and convexity of $\mathcal{D}^{\bar{\gamma}}$ are established in the two propositions below.

**Proposition 3** (Symmetry of $\mathcal{D}^{\bar{\gamma}}$)**.** The region of guaranteed cost $\mathcal{D}^{\bar{\gamma}}$ is symmetric about the origin, for any given $\bar{\gamma} > 0$.

*Proof:* Symmetry can be demonstrated by applying the Schur complement to LMI (7), as in the proof of Proposition 1. ■

**Proposition 4** (Convexity of $\mathcal{D}^{\bar{\gamma}}$)**.** The region of guaranteed cost $\mathcal{D}^{\bar{\gamma}}$ is a convex set, for any given $\bar{\gamma} > 0$.

*Proof:* Let $\xi_1 \in \mathbb{R}^{n_x}$ and $\xi_2 \in \mathbb{R}^{n_x}$ be two elements of $\mathcal{D}^{\bar{\gamma}}$ for a given $\bar{\gamma} > 0$, and let $(\gamma_1^*, Q_1^*, Y_1^*, X_1^*, Z_1^*)$ and $(\gamma_2^*, Q_2^*, Y_2^*, X_2^*, Z_2^*)$ be the optimal solutions of $\mathbb{P}(\xi_1)$ and $\mathbb{P}(\xi_2)$. From the definition of $\mathcal{D}^{\bar{\gamma}}$, it follows that

$$\gamma_1^* \leq \bar{\gamma},\ \gamma_2^* \leq \bar{\gamma}. \tag{21}$$

Now, let $\xi_3 = \lambda\xi_1 + (1-\lambda)\xi_2$ and $(\gamma_3, Q_3, Y_3, X_3, Z_3) = \lambda(\gamma_1^*, Q_1^*, Y_1^*, X_1^*, Z_1^*) + (1-\lambda)(\gamma_2^*, Q_2^*, Y_2^*, X_2^*, Z_2^*)$, with $\lambda \in [0,1]$. In view of (21), one has $\gamma_3 = \lambda\gamma_1^* + (1-\lambda)\gamma_2^* \leq \bar{\gamma}$. Moreover, a demonstration similar to that of Proposition 2 can be used to prove that $(\gamma_3, Q_3, Y_3, X_3, Z_3)$ is a feasible solution to $\mathbb{P}(\xi_3)$. Finally, let $\gamma_3^*$ be the optimal value of the cost $\gamma$ for $\mathbb{P}(\xi_3)$. Given that the minimal value of the cost must be smaller or equal to the cost of any feasible solution, it follows that $\gamma_3^* \leq \gamma_3 \leq \bar{\gamma}$. Therefore, $\xi_3 \in \mathcal{D}^{\bar{\gamma}}$, which shows that $\mathcal{D}^{\bar{\gamma}}$ is a convex set. ■

Extreme points of $\mathcal{D}^{\bar{\gamma}}$ can be obtained by solving an SDP problem of the form

$$\min_{\beta,\gamma,Q>0,Y,X,Z} \beta \tag{22}$$

subject to

$$\gamma \leq \bar{\gamma} \tag{23}$$

$$\begin{bmatrix} Q & \beta\xi \\ * & 1 \end{bmatrix} \geq 0 \tag{24}$$

and the remaining LMIs (8) – (12) of the original RMPC optimization problem. As in the previous section, the extreme point will be given by $\beta^*\xi$, where $\xi \in \mathbb{R}^{n_x}$ is a constant vector that defines the direction along which the extreme point is to be found. An inner approximation of $\mathcal{D}^{\bar{\gamma}}$ can be generated as the convex hull of extreme points obtained with different $\xi$ vectors.

## V. Numerical example

The angular positioning system described in [1] will be adopted to illustrate the proposed method. The problem under consideration involves the control of a rotating antenna driven by an electric motor. The plant dynamics are described by the following discrete-time state equation:

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \end{bmatrix} = \\ \begin{bmatrix} 1.0 & 0.1 \\ 0 & 1-0.1\alpha(k) \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix} + \begin{bmatrix} 0 \\ 0.0787 \end{bmatrix} u(k) \tag{25}$$

where the state variables $x_1$ and $x_2$ denote the angular position ($rad$) and velocity ($rad/s$) of the antenna, respectively, and the control variable $u$ corresponds to the input voltage ($V$) of the electric motor. The viscous friction in the moving parts of the antenna is associated to the uncertain parameter $\alpha(k)$, which is assumed to be in a given range $[\alpha_{min}, \alpha_{max}]$. Therefore, the model is of the form (1), with $A(k) \in Co\{A_1, A_2\}$ where

$$A_1 = \begin{bmatrix} 1.0 & 0.1 \\ 0 & 1-0.1\alpha_{min} \end{bmatrix},\ A_2 = \begin{bmatrix} 1.0 & 0.1 \\ 0 & 1-0.1\alpha_{max} \end{bmatrix} \tag{26}$$

It is worth noting that the actual dependence of $\alpha(k)$ on the time $k$ does not affect the RMPC formulation, because the LMIs only involve the vertices $A_1$, $A_2$ of the uncertainty polytope. In what follows, the uncertain parameter will be denoted simply by $\alpha$ for brevity.

The cost function weights $S$ and $R$ were set to $I_{2\times 2}$ and 1, respectively. Moreover, the control variable $u$ was constrained to the range $[-2V, +2V]$, as in [1], and the position $x_1$ was constrained to the range $[-1rad, +1rad]$. Such a constraint can be cast into the form (5) by defining an output variable $y = Cx$, with $C = [1\ 0]$. All numerical results were obtained by using the LMI Lab package for Matlab.

### A. Results

Figure 2a presents the inner approximations of the attraction domain $\mathcal{D}$ obtained by using 4 and 16 extreme points. In this case, the bounds on the uncertain parameter $\alpha$ were set to $\alpha_{min} = 0.1$ and $\alpha_{max} = 10$. Moreover, Fig. 2a shows a grid of states that were employed to test the feasibility of the original SDP problem (6) - (12). As can be seen, all grid points inside the obtained polygons correspond to feasible initial conditions. It is worth noting that points $[-1\ 0]^T$ and $[+1\ 0]^T$, which correspond to unfeasible initial conditions, are outside the polygons, as shown in Fig. 2b. The use of 16 extreme points provides a better approximation of the attraction domain, in that the resulting polygon encompasses a larger region of feasible initial conditions, as compared to the 4-point approximation.

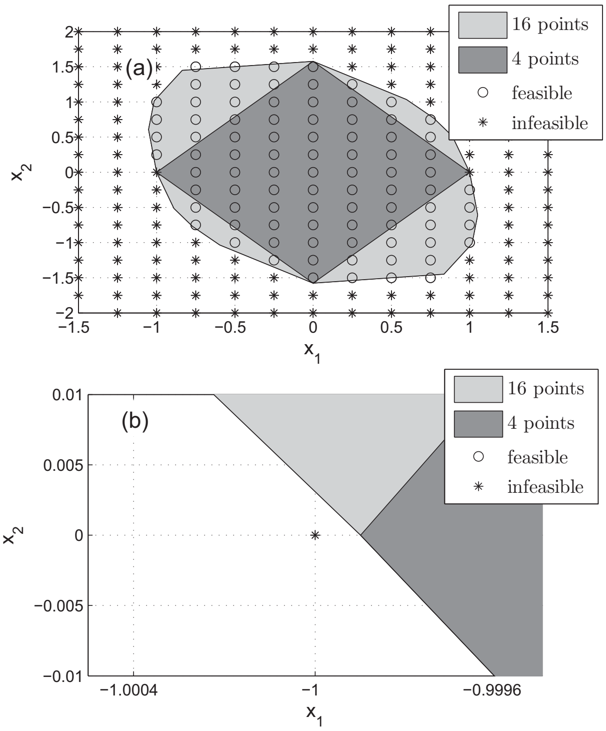


Fig. 2. (a) Inner approximations of the attraction domain $\mathcal{D}$. (b) Detail of the region around point $[-1\ 0]^T$.

From a physical point of view, unfeasibility arises if the position $x_1$ is close to the ±1 bounds and the velocity is such that the position is changing towards the bound. It is interesting to notice that part of the 16-point polygon is located outside the $[-1, +1]$ range of admissible values for $y = x_1$. This result can be explained by noting that the output constraints in (5) are only enforced after one time step ahead of the present time. Therefore, if the initial condition is such that the output $y$ can be steered to the admissible range in a single time step, the output constraint in (5) will be satisfied.

An interesting investigation that could be performed at this point concerns the relation between the domain of attraction $\mathcal{D}$ and the range of values for the uncertain model parameter $\alpha$. It is expected that $\mathcal{D}$ will be larger if the characterization of $\alpha$ is more precise, i.e., if the range of uncertainty is smaller. To investigate this issue, the proposed method was employed to obtain new extreme points of $\mathcal{D}$, with bounds on $\alpha$ set to $\alpha_{min} = 0.5$ and $\alpha_{max} = 2$. The results are presented in Fig. 3. As expected, the polygon obtained for $0.5 \leq \alpha \leq 2$ contains the polygon obtained for the wider range $0.1 \leq \alpha \leq 10$.

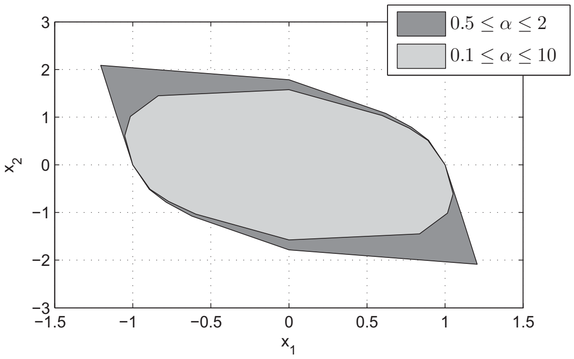


Fig. 3. Inner approximations of the attraction domain $\mathcal{D}$ for two different uncertainty ranges.

Figure 4 presents the inner approximations of the region of guaranteed cost $\mathcal{D}^{\bar{\gamma}}$ obtained by using 4 and 16 extreme points with $\bar{\gamma} = 100$. In this case, the bounds on the uncertain parameter $\alpha$ were again set to $\alpha_{min} = 0.1$ and $\alpha_{max} = 10$. The grid of states shown in this figure was employed to solve the SDP problem (6) - (12) in order to obtain the minimal value of $\gamma$, which is denoted by $\gamma^*$. As expected, all grid points inside the obtained polygons correspond to initial conditions for which $\gamma^* \leq \bar{\gamma} = 100$. As in Fig. 2a, the use of 16 extreme points provides a better approximation of the region under consideration.

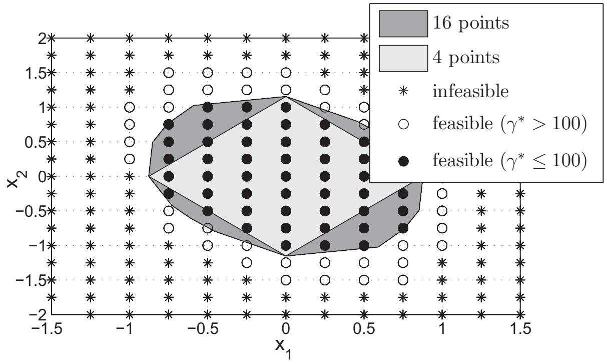


Fig. 4. Inner approximations of the region of guaranteed cost $\mathcal{D}^{\bar{\gamma}}$ for $\bar{\gamma} = 100$.

Figure 5 shows the 16-point polygons obtained for different values of $\bar{\gamma}$. As can be seen from these inner approximations, the region of guaranteed cost $\mathcal{D}^{\bar{\gamma}}$ tends to increase with $\bar{\gamma}$. It can also be seen that, as $\bar{\gamma}$ is increased, $\mathcal{D}^{\bar{\gamma}}$ converges to the domain of attraction $\mathcal{D}$. Indeed, if $\bar{\gamma}$ is made arbitrarily large, the SDP problem becomes equivalent to that involved in the determination of extreme points for $\mathcal{D}$ (i.e. without the $\gamma$ constraint in (23)).

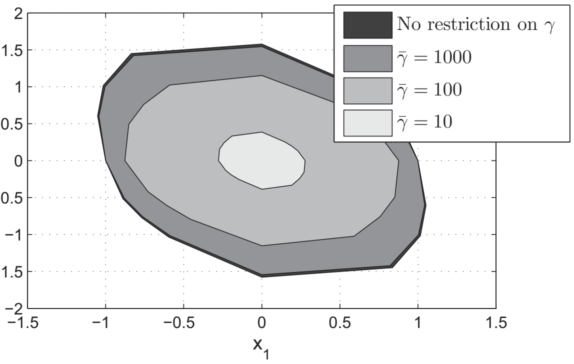


Fig. 5. Inner approximations of the region of guaranteed cost $\mathcal{D}^{\bar{\gamma}}$ for different values of $\bar{\gamma}$. The region obtained with no restriction on $\gamma$ is an inner approximation of $\mathcal{D}$.

Finally, Fig. 6 presents the 16-point polygonal approximations to the region $\mathcal{D}^{\bar{\gamma}}$ for $\bar{\gamma} = 100$ and two different ranges for the uncertain parameter $\alpha$, namely $0.1 \leq \alpha \leq 10$ and $0.5 \leq \alpha \leq 2$. As in the case of the attraction domain, the region of guaranteed cost increases as the uncertainty range is reduced.

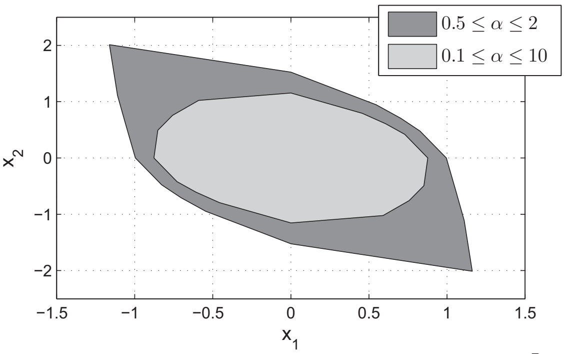


Fig. 6. Inner approximations of the region of guaranteed cost $\mathcal{D}^{\bar{\gamma}}$ for $\bar{\gamma} = 100$ and two different uncertainty ranges.

## VI. Concluding Remarks

This paper demonstrated some properties of the domain of attraction $\mathcal{D}$ and regions of guaranteed cost $\mathcal{D}^{\bar{\gamma}}$ for the LMI-based RMPC formulation originally proposed in [1]. More specifically, $\mathcal{D}$ and $\mathcal{D}^{\bar{\gamma}}$ were shown to be convex and symmetric about the origin, which allowed the determination of extreme points through the solution of modified versions of the original SDP problem involved in the RMPC formulation. Inner approximations of these sets could then be generated as the convex hull of the extreme points. In the numerical example presented for illustration, such approximations were found to be in agreement with the feasibility and cost results obtained in a pointwise manner for a grid of initial conditions.

Future investigations could be concerned with the extension of the present work to other LMI-based RMPC formulations with less conservatism, such as those proposed in [17], [18], [19], [20]. It is expected that a reduction in conservatism should lead to an enlargement in the domain of attraction, as well as the regions of guaranteed cost, which would be an additional advantage of those formulations with respect to [1].

## Acknowledgements

The authors acknowledge the support of FAPESP (research grant 2011/17610-0 and doctoral scholarship 09/52400-7), CAPES (Pró-Engenharias MSc scholarship) and CNPq (research fellowship).

## References

[1] M. V. Kothare, V. Balakrishnan, and M. Morari, “Robust constrained model predictive control using linear matrix inequalities,” *Automatica*, vol. 32, no. 10, pp. 1361–1379, 1996.

[2] J. M. Maciejowski, *Predictive control with constraints*. Harlow: Prentice Hall, 2002.

[3] S. J. Qin and T. A. Badgwell, “A survey of industrial model predictive control technology,” *Control Engineering Practice*, vol. 11, pp. 733–764, 2003.

[4] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. M. Scokaert, “Constrained model predictive control: Stability and optimality,” *Automatica*, vol. 36, pp. 789–814,, 2000.

[5] P. J. Campo and M. Morari, “Robust model predictive cotrol,” *In Proc. American Control Conf., Minneapolis, MN*, pp. 1021–1026, 1987.

[6] J. C. Allwright and G. C. Papavasiliou, “On linear programming and robust model-predictive control using impulse-responses,” *Syst. Control Lett.*, vol. 18, pp. 159–164, 1992.

[7] Z. Q. Zheng and M. Morari, “Robust stability of constrained model predictive control,” *In Proc. American Control Conf., San Francisco, CA*, pp. 379–383, 1993.

[8] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*. Philadelphia: SIAM, 1994.

[9] Z. Wan and M. V. Kothare, “Robust output feedback model predictive control using off-line linear matrix inequalities,” *Journal of Process Control*, vol. 12, pp. 763–774, 2002.

[10] ——, “A framework for design of scheduled output feedback model predictive control,” *Journal of Process Control*, vol. 18, pp. 258–264, 2008.

[11] B. Ding, Y. Xi, M. T. Cychowski, and T. O’Mahony, “A synthesis approach for output feedback robust constrained model predictive control,” *Automatica*, vol. 44, pp. 258–264, 2008.

[12] Z. Wan and M. V. Kothare, “Efficient scheduled stabilizing model predictive control for constrained nonlinear systems,” *International Journal of Robust and Nonlinear Control*, vol. 13, pp. 331–346, 2003.

[13] B. Ding and B. Huang, “Constrained robust model predictive control for time-delay systems with polytopic description,” *International Journal of Control*, vol. 80, no. 4, pp. 509–522, 2007.

[14] B. Ding, L. Xie, and W. Cai, “Robust MPC for polytopic uncertain systems with time-varying delays,” *International Journal of Control*, vol. 81, no. 8, pp. 1239–1252, 2008.

[15] B. Ding, “Robust model predictive control for multiple time delay systems with polytopic uncertainty description,” *International Journal of Control*, vol. 83, no. 9, pp. 1844–1857, 2010.

[16] M. S. M. Cavalca, R. K. H. Galvão, and T. Yoneyama, “Robust model predictive control using linear matrix inequalities for the treatment of asymmetric output constraints,” *Journal of Control Science and Engineering*, pp. 1–7, Article ID 485784, 2012.

[17] F. A. Cuzzola, J. C. Geromel, and M. Morari, “An improved approach for constrained robust model predictive control,” *Automatica*, vol. 38, pp. 1183–1189, 2002.

[18] N. Wada, K. Saito, and M. Saeki, “Model predictive control for linear parameter varying systems using parameter dependent Lyapunov function,” *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 53, pp. 1446 – 1450, 2006.

[19] S. Lee and J. H. Park, “Output feedback model predictive control for LPV systems using parameter-dependent Lyapunov function,” *Applied Mathematics and Computation*, vol. 190, pp. 671–676, 2007.

[20] W. J. Mao, “Robust stabilization of uncertain time-varying discrete systems and comments on “an improved approach for constrained robust model predictive control”,” *Automatica*, vol. 39, pp. 1109–1112, 2003.

[21] Z. Wan and M. V. Kothare, “An efficient off-line formulation of robust model predictive control using linear matrix inequalities,” *Automatica*, vol. 39, pp. 837–846, 2003.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# A Control Method to Eliminate Polarization-Induced Phase Distortion in Dual Mach-Zehnder Fiber Interferometer

Yang An[1] , Hao Feng[1] , Yan Zhou[2] , Shi-jiu Jin[1] , Zhou-mo Zeng[1]
1 State Key Laboratory of Precision Measuring Technology & Instruments, Tianjin University, Tianjin, 300072, China
2 R and D Center of PetroChina, Pipeline Company, LangFang, HeBei, 065000, China
Email: fhlele256@tju.edu.cn

*Abstract*—**Polarization-induced Phase distortion happens in the distributed dual Mach-Zehnder fiber interferometric sensing system, which results in big locating errors. Based on the investigation in the cause, we propose and demonstrate a scheme using automatic polarization controller together with simulated annealing as control algorithm to obtain high positioning accuracy in the system. Through laboratory experiments, the key parameters of the algorithm are analyzed and the optimal settings of them are fixed. Field trial results show that this method can make a fast search for the optimal state of polarization and maintain the strong signal correlation for a long time.**

***Keywords-Distributed sensor, polarization control, Mach-Zehnder fiber interferometer, polarization, simulated annealing***

## I. INTRODUCTION

With the exploitation of oil and gas resources as well as the surge in energy demand, the total pipeline mileage is growing constantly, and pipeline safety has become an important research topic. Our group has developed the distributed oil and gas pipeline leak detection and pre-warning system based on dual Mach-Zehnder Optical Fiber Interferometer for detecting and locating intrusions which may threaten pipeline safety[1]. Positioning the illegal invasion, as a key technology in the system, uses cross-correlation function to estimate time delay which requires strong correlation between signals. However, due to the phenomenon of polarization-induced fading which can cause phase distortion, correlation bewteen detection singals often degenerates and big locating errors then arise in pratical applications[2].

To solve the problem, passive phase demodulation method was studied in previous papers, in which the vibration phase signal are demodulated from every two original signals and thus avoid the negative effect of polarization-induced phase shift.[3] However, as this method costs much time in signal processing, it's hardly suitable in practical applications which demand good real-time performance. In this paper we propose and demonstrate a new method based on polarization control to eliminate phase distortion and furthermore to increase positioning accuracy of the system.

The existing polarization control algorithms mainly aim to eliminate the polarization mode dispersion(PMD) in the optical fiber and polarization dependent modulation(PDM) in the electrooptical modulator(EOM)[4],[5]. These feedback control algorithms are based on the peak value search of light intensity and can get control effect instantly with the output of control words. In the dual Mach-Zehnder interferometric system, however, the polarization control objective is the correlation between two detection signals which is not directly related to the light intensity. Therefore, the above mentioned algorithms can not apply to the pipeline security monitoring system. In this paper, the simulated annealing(SA) based on the correlation coefficient between two detection signals is employed as the polarization control algorithm to search for the input polarization operating point which results in little degeneration of detection signals' correlation. Field experiments show that the algorithm can rapidly find the operating point and continuously stabilize the correlation between the system detection signals.

## II. RESEARCH ON THE CAUSE

Fig. 1 shows the equivalent optical path of the dual Mach-Zehnder fiber interferometer sensing system.

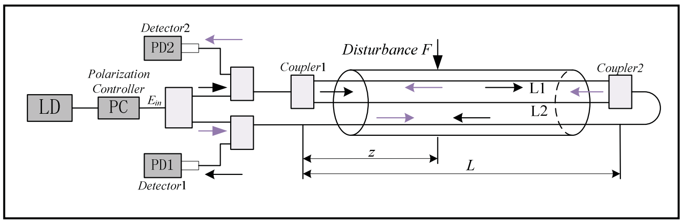


Figure 1. Conventional diagram of dual Mach-Zehnder system structure

The system input light can be represented by the Jones vector:

$$E_{in} = \begin{bmatrix} E_x \\ E_y \end{bmatrix} = E_0 \begin{bmatrix} \cos\theta\cos\varepsilon - j\sin\theta\sin\varepsilon \\ \sin\theta\cos\varepsilon + j\cos\theta\sin\varepsilon \end{bmatrix} \qquad (1)$$

Where $E_0$ denotes the lightwave amplitude, $\theta$ and $\varepsilon$ are the azimuth and ellipticity angle respectively which determine the input polarization state together.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

Assuming that the forward and backward equivalent Jones Matrices of the two sensing arms are separately $E_1$,$E_2$ and $E_1^T$ ,$E_2^T$ which are determined by polarization property of the sensing fiber, the optical signals detected by photodetectors PD1 and PD2 can be represented by:

$$\begin{cases} E_{out1}=\left(E_1+E_2\cdot e^{j\delta(t)}\right)\cdot E_{in} \\ E_{out2}=\left(E_1^T\cdot e^{j\delta(t)}+E_2^T\right)\cdot E_{in} \end{cases} \tag{2}$$

Where $\delta(t)$ is the phase retardation difference between two sensor fibers caused by disturbance. The light intensity signals $I_1(t)$、$I_2(t)$ can be represented by:

$$\begin{cases} I_1(t)=\left|E_{out1x}\right|^2+\left|E_{out1y}\right|^2=I_0\cos\left[\delta(t)+f_1\left(\theta,\varepsilon,\alpha\right)\right] \\ I_2(t)=\left|E_{out2x}\right|^2+\left|E_{out2y}\right|^2=I_0\cos\left[\delta(t)+f_2\left(\theta,\varepsilon,\beta\right)\right] \end{cases} \tag{3}$$

Where $I_0$ is the light intensity of light source, $\alpha$ and $\beta$ are the polarization induced phase shift caused by fiber birefringence respectively in forward and backward fiber path.

Taking the length of time $t$ of the detected signals, the correlation coefficient between two channel signals is calculated by the following formula:

$$\rho_{xy}=\frac{\int_0^t I_1(t)I_2(t)dt}{\left[\int_0^t I_1^2(t)dt\int_0^t I_2^2(t)dt\right]^{\frac{1}{2}}}=\frac{\int_0^t \cos^2\left(\delta(t)+f_1\left(\theta,\varepsilon,\alpha\right)\right)\cdot\cos^2\left(\delta(t)+f_2\left(\theta,\varepsilon,\beta\right)\right)dt}{\left[\int_0^t \cos^2\left(\delta(t)+f_1\left(\theta,\varepsilon,\alpha\right)\right)dt\cdot\int_0^t \cos^2\left(\delta(t)+f_2\left(\theta,\varepsilon,\beta\right)\right)dt\right]^{\frac{1}{2}}} \tag{4}$$

The correlation coefficient $\rho_{xy}$ is mainly determined by the phases of two detected signals, so it's feasible to alter the correlation between signals through changing the additional phase differences $f_1(\theta,\varepsilon,\alpha)$ and $f_2(\theta,\varepsilon,\beta)$. Since the additional phase difference are determined together by the input polarization state, the polarization property of fiber and the system laying environment in which the latter two are difficult to adjust artificially, it can only be changed by altering the input polarization state.

Fig. 2 shows the various interference light intensity in different input light polarization state which are respectively linearly polarized($\theta$=0.25π, $\varepsilon$=0), elliptically polarized($\theta$=0.25π, $\varepsilon$=0.25π), and circularly polarized($\theta$=0.25π, $\varepsilon$=0.125π) and $\delta(t)$ is assumed changing from -2π to 2π. From Fig 2, we can confirm that not only the amplitude but also the phase of the signal varies in different input polarization states. It's conceivable that there must be one input polarization state or several of them which can decrease the additional phase difference to an acceptable range. Therefore, it's feasible to improve the correlation between the detection signals by controlling the polarization state of input light and searching for the specific one as mentioned above.

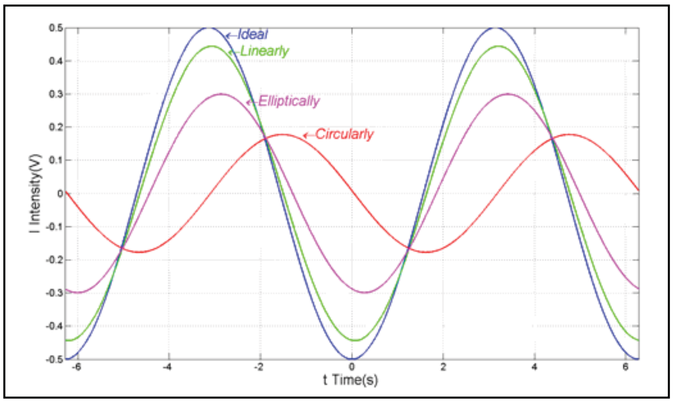


Figure 2. The interference light intensities in different input polarization states

## III. ALGORITHM CLARIFICATION

### *A. Algorithm Principle*

Setting correlation coefficient $\rho_{xy}$ as the objective function, the polarization control model can be represented by:

$$\max(\rho_{xy})=f(X) \qquad X\in\Theta \tag{5}$$

Where $X$ denotes a combination of the polarization controller's retardation $x_i$(i=1,2,3,4). It's an optimization problem which requires global search capability of the control algorithm. simulated annealing[6]-[8] can probabilistically jump out of local optimum and achieve global optimum according to Metropolis criterion and thus is applied as the feedback control algorithm. The initial solution $X_i$ is randomly generated whose objective function value is $f(X_i)$ and the new solution $X_j$ is generated by state generator function whose objective function value is $f(X_j)$. The probability of accepting the new solution $p_r$ is determined by the Metropolis criterion[7]:

$$p_r\left(X_i\Rightarrow X_j\right)=\begin{cases}1 & f\left(X_j\right)\le f\left(X_i\right) \\ \exp\left(-\dfrac{f\left(X_j\right)-f\left(X_i\right)}{T_k}\right) & f\left(X_j\right)>f\left(X_i\right)\end{cases} \tag{6}$$

Where $T_k$ is the current temperature state. Comparing $p_r$ with a random number which ranges between 0 and 1, if $p_r$ is larger, the new solution $X_j$ is accepted, otherwise $X_i$ is retained. After specified rounds of solution changing, local optimum polarization state under current temperature state can be found. Then $T$ value is gradually reduced according to the temperature update function and local optimum solution under every temperature state can be derived. When $T$ value is close to zero, calculation stops and current optimum solution is global optimum solution. The temperature update function is defined as:

$$T_{k+1}=p\cdot T_k \tag{7}$$

Where the temperature update coefficient $p$ has a range of 0-0.95[8].

Since the phase retardation caused by polarization controller in practice has a range of 0-5π which means the solution space is bounded. We define the state generator function as follows:

$$X_{i+1} = \begin{cases} X_i - f(X_i)\cdot s & X_i \geq 5\pi \\ X_i + f(X_i)\cdot c\cdot s & 0 < X_i < 5\pi \\ X_i + f(X_i)\cdot s & X_i \leq 0 \end{cases} \tag{8}$$

Where *s* is searching step size, *c* is plus or minus one randomly.

### *B. Setting of Key Parameters*

Simulated annealing algorithm performance does not rely on the value of initial solution but can be affected by settings of key parameters such as the initial temperature $T_0$, the search step size *s* in the state generator function and the temperature update coefficient *p*.

In order to obtain the optimum setting of key parameters so as to ensure the global search capability of the control algorithm, experiments have been carried out to test the optimization in different parameter settings.

Fig. 1 shows the installation position of the polarization controller which can control both the forward and backward optical path and avoid changing the system structure as well. The model of the polarization controller applied is polaRITE Ⅲ from General Photonics.

#### *1) The choice of the initial temperature*

In simulated annealing algorithm, higher initial temperature $T_0$ will enhance the probability of obtaining high quality solution but increase the number of outer loop at the same time. With comprehensive consideration of both quality and efficiency of optimization, the range of $T_0$ is set less than one.

Fig. 3 shows the different search areas in Poincare Sphere with different initial temperature settings in stimulated annealing algorithm. When $T_0$ is small, the algorithm can only search part of the region in Poincare Sphere with a large blind area. According to the experiment results, the initial temperature is set as 1.

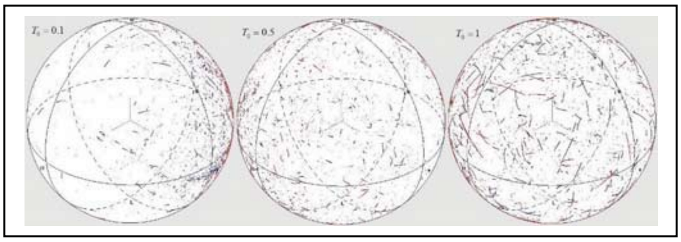


Figure 3. Initial temperature effect on search area, $T_0$=0.1,0.5,1.

#### *2) The choice of temperature update coefficient*

Simulated annealing is a process of temperature slowly decreasing in which higher value of temperature update coefficient *p* will bring about a larger searching area but lower convergence rate .

Fig. 4 shows the different search areas in Poincare Sphere with different settings of *p* in stimulated annealing process. Since they do not have much difference as Fig. 4 shows, taking into account that the higher value of *p* will slow down the convergence speed and then degrade the real-time performance of the system, the temperature update coefficient is set as 0.5.

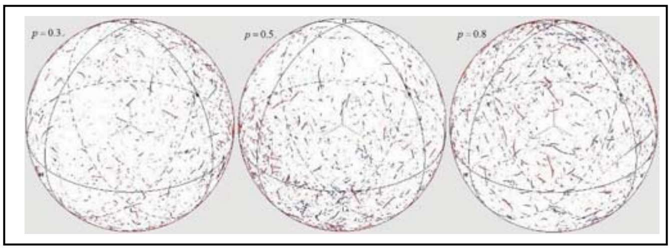


Figure 4. Temperature update coefficient effect on search area, *p*=0.3,0.5,0.8.

#### *3) The choice of search step size*

State generator function determines whether the system can search for the optimal value in the whole solution space. Large step size brings about big change of the solution which may deviate from the solution space, while small one brings about little change of the solution which can hardly jump out of local optimum to achieve global optimum.

Fig. 5 shows the effects of different settings of search step size. When *s* increases to 1.10 can it jump out of local optimum and search the whole Poincare Sphere. Therefore, *s* is set as 1.10.

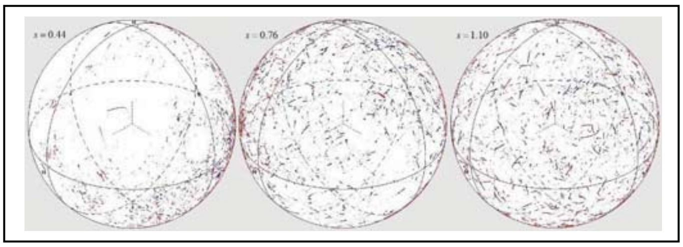


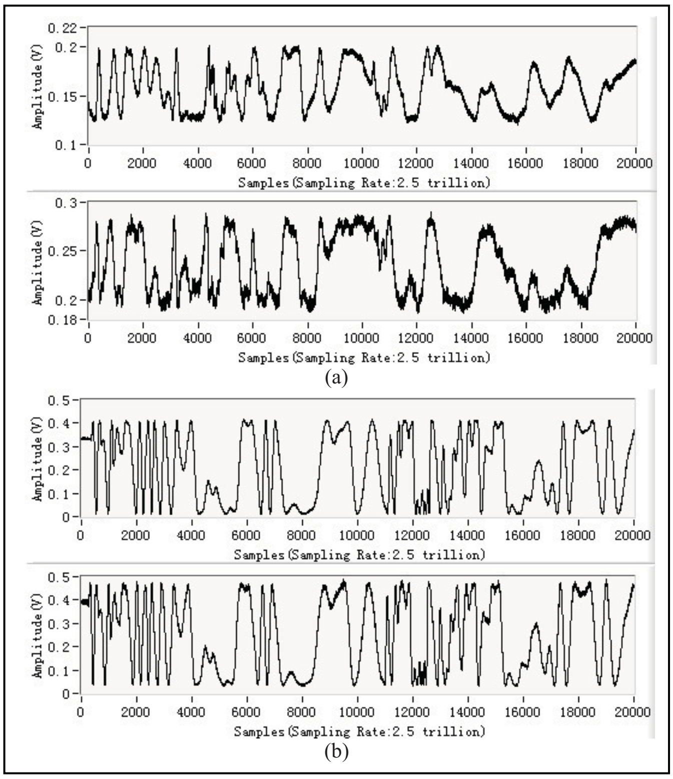
Figure 5. Search step size effect on search area, *s* =0.44,0.76,1.10.

## IV. FIELD EXPERIMENT VERIFICATION

The approach is applied in the oil pipeline safety monitoring and pre-warning system which covers a monitoring distance of 43km.

Fig. 6 shows the contrast between the original detection signal waveforms with and without employment of the polarization control method. When the approach is not applied, the signals in Fig. 4(a) has a calculated correlation coefficient $\rho_{xy}$ =0.28, with employment of the polarization control approach, the correlation coefficient is 0.95 as shown in Fig. 4(b).

Fig. 7 shows the correlation coefficient between detection signals within 24 hours recorded every second after using the polarization control method. which verifies the stability of the approach.



(a)

(b)

Figure 6. Origianl detection signal waveforms. (a)Polarization control is not applied. (b)Polarization control is applied.

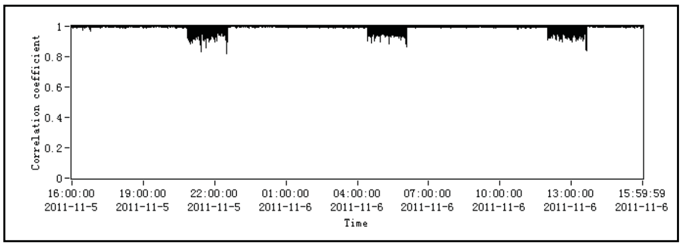


Figure 7. Correlation coefficient within twenty-four hours under polarization control.

## V. CONCLUSION

Based on the investigation in the cause of positioning errors, we present a new approach to solve the degeneracy problem of detection signals' correlation in the distributed dual Mach-Zehnder fiber interferometric sensing system using automatic polarization controller together with simulated annealing as control algorithm. Using correlation coefficient between two detection signals as the feedback quantity, this method can control the input polarization state in real-time and find the operating point rapidly when the correlation between signals degenerates. Through laboratory experiments, the key parameters of the algorithm are analyzed and the optimal settings of them are fixed. Field trial shows significant improvement and stability of the correlation between detection signals after employing the approach.

## REFERENCES

[1] Y. Zhou, S. J. Jin, Z. M. Zeng and H. Feng, "Study on the distributed optical fiber sensing technology for pipeline safety detection and location," *Guangdianzi Jiguang/Journal of Optoelectronics Laser*, vol.19, pp. 922-924, 2008.

[2] A. A. Chtcherbakov and P. L. Swart, "Polarization effects in the Sagnac-Michelson distributed disturbance location sensor," *Lightwave Technology, Journal of*, vol.16, pp. 1404-1412, 1998.

[3] H. Feng, S. J. Jin, Y. An, Z. M. Zeng, and Z. G. Qu, "Phase distortion analysis and passive demodulation for pipeline safety system based on Jones matrix modeling," *Measurement,* vol. 44, pp. 1531-1538, 2011.

[4] W. Rong and L. Dupont, "A feedback algorithm for polarization control using two rotatable wave plates with variable birefringence," *Optics Communications*, vol.259, pp. 603-611, 2006.

[5] M. Martinelli and R. A. Chipman, "Endless polarization control algorithm using adjustable linear retarders with fixed axes," *Lightwave Technology, Journal of*, vol.21, pp. 2089-2096, 2003.

[6] G. B. Gao, W. Wang, K. Lin and Z. C. Chen, "Parameter identification based on modified annealing algorithm for articulated arm CMMs," *Guangxue Jingmi Gongcheng/Optics and Precision Engineering*, vol.17, pp. 2499-2505, 2009.

[7] L. S. Kang, Y. Xie, S. Y. You, Z. H. Luo, Non-numerical Parallel Algorithms(Volume Ⅰ)Simulated Annealing. Beijing: Science Press, 1998.

[8] P. Huang, Optimal Theories and Methods. Beijing: Tsinghua University Press, 2009

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Designing and Finite Element Analysis of customized titanium plate for Mandible

Tianbiao Yu, Guoqiang Ma
Tianhua Wei, Wanshan Wang
School of Mechanical Engineering & Automation
Northeastern University
Shenyang, China
tbyu@mail.neu.edu.cn

Xingjun Qin
Department of Oral and Maxillofacial Surgery
China Medical University
Shenyang, China

*Abstract*—**The Rigid Internal Fixation is one of the commonly used methods to reconstruct the mandible in bone repair surgery. This research proposed a complete flow of the designing and Finite Element Analysis (FEA).The customized titanium plate and screws model were imported in Mimics, then fitted and simulated with the shape of mandible 3D model. The simulation model would be imported into Magic's and the screw holes were located and punched. In order to ensure the validity of model for surgical guidance, the FEA was employed by FEA software Abaqus in this study. Different angles were selected and different loads were loaded. Comparing with the nomal mandible mechanical data, the von mises stress and deformation which was from the reconstruction of mandible were in the regular range and the RP model which would be used in the medical guidence was manufactured according to the 3D model. The research for this processing has important guide significance and good business prospects.**

***Keywords-customized ti-plate, fixation and simulation, FEA in Abaqus, model buliding***

## I. INTRODUCTION

The reconstruction of mandible is a great challenge in oral and maxillofacial surgery. The loss of mandibular bone leads to severe deformation and dysfunctions such as swallowing, chewing and speaking. Usually bone defects in mandible region are caused by tumors and bone infections [1]. The traditional reconstruction plates are difficult to form into the mandible anatomical shape. Plate failures always occur at the plate bending area and usually lead to unsatisfactory functional results to patients. Because of the complexity of human anatomy and individual difference in general, the construction of true anatomical model based on the individual patient has great significance. Combining with image processing and visualization techniques, the customized model as well as the finite element biomechanical simulation model has possibility to be established by medical images which have no trauma. The customized biomechanical modeling based on CT images provides accurate and simple way to establish the finite element model of organisms which try its best to be faithful to the customized anatomy structure and simulate the biomechanical behavior of living tissue with high precision.

In this study, the mandible reconstruction was conducted by using the segmental fibulas, which were similar to the defect of mandible in morphology and fixed by titanium plate and screws. This method involves Rigid Internal Fixation (RIF), which takes bony cortex of lateral law as basis of fixation. Michelet et al. (1973) proposed the notion of simple cortex of bone osteosynthesis. RIF achieves rigid fixing in 3D and has the advantages of fast healing, no external callus and direct cicatrization.

In order to fix the reconstruction mandible with the fibular, a customized titanium plate will be designed. The reconstructive mandible FE model will be established and prepare to be analysed in Abaqus. After the mandible reconstructed, the mandible model will be disposed by a Finite Element Analysis. Comparing with the normal mandible mechanical data, if the FEA data is regular, the RP model will be manufactured according to the 3D model. The whole process is represented by the Fig.1.

As in engineering, the initial work with FEA in medicine was the analysis of the macroscopic solid structures [2].The FEA was also used some years ago to predict dental implant loading [3], [4] and for micro structural analysis [5].

## II. METHODS

### A. Mandible Fixation and Simulation.

The reconstructive mandible 3D virtual model was established from the patient's CT data in Mimics [6].According to the mandible shape, the customized titanium plate was modeling in Pro/E condition. The titanium screw model also was built and assembled with the titanium plate. In order to accurately conduct FEA, the thread should be omitted to prevent irregular triangles and other errors appearing. It was the titanium screw model in Fig.2 and the assembly of titanium plate and titanium screws was in Fig.3.

The assembled titanium plate and titanium screws were imported Mimics with STL format, then fitted and simulated with the shape of mandible 3D model as shown in Fig. 4.

After the simulation, the 3D model would be input Magics software (Materialise, Inc., Leuven, Belgium). Depending on the location of titanium screws, the screw holes would be located and punched as shown in Fig. 5 and Fig.6. In Magics, the STL file was edited and simplified so that it was only one shell and contained a regular and reduced number of triangles.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

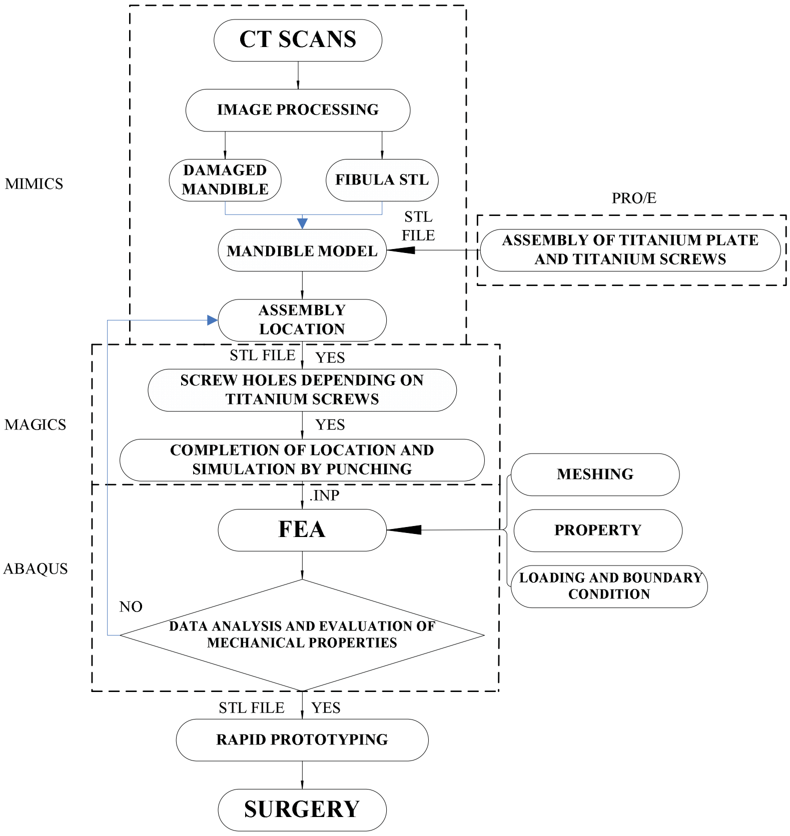

MIMICS
CT SCANS
IMAGE PROCESSING
DAMAGED MANDIBLE
FIBULA STL
PRO/E
STL FILE
MANDIBLE MODEL
ASSEMBLY OF TITANIUM PLATE AND TITANIUM SCREWS
ASSEMBLY LOCATION
STL FILE
YES
MAGICS
SCREW HOLES DEPENDING ON TITANIUM SCREWS
YES
COMPLETION OF LOCATION AND SIMULATION BY PUNCHING
.INP
MESHING
ABAQUS
FEA
PROPERTY
LOADING AND BOUNDARY CONDITION
NO
DATA ANALYSIS AND EVALUATION OF MECHANICAL PROPERTIES
STL FILE
YES
RAPID PROTOTYPING
SURGERY

Figure 1. Whole process of designing and force analysis

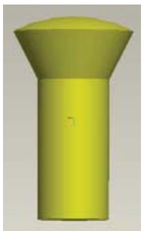


Figure 2. Titanium screw model

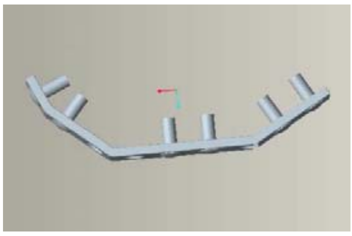


Figure 3. Assembly of titanium plate and titanium screws

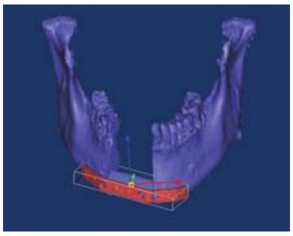


Figure 4. Fitting of titanium and mandible 3D model

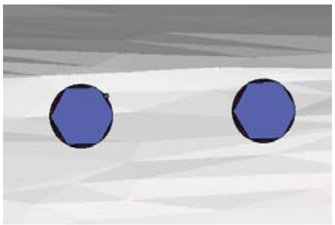


Figure 5. Screw holes

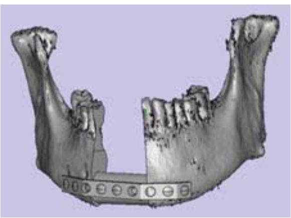


Figure 6. Reconstructive mandible and titanium plate

This research has its advantage of mandible fixation and simulation. The screw holes were punched firstly in traditional CAD model, then the titanium screws were located depending on the screw holes. In this research, the mandible shape was irregular and the titanium plate and titanium screws were modeling in Pro/E condition. If one screw hole was located and punched at first, the location of the next screw hole must be wrong because the titanium plate is limited. Contrary to the traditional way, the assembly of titanium plate and titanium screws was located at first, then the screw holes would be fixed and punched depending on the titanium screws. This method ensures the accuracy of screw hole’s location in this non-traditional CAD model.

### *B. FE Model Establishment*

In Magics, the repaired model was imported to the remeshing module which could divide the model into triangle meshes and finish the repair and optimization, and then the preprocessing documents of finite element method were exported. In this paper, a nonlinear finite element analysis software named Abaqus was adopted, so the exported document should be stored as .inp. Then the .inp model would be imported into Abaqus and converted from triangular meshes to tetrahedral meshes through the mesh module of Abaqus. The mesh model is shown in Fig.7. The tetrahedral meshes totally had 138866 tetrahedral elements and 30586 nodes as shown in Table I.

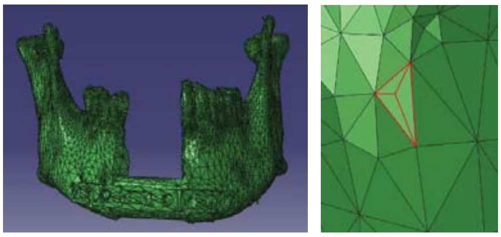


Figure. 7. Reconstruction mandible after remeshed and its grid cell

TABLE I. PARAMETER OF MANDIBLE FE MODEL

| Number of nodes | Number of elements | Element types |
|---|---|---|
| 30586 | 138866 | C3D4 |

### C. *Property*

In Abaqus, a new job in FE model would be created and exported to Mimics with .inp file. The property was given by module which the Mimics came with. The following is the formula for calculating the property parameters.

$$\begin{cases}\text{Density} = \text{-}13.4 + 1017 \times \text{Grayvalue} \\ \text{E - Modulus} = \text{-}388.8 + 5925 \times \text{Density}\end{cases} \qquad (1)$$

After the property was given completely, the model should be exported from Mimics to Abaqus with .inp file. Examining the property module, the model's color had turned light blue, which proved that the property was given successfully. The model with property is shown in Fig.8.

This is the essence: CT images of bone were calculated according to the pixel and the CT value would be obtained. The density of each part would be got from CT value and the property parameters of each skeletal part could be calculated successfully relying on the formula (1).The mandible consists of many composite biomaterials,so endowing property artificially would caused unexpected inaccuracy and errors. Comparing with the artificial way, it is a good choice to endow property with Mimics software.

### D. *Loading and Boundary Condition*

Before the FE model was done an analysis, the loading and boundary condition need to be set.

Boundary constraint is a very important issue related to the accuracy of FE model. The exactness or not of boundary constraint will directly affect the accuracy of the result. The perfect instance is that all boundary conditions are the same with actual instance, but it is almost impossible owing to the complex configuration. Therefore, the boundary conditions need to be predigested. As to the definition of mandibular boundary conditions, there was no uniform conclusion. Li ling et.al [7] achieved reasonable result with such boundary conditions: the mandibular lower edge as the underside boundary, acetabulum plane as the upside boundary and back of mandibular rise-offset as definition. This paper set the boundary conditions as shown in Fig. 9.

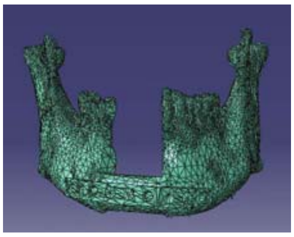


Fig.ure 8. Model with property

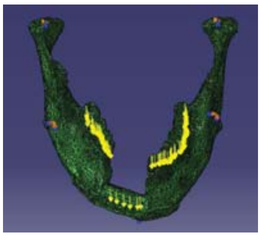


Figure 9. Loading and boundary conditions

The left tooth and the fibular surface were loaded in this analysis. The following are the boundary conditions.

$$\begin{cases}\text{acetabulum plane:} \\ ZASYMM\big(\ U1 = U2 = UR3 = 0\big) \\ \text{The reconstructive mandible middle line:} \\ XSYMM\big(\ U1 = UR2 = UR3 = 0\big) \\ \text{The mandible trailing edge:} \\ ENCASTRE\big(\ U1 = U2 = U3 = UR1 = UR2 = UR3 = 0\big)\end{cases}$$

### E. *Finite Element Analysis in Abaqus*

According to the actual mandibular occlusal force condition, the elastic mechanics theory was used to analyze the mandibular biomechanics. Spatial problem has 15 unknown numbers: six stress components $\sigma_x$ 、 $\sigma_y$ 、 $\sigma_z$ 、 $\tau_{yz} = \tau_{zy}$ 、 $\tau_{zx} = \tau_{xz}$ 、 $\tau_{xy} = \tau_{yx}$ ; six deformation components $\xi_x$ 、 $\xi_y$ 、 $\xi_z$ 、 $\gamma_{yz}$ 、 $\gamma_{zx}$ 、 $\gamma_{xy}$ ; three displacement components $u$ 、 $v$ 、 $w$ . These 15 unknown functions should meet the 15 basic equations.

The space problem of equilibrium differential equation:

$$\left.\begin{aligned}\frac{\partial\sigma_x}{\partial x} + \frac{\partial\tau_{yx}}{\partial y} + \frac{\partial\tau_{zx}}{\partial z} + f_x = 0 \\ \frac{\partial\sigma_y}{\partial y} + \frac{\partial\tau_{zy}}{\partial z} + \frac{\partial\tau_{xy}}{\partial x} + f_y = 0 \\ \frac{\partial\sigma_z}{\partial z} + \frac{\partial\tau_{xz}}{\partial x} + \frac{\partial\tau_{yz}}{\partial y} + f_z = 0\end{aligned}\right\} \qquad (2)$$

According to the elastic mechanics theory, the elastomer stress boundary condition:

$$\left.\begin{aligned}l(\sigma_x)_s + m(\tau_{yx})_s + n(\tau_{zx})_s = \overline{f_x} \\ m(\sigma_y)_s + n(\tau_{zy})_s + l(\tau_{xy})_s = \overline{f_y} \\ n(\sigma_z)_s + l(\tau_{xz})_s + m(\tau_{yz})_s = \overline{f_z}\end{aligned}\right\} \qquad (3)$$

It demonstrated the relationship between the stress components of the boundary value and the plane force components [8].

Considering the above points, the steps were set for Finite Element Analysis after importing mandibular model into Abaqus. Two steps were set separately to raise the simulation reality. In order to raise the load regularly, the first step load is set low, while the second normal. This research mainly analyzed the force of chewing, as the force of not chewing was too low to consider; thus, we adopt Static-General analysis.

Calculation formula of Von Mises ( $\sigma_e$ ) is：

$$\sigma_e=\left\{\frac{1}{2}\left[\left(\sigma_1-\sigma_2\right)^2+\left(\sigma_2-\sigma_3\right)^2+\left(\sigma_3-\sigma_1\right)^2\right]\right\}^{\frac{1}{2}} \qquad (4)$$

$\sigma_1$ 、 $\sigma_2$ 、 $\sigma_3$ is the first, second and third principal stress [8].

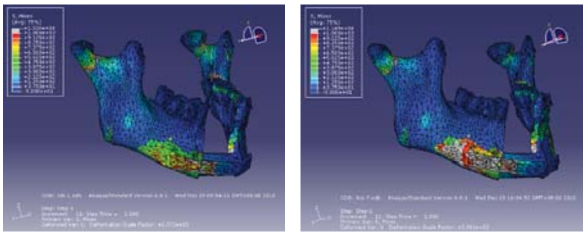
### F. *Results and Discussion*

When occluding, chewing muscles drive teeth to chew the food so as to make the mandible suffer load; and when not occluding, mandible hardly support load. Therefore, the analysis of mandibular force is mainly the force analysis of occluding. Bai shizhu, Li dichen et.al [9] applied force on the occluding plane to FEA the mandible, and got relatively accurate result.

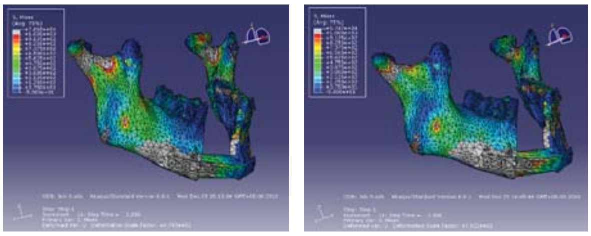
In this study, the force was applied separately in the direction of vertical, 45° with the occluding plane (teeth force range when occluding) to analyze the force. The load magnitudes are 100N and 500N.

Depending on the magnitude of the load and the analysis of different loading directions, we got reconstruction mandible stress simulation results as shown in the reconstruction of mandibular force after the simulation process. The maximum stress was at mandibular condylar neck.

According to the Von Mises stress diagram of reconstructive mandible, we can get the results
with the magnitude of load increasing, right side of the mandible and fibula junction, and the mandible condyle the stress concentration region increased. The maximum stress of the Von Mises stress diagram of reconstructive mandible increased as the load increased gradually. The Von Mises stress is shown in Fig.10.



（A）100N-0° （B）100N-45°



（C）500N-0° （D）500N-45°

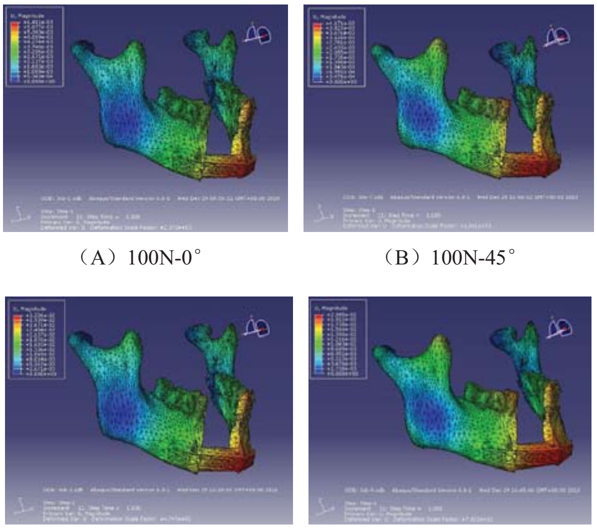
Figure 10. Result of reconstructive mandible Von Mises stress on different conditions

Reference to the classification criteria from Urken et al [10], this model belongs to B-type damage. Based on the results from Abaqus, the maximum stress of the Von Mises is shown in Table II.

TABLE II. MAXIMUM STRESS OF THE VON MISES

| Angles and loads | Maximum stress of the Von Mises (Mpa) |
|---|---|
| 100N-0° | 0.01534 |
| 500N-0° | 0.07668 |
| 100N-45° | 0.01149 |
| 500N-45° | 0.05747 |

According to the deformation results, with the magnitude of load increasing, the maximum deformation increased at the mandible midline. The mandible maximum deformation was increasing as the load was increasing gradually. The mandible deformation is shown in Fig.11.



（A）100N-0° （B）100N-45°

（C）500N-0° （D）500N-45°

Figure 11. Result of reconstructive mandible deformation on different conditions

The maximum deformation was at left side of the mandible and fibula junction and it is shown in Table III.

TABLE III. MAXIMUM DEFORMATION

| Angles and loads | Maximum deformation (mm) |
|---|---|
| 100N-0° | 0.06411 |
| 500N-0° | 0.32060 |
| 100N-45° | 0.04171 |
| 500N-45° | 0.20850 |

## III. Model building

Refer to the reference [11], the von mises stress and deformation were in the regular range and the RP model can be manufactured with 3DP rapid prototyping which is from Israel. The model with support materials is shown in Fig.12 and the model removing support materials is shown in Fig.13. The doctor could measure kinds of datum and practice operation to enhance the efficiency and success rate of surgery by 3D model which had been manufactured.

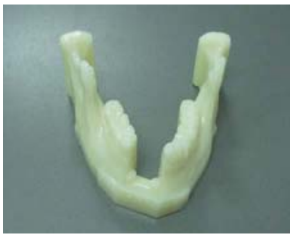


Figure 12. Model with support materials

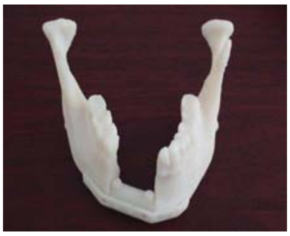


Figure 13. Model without support materials

## IV. Conclusions

Sekou Singare and Dichen Li [12] had established the mandible model successfully with CT data and curve fitting. The RP model had been manufactured at first, then the mold would be reproduced. After that, the titanium implant had been casted and implanted into the human body completely as shown in Fig.14.Comparing with the method which Sekou Singare had proposed, this set of methods described in this study have four advantages.

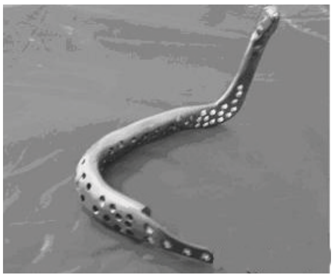


Figure 14. Titanium imported part

1) This set of methods were made up of kinds of professional software and largely reduced uncertain factors and errors which came from researcher's action. The researcher could build accurate outer contour and precise FEA, so the accuracy of model and the success rate of operation could be enhanced. This research is applicable to most of bone repair and large-scale promotion, so they have good business prospects.

2) Because the titanium implant was in the human body for a long time, the metal ions may leak and have a bad effect on human health [13]. Long-term use would also lead to increasing the internal stress and the rate of fatigue fracture. When the RIF was used, the titanium plate could be removed by the second operation after the mandible grew together with the fibula and the repaired mandibular could stand kinds of normal bite force.

3) After the FE model was established, the mechanical data would be analyzed to judge whether the model was qualified. It could decrease the risk of surgery, optimize the surgery program and alleviate the patient's suffering.

4) Autogenous bone graft has the capacity of becoming its marked osteogenic bone and resisting against infections.

This study represents a new approach for surgical planning and simulation, and both the designer and surgeon appreciate the meaning for comprehending the model. The mandible FE model has the ability to do the biomechanical analysis for the mandible model, and increase surgical precision over surgeries performed without the aid of mandible FE model. It should be recognized that there are a few shortcomings to improve such as: the intermediate links which were prone to lose some datum, higher operating skills which need higher knowledge to master, so it remains a big challenge and deserves the need for more attention and further research.

## V. References

[1] S. Singare, D. Li, B. Lu, Y. Liu, Z. Gong and Y. Liu, "Design and fabrication of custom mandible titanium tray based on rapid prototyping", Medical Engineering & Physics, Vol. 26, p. 671-676, 2004.

[2] Kavanagh EP, Frawley C, Kearns G, McGloughlin T and Jarvis J, "Use of finite element analysis in presurgical planning: treatment of mandibular fractures", Ir J Med Sci, Vol.177, p. 325-31, 2008.

[3] O'Mahony AM, Williams JL and Spencer P, "Anisotropic elasticity of cortical and cancellous bone in the posterior mandible increases peri-implant stress and strain under oblique loading", Clin Oral Implants Res, Vol.12, p. 648-57, 2001.

[4] Huang HL, Lin CL, Ko CC, Chang CH, Hsu JT and Huang JS, "Stress analysis of implant-supported partial prostheses in anisotropic mandibular bone: in-line versus offset placements of implants", J Oral Rehabil, Vol.33, p.501-8, 2006.

[5] Korioth TW, Romilly DP and Hannam AG, "Three dimensional finite element stress analysis of the dentate human mandible", Am J Phys Anthropol, Vol.88, p. 69-96, 1992.

[6] W.Z. Wu, X.J. Qin, Y. Zhang and W.S. Wang, "Mandibular Virtual Reconstruction Surgery Guided by Mimics", Applied Mechanics and Materials, Vol. 16-19, p. 842-846, 2009.

[7] L. Li, S. Xue and F. Q. Zhang, "The establishment of three-dimension finite element model for Upper and lower jaws and dentition", Dental Materials and Devices, Vol.12, p.117-121, 2003. (in Chinese)

[8] Z. L. Xu, Theory of elasticity(volume 1)[M], China: Higher Education Press, 2008.

[9] S. Z. Bai, D. C. Li, Y. M. Zhao and X. Li, "Study on the finite element analysis boundary conditions of exognathion", J Oral Maxillofac Surg, Vol.22, p. 720-723, 2006. (in Chinese)

[10] Urken ML, Buchbinder D, Weinberg H, et al., "Functional evaluation following microvascular oromandibular reconstruction of the oral cancer patient : A comparative study of reconstructed and nonreconstructed patients", Laryngoscope, Vol.101, p. 935-950, 1991.

[11] T. Ji, Y. Tie, D. M. Wang and C. P. Zhang, "Three -dimensional finite element analysis of the mandible reconstruction with fibula", West China Journal of Stomatology, Vol.27, p.143-146, 2009. (in Chinese)

[12] S. Singare, Liu Yaxiong, Li Dichen, et al., "Fabrication of customised maxillo-facial prosthesis using computer-aided design and rapid prototyping techniques", Rapid Prototyping Journal, Vol.12, p. 206–213, 2006.

[13] Ajit K M, James A D, Paul K and RobertA P, "In:Eylon D, Boyer R R, Koss D A eds", Beta Titanium Alloys in 1990's, TMS, Warrendale, PA, p. 61, 1993.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# A RFID mutual authentication protocol based on AES algorithm

Tuan Anh Pham, Mohammad S. Hasan and Hongnian Yu
Faculty of Computing, Engineering and Technology, Staffordshire University, Stafford, UK
tapham@live.com, {m.s.hasan, h.yu}@staffs.ac.uk

*Abstract*— **The emergence of RFID applications has huge influence to become pervasive in modern life. However the vulnerability of the transmission through the air and the unique identification number of RFID tag are the drawbacks that impact the popularity of RFID technology. In this paper, a mutual authentication protocol is proposed based on the challenge – response model. The Advanced Encryption Standard (AES) is used as a cryptographic primitive to secure the data. The experimental works are carried out to validate the protocol in term of security and privacy. The timing analysis is also presented and applied to a case study of conveyor belt system.**

***Keyword: RFID, AES algorithm, mutual authentication***

## I. Introduction

Recently Radio Frequency Identification (RFID) systems have been becoming popular and aiming to be ubiquitously applied in many areas including library, banking, logistics, transportation, manufacturing, supply chain system, military etc. Some large corporations who have deployed this technology are Wal-Mart, Procter and Gamble, and the United State Department of Defense [1] etc.

A RFID tag can be read as long as the item is within the range of the reader without requiring the line-of-sight operation as bar code technology. However, one of the biggest difficulties to the adoption of RFID technology is the lack of security and privacy. There is little security on the RFID tags or during the communication with reader which causes the RFID system vulnerable to many types of attacks e.g. information leakage, replay, denial of service [2], [3].

Many authentication protocols have been proposed to enhance the robustness of a RFID system. Some of them were developed based on cipher algorithms such as Advanced Encryption Standard (AES) or Elliptic Curve Cryptography (ECC) [4], [5], [6], [7]. Some utilised the hash-based algorithm, pseudo random number generator, Cyclic Redundancy Check (CRC) function and/or some ExOR and rotation operations [8], [9], [10], [11], [12]. Normally, cipher-based approaches are not preferred for passive tag because of its high computational cost and the large hardware area. However, with the advances in technology, it is feasible to embed the cipher engine on the passive tags but still guarantee the low tag's cost [5].

The aim of this paper is to develop a protocol to provide a strong, high security and trustful authentication scheme which can protect against most of well-known RFID system attacks. Thus, a novel mutual authentication protocol based on AES primitives and challenge – response method is proposed. The symmetric block cipher AES-128 is utilised in this proposal because it has been standardised and proved to be secure [13].

The rest of the paper is organised in six parts from section II to section VII. Section II briefly introduces the recent works. Section III proposes the novel protocol. In section IV, the analyses of the protocol in term of security and privacy are presented. Section V provides some experimental results. And section VI introduces a case study on conveyor belt system. The last section, section VII, is the conclusion of the paper.

## II. Related works

### A. *RFID security proposals*

There are the number of researches to address RFID security and privacy issues. This paper roughly categorises them into two categories: **cipher-based protocols** which are developed on AES or ECC algorithms and **hash-based protocols and others** which are based on hash functions and/or some simple operators such as ExOR, rotation etc.

#### 1) *Cipher-based proposals*

A strong authentication for RFID systems is introduced in [5] by implementing one-way encryption AES algorithm on RFID passive tag. A modified communication method between the reader and tag is also proposed in order to satisfy the strict timing requirement. However, this research reveals a possibility for the adversary to achieve the shared key of the AES encryption block [7]. In addition, the identical challenge always results in the identical cipher response which causes it to be susceptible to replay attack and tag clone. For example, assuming the challenge is the 16-bit output of Pseudo Random Number Generator [14], it is feasible for attackers to collect a database of 65536 entries to impersonate the legitimate devices to obtain the authentication.

Extending from the research of [5], [7] offers an advanced mutual authentication using the AES algorithm as a cryptographic primitive. The main issue of this protocol is easy to lose the synchronisation between the reader and the tag if the response from tag is blocked. Another security issue of this research is the man in the middle attack [6].

[4] develops a mutual authentication protocol based on Elliptic Curve Cryptography (ECC) which is an asymmetric

cryptographic algorithm. However, it is susceptible to DOS attack. The attackers can modify the counter value in the message sent from the server to the tag by a much larger value in order to deceive the tag into updating it. In next authentication, because the counter value in the tag is greater than the one the server sends, the tag terminates the authentication process right away.

### 2) *Hash-based proposals and others*

[11] proposed a mutual authentication protocol for passive tag which is based on cryptographic hash functions. A method to prevent the desynchronisation between a reader and a tag was proposed as well. This proposal however might disclose the issues of tracking, tag impersonation and the DOS attack [15].

Another hash-based two-way authentication scheme which employs Cyclic Redundancy Check (CRC) and Pseudo Random Number Generator (PRNG) function was presented in [9]. Although it was claimed to be robust, the attackers are still able to track the tag due to its identical response. Furthermore, the replay attack and tag impersonation can be carried out to compromise the valid reader. For example, the attackers possibly play on the response from the tag embedded in the cheap product and then replay that to the standard item.

[12] proposed a scheme for RFID system conforming to EPC Class 1 Generation 2 specifications. However, [16] has pointed out many security failures in this protocol. Moreover, the DOS attack can be employed to desynchronise the database server and the tag [10]. [16] has also indicated the possibility of auto-desynchronisation at backend database up to 0.93 if the population of tags is greater than $2^{18}$ tags.

To overcome the security issues of [12], [10] enhanced the scheme to defend against DOS attack, replay attack, forward secrecy and privacy concern. However, [17] and [8] claimed that this proposed protocol is not invulnerable due to the insufficient size of the secure keys. Consequently, it is susceptible to Tag/Reader impersonation attack and desynchronisation as [17] also claimed.

## III. THE PROPOSED PROTOCOL

### A. *Assumptions*

- A tag has the AES-128 encrypting block on-board proposed in [26].
- A reader and database server can perform AES-128 encryption and decryption.
- The channel between the reader and tag, the reader and backend database are vulnerable.

### B. *Initialisation phase*

At the beginning, the backend database, the reader and the tag have the same shared key *K*. The seed *s* of each tag is stored on the server and on the rewriteable memory of tag. The *ID* is placed in tag's read-only memory and server as well. In addition, the server and reader keep the reader identification number $ID_R$ for reader authentication purpose.

### C. *Authentication phase*

The authentication phase is illustrated in Figure 1.

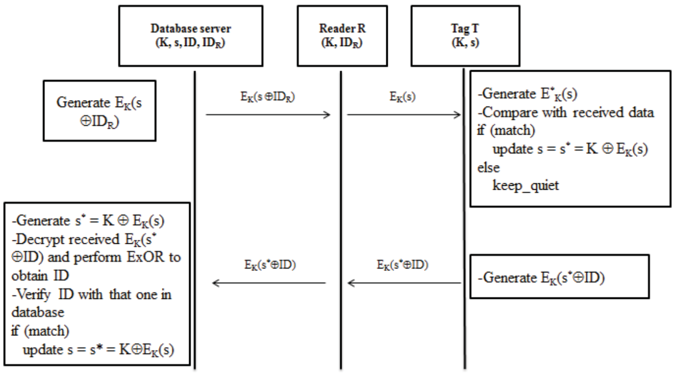


Figure 1. The proposed protocol

- Step 1: Query

Initially, the server generates $E_K(s⊕ID_R)$ and sends to the reader. The reader will decrypt this value to obtain the seed *s* by performing $s=ID_R⊕(ID_R⊕s)$. Afterward, reader encrypts the seed *s* and forward to the queried tag.

- Step2: Reader-to-Tag authentication

On receiving $E_K(s)$, the tag employs the on-board AES encrypting block to produce its own $E^*_K(s)$ and then check whether $E^*_K(s)=E_K(s)$ holds. If not, the tag keeps quiet. On the contrary, tag updates the seed *s* by $s^*=K⊕E_K(s)$ and again performs AES encrypting block to compute $E_K(s^*⊕ID)$ in order to convey to the reader.

- Step 3: Tag-to-Reader authentication

Upon obtaining $E_K(s^*⊕ID)$, the reader will pass it to the backend database to decrypt it. The decrypted value $s^*⊕ID$ is used to extract the tag's *ID* by carrying out the simple operation $ID = K⊕E_K(s⊕(s^*⊕ ID)$. Then this *ID* is verified by comparing with the *ID* existing in database to check whether this tag is legally acceptable or not. If the mismatch occurs, the server discards any data it has received and declines the authentication of the tag. Oppositely, the server updates the seed value of reader by $s^*=K⊕E_K(s)$ to guarantee the synchronisation of the system.

## IV. SECURITY AND PRIVACY ANALYSES

### A. *Information leakage*

Due to the insufficient protections and unreliable security levels, the data transmitted through the air are easily compromised. In this protocol, there are two messages interchanged between the reader and the tag. However, these sensitive data are encrypted by AES-128 cryptographic block. The attackers cannot get the plaintext or the raw data. So the information leakage is evitable.

### B. *Tag tracking and tracing*

For even EPC or IEEE standard, RFID tag is designed to have a unique number which can be tracked in range of any readers. However, this protocol provides the mechanism that whenever the counterfeit reader queries the tag, the tag does not send any response back. Consequently, the tracking and tracing privacy is secure in this protocol.

### *C. Tag clone*

In theory, there is no tag which has similar unique identifier number to another one. However, adversaries can replicate the forged tag without much effort and expense [18]. To conduct the tag clone attack, the attackers have to obtain the key $K$, seed $s$ and the $ID$ of the tag. However, these shared data are kept safely and privately in the backend database and inside the tag; the attackers have no information to perform the same encryption block to acquire the authentication.

### *D. Man-in-the-middleattack*

The adversative readers can impersonate the valid one in order to intercept, change and obtain the messages going between the parties which they need. In the proposed protocol, before querying the tag, the server sends the message $E_K(s \oplus ID_R)$ to the reader. Because only valid reader has the key $K$ to decrypt this cipher text, there is no possibility for attackers' readers to achieve the seed $s$ in order to communicate with the tag. Thus the man-in-the-middle attack can be avoided.

### *E. DOS(Denial of Service) attack*

It affects any wireless communication e.g. WiFi, RFID etc. by obstructing or intercepting the wireless signal that causes loss of synchronisation among the devices. Let us assume the scenario in which the response $E_K(s \oplus ID)$ is intercepted by the adversary. Because the reader has not got the response from the tag, it does not update its seed. In the next authentication, reader sends $E_K(s)$ to challenge the tag. But the seed in tag now is $s^* = K \oplus E_K(s)$, therefore, $E_K(s) = E_K(s^*)$ is not satisfied and the tag keeps quiet. So the solution is that reader will attempt with $E_K(K \oplus E_K(s))$. At this attempt, the synchronisation is re-established. Hence, the proposed protocol is insusceptible to DOS attack.

### *F. Replay attack*

In replay attack, the adversaries stand in the middle of communication channel to duplicate the valid transmission which will be fraudulently repeated later. However, the seed $s$ is automatically updated after each successful authentication session. This leads to the fact that the cipher text $E_K(s)$ changes in every authentication cycle. Thus, the attackers cannot utilise the former data in order to deceive the authorised reader or tag to overcome the authentication process.

Overall, the proposed protocol is able to protect against many types of the known attacks. The comparison with other researches is shown in Table I to have a general view of the robustness of this protocol.

## V. SIMULATION AND VALIDATION

The simulation is executed on a computer with Intel T3200 2GHz processor, 2GB of Ram memory on Windows 7 Ultimate. The C programming language has been used to develop the simulations. The AES core is the open source program designed by Niyaz PK [19].

TABLE I. COMPARISONS IN TERM OF SECURITY AND PRIVACY

| | *[5]* | *[7]* | *[11]* | *[9]* | *[12]* | *[10]* | **Proposed protocol** |
|---|---|---|---|---|---|---|---|
| Information leakage | Δ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Tag tracking and tracing | ✓ | ✓ | Δ | × | × | × | ✓ |
| Tag clone | × | ✓ | Δ | × | × | × | ✓ |
| Denial Of Service (DOS)/Desynchronisation | ✓ | × | Δ | ✓ | × | × | ✓ |
| Replay | × | × | ✓ | × | ✓ | ✓ | ✓ |
| Mutual authentication | × | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

✓: fully satisfied; ×: not satisfied; Δ: partially satisfied as assumption

### *A. Performance analysis*

In this proposed protocol, the messages transmitted on the channel are always encrypted. However, the adversaries can play brute-force attack (BFA) to attempt to get the answer from the tag. But since the size of the message is 128 bits, it is not feasible to perform BFA entirely with all the possible cases (more than $3x10^{38}$). Therefore, the simulation will randomly pick up an amount of random 128-bit input vectors (it is called sub-BFA) in order to verify the possibility of unintentional matches. The results are shown in Figure 2.

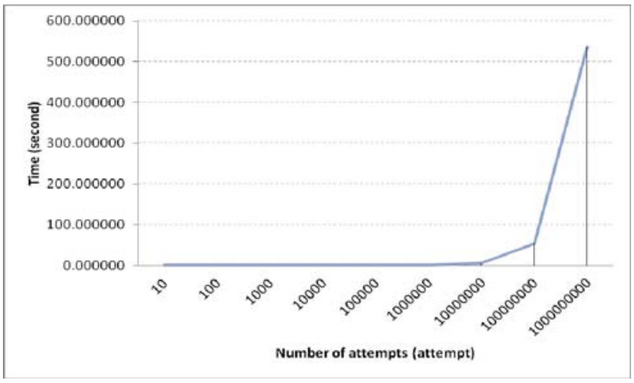


Figure 2 Time taken for sub brute force attack

Another simulation is conducted to measure the average time to perform one typical AES-128 encrypting operation. It is shown that it takes about **37µs** to complete.

From the simulation results, the protocol is robust to brute-force attack. And in case the server desynchronises, it takes just a few seconds to perform the AES encrypting computations to obtain the synchronisation.

### *B. Timing analysis*

#### *1) Timing components*

In this section, the minimum and maximum time requirements of the communication among the server, the reader and the tag are analysed. The time complexity is calculated based on the parameters in [14]. The timing model is shown in Figure 3.

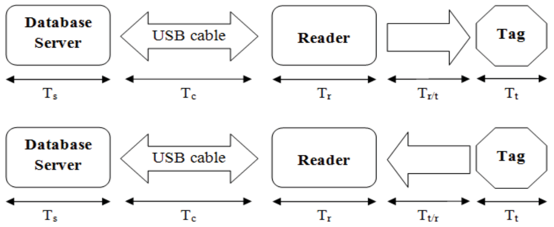


Figure 3. Timing model

In which:

$T_s$, $T_r$, $T_t$ : time for the database server, reader and tag to perform the AES algorithm respectively.

$T_c$ : time to transfer 128-bit message from the database to the reader through high-speed USB cable

$T_{r/t}$, $T_{t/r}$ : time to transmit 128-bit message from the reader to the tag and from the reader to the tag respectively.

The values of $T_s$, $T_r$, $T_t$ and $T_c$ can vary depending on the configuration of the server, reader, tag and the length of the cable, respectively.

#### *2) Assumptions*

- $T_s$, $T_r$, $T_t$ and $T_c$ are assumed to be constants.
- The computational times for AES encryption and decryption are the same.

#### *3) The minimum and maximum values*

Figure 4 shows the timing details of communication between the reader and the tag.

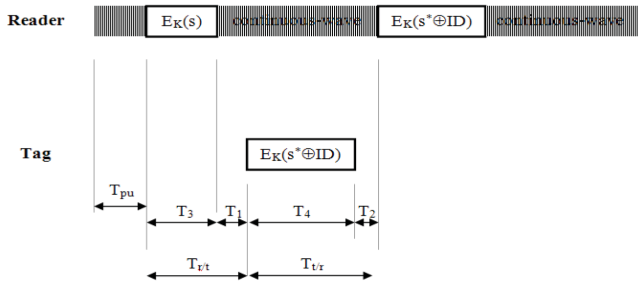


Figure 4. Time of reader & tag communication

In which:

$T_{pu}$ : time for tag to power-up

$T_1$, $T_2$ : time from reader transmission to tag response and from tag response to reader transmission respectively

$T_3$, $T_4$ : time to transmit 128-bit message from the reader to tag and from the tag to reader

Let T be the total time of an authentication cycle.

$$T = T_{\text{database}\to\text{tag}} + T_{\text{tag}\to\text{database}}$$

$$= (T_s + T_c + T_r + T_{r/t} + T_t) + (T_t + T_{t/r} + T_r + T_c + T_s) \qquad \textbf{(1)}$$

In tag-to-database direction, the value $T_r$ is 0 because the reader just transfers the tag's response to the database without performing any AES operation.

Therefore, (1) can be written as (2).

$$T = 2 \text{ x } (T_s + T_c + T_t) + T_r + (T_{pu} + T_1 + T_2 + T_3 + T_4) \qquad \textbf{(2)}$$

#### *4) $T_s$ & $T_r$*

The database requires one AES operation in either database-to-tag or tag-to-database authentication. According to section V.A, the time requirement to perform one AES encrypting function is about 37µs, therefore $T_s$ = 37µs.

The reader requires two AES operations on the way from the database to tag. Based on the assumptions presented in section V.B.2), $T_r$ can be presented by (3).

$$T_r = 2 \text{ x } 37 = 74\ \mu s \qquad \textbf{(3)}$$

#### *5) $T_t$*

The AES encrypting operation at the tag takes 356 clock cycles at the frequency of 100 KHz [20]. So $T_t$ is calculated as below.

$$T_t = 356 \text{ x } \frac{1}{100\times10^3} = 3560\mu s \qquad \textbf{(4)}$$

#### *6) $T_c$*

$T_c$ is computed by adding the time of transferring 128 bits from the server to the reader and the delay of the cable. Because the USB 2.0 can transfer at 480 Mbps and the cable delay is 26ηs [21]. So the value of $T_c$ is:

$$T_c = 128 \text{ x } \frac{1}{480\times10^6} + 26\eta s = 0.29267\mu s \qquad \textbf{(5)}$$

#### *7) $T_{pu}$*

[14] does not define the minimum of $T_{pu}$, it only indicates the maximal value of 1500µs for powering up the tag. Therefore, let us assume the min($T_{pu}$) = max($T_{pu}$) = 1500 µs.

#### *8) $T_1$*

[14] has given the equation to compute $T_1$ as (6).

$$\max(RTcal, 10T_{pri})x(1 - FT)\text{-}2\mu s \le T_1 \le \max(RTcal, 10T_{pri}) \qquad \textbf{(6)}$$

To calculate RTcal and $T_{pri}$, some values are given from [14] such as 6.25µs≤Tari≤25µs, 2.5Tari≤RTcal≤3.0Tari, 1.1RTcal≤TRcal≤3RTcal, BLF = $\frac{DR}{TRcal}$ , $T_{pri} = \frac{1}{BLF}$ .

Therefore, the minimum and maximum values for RTcal and TRcal can be expressed in (7).

$$\min(RTcal) = 15.625\mu s;\ \max(RTcal) = 75\mu s$$

$$\min(TRcal) = 17.1875\mu s;\ \max(TRcal) = 225\mu s \qquad \textbf{(7)}$$

Based on the minimum and maximum values of TRcal, the values of BLF and $T_{pri}$ can be found according to [14].

$$\max(BLF) = 465KHz;\ \min(BLF) = 95KHz$$

$$\min(T_{pri}) = \frac{1}{465\times10^3} = 2.15\mu s;\ FT = 19\% \qquad \textbf{(8)}$$

$$\max(T_{pri}) = \frac{1}{95\times10^3} = 10.53\mu s;\ FT = 5\%$$

Hence the minimum and maximum values of $T_1$ can be derived from (6) as (9) and (10), respectively.

$$\min(T_1) = \min(\max(RTcal,\ 10T_{pri}) \text{ x } (1 - FT) - 2\mu s) \quad \textbf{(9)}$$

$$= \max(\min(RTcal),\ 10 \text{ x } \min(T_{pri})) \text{ x } (1 - FT) - 2\mu s = 15.415\mu s$$

$$\max(T_1) = \max(\max(RTcal,\ 10T_{pri})) \quad \textbf{(10)}$$

$$= \max(\max(RTcal),\ 10 \text{ x } \max(T_{pri})) = 105.3\mu s$$

*9) $T_2$*

The equation used to calculate $T_2$ is referred from [14].

$$3T_{pri} \le T_2 \le 20T_{pri}$$

Therefore, the minimum and maximum values of $T_2$ can be computed as (11).

$$\min(T_2) = 3 \text{ x } \min(T_{pri}) = 6.45\mu s$$

$$\max(T_2) = 20 \text{ x } \max(T_{pri}) = 210.06\mu s \quad \textbf{(11)}$$

*10) $T_3$ & $T_4$*

In the reader-to-tag transmission, the reader starts signalling with either a preamble or a frame-sync as specified in [14]. This investigation assumes that the reader uses the frame sync which includes 3 components: delimiter=12.5µs; data-0=1Tari; RTcal [14]. The 128-bit data are sent afterward. At the end of the signalling either 2-bit 00 or 11 indicates the end of a communication. On the other hand, the tag starts the tag-to-reader signalling with a 6-bit preamble, then 128-bit data and 2-bit end of signalling.

The minimum and maximum values of $T_3$ are shown in (12) and (13) at the data rate of 465 KHz and 95 KHz respectively.

$$\min(T_3) = t_{delimiter} + \min(1Tari) + \min(RTcal) + \min t_{130\text{ bits}}$$

$$= 12.5 + 6.25 + 15.625 + \frac{130}{465 \times 10^3} \times 10^6 = 313.94\mu s \quad \textbf{(12)}$$

$$\max(T_3) = t_{delimiter} + \max(1Tari) + \max(RTcal) + \max t_{130\text{ bits}}$$

$$= 12.5 + 25 + 75 + \frac{130}{95 \times 10^3} \times 10^6 = 1480.92\mu s \quad \textbf{(13)}$$

Likewise, the minimum and maximum values of $T_4$ are computed as (14) and (15).

$$\min(T_4) = \min t_{136\text{ bits}} = \frac{6+128+2}{465\times10^3}\times10^6 = 292.47\mu s \quad \textbf{(14)}$$

$$\max(T_4) = \max t_{136\text{ bits}} = \frac{6+128+2}{95\times10^3}\times10^6 = 1431.58\mu s \quad \textbf{(15)}$$

Finally, deriving from (2) and the calculations above, the minimum and maximum values of T now can be resulted as (16) and (17).

$$\min(T) = 2 \text{ x } (T_s + T_c + T_t) + T_r + \min(T_{pu}) + \min(T_1) + \min(T_2) + \min(T_3) + \min(T_4) \approx 0.0094s \quad \textbf{(16)}$$

$$\max(T) = 2 \text{ x } (T_s + T_c + T_t) + T_r + \max(T_{pu}) + \max(T_1) + \max(T_2) + \max(T_3) + \max(T_4) \approx 0.012s \quad \textbf{(17)}$$

## VI. CASE STUDY: CONVEYOR BELT SYSTEM

The analysis and performance explained in section V.B are applied to estimate the maximum number of attempts for resynchronisation to a typical belt system as in Figure 5. The items with RFID tags are moving evenly from position A to position B on the conveyor belt. The range of the reader is L and θ is the angle covered by the antenna of the reader.

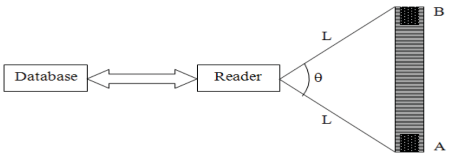


Figure 5. Model of RFID-based conveyor belt

The speed of the conveyor belt, $V_b$, is 2.5m/s which is the maximum speed in [22]. Hence, the time an item needs to move from A to B, denoted by t, is calculated as (18).

$$t = \frac{AB}{V_b} = \frac{2 \times L \times \sin\left(\frac{\theta}{2}\right)}{V_b} \quad \textbf{(18)}$$

[1] claimed that the read range varies from nominal range of 10cm to extended range of 50cm, so min(L) = 10cm and max(L) = 50cm. In addition, theoretically $0 \le \theta < 180$, it is inferred that $0 \le \sin\left(\frac{\theta}{2}\right) < 1$.

The maximum number of resynchronisation attempts, N is computed as (19) and (20) which are the best case scenario and worst case scenario respectively. These computations are based on (16), (17) and (18).

$$N = \text{floor}\left(\frac{t}{\min(T)}\right) = \text{floor}\left(\frac{L \times \sin\left(\frac{\theta}{2}\right)}{0.0094}\right) \quad \textbf{(19)}$$

$$N = \text{floor}\left(\frac{t}{\max(T)}\right) = \text{floor}\left(\frac{L \times \sin\left(\frac{\theta}{2}\right)}{0.012}\right) \quad \textbf{(20)}$$

Figure 6 and Figure 7 shows the value of N in (19) and (20) respectively with different values of L and θ.

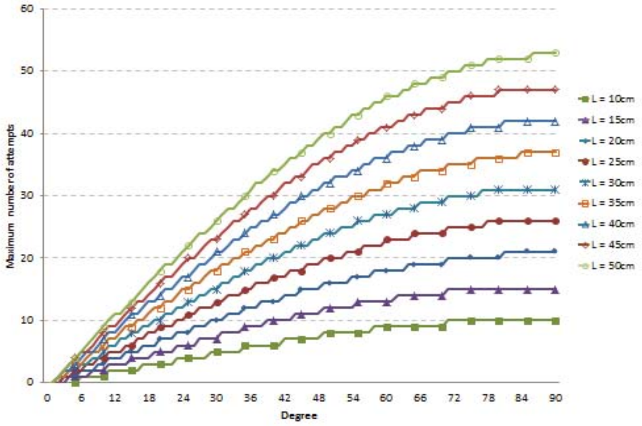


Figure 6. Maximum number of attempts to resynchronise a tag for various L and θ in case of min(T)

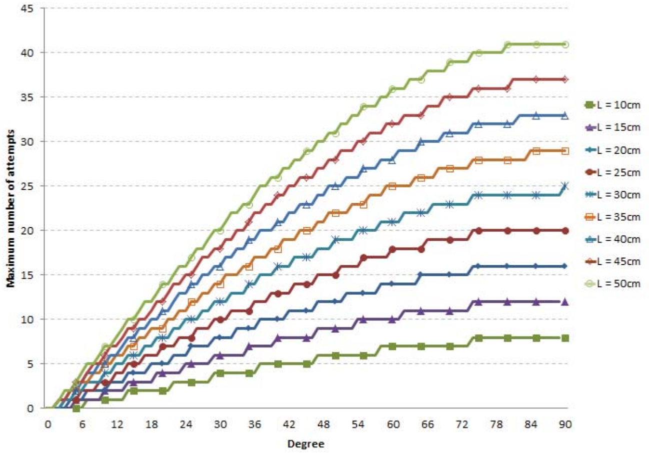


Figure 7. Maximum number of attempts to resynchronise a tag for various L and θ in case of max(T)

If θ = 68° is chosen as a reliable read angle as mentioned in [23], the values of N for min(L) = 10cm and max(L) = 50cm, are 6 and 29, respectively for the best case scenario or min(T). On the other hand, for the worst case scenario or max(T), the values are 4 and 23, respectively.

## VII. Conclusion

The proposed protocol in this paper is a mutual authentication protocol which utilises AES-128 as a primitive to encrypt the messages transmitted on the channel. With that cipher block, the protocol can protect against many types of attacks such as information leakage, tag tracking etc. In addition, the secure keys stored in tag and sever are always updated in each authentication session, it is impossible for attackers to play the replay attack or trace back the previous data.

## References

[1] A. Juels, "RFID security and privacy: a research survey," Selected Areas in Communications, IEEE Journal on, vol. 24, pp. 381-394, 2006.

[2] D. Dang Nguyen, L. Hyunrok, D. M. Konidala, and K. Kwangjo, "Open issues in RFID security," in Internet Technology and Secured Transactions, 2009. ICITST 2009. International Conference for, 2009, pp. 1-5.

[3] R. K. Pateriya and S. Sharma, "The Evolution of RFID Security and Privacy: A Research Survey," in Communication Systems and Network Technologies (CSNT), 2011 International Conference on, 2011, pp. 115-119.

[4] J.-S. Chou, Y. Chen, C.-L. Wu, and C.-F. Lin, "An efficient RFID mutual authentication scheme based on ECC," IACR Cryptology ePrint Archive, vol. 2011, p. 418, 2011.

[5] M. Feldhofer, S. Dominikus, and J. Wolkerstorfer, "Strong Authentication for RFID Systems Using the AES Algorithm," presented at the Workshop on Cryptographic Hardware and Embedded Systems CHES 2004, Boston Marriott Cambridge, Cambridge (Boston), USA, 2004.

[6] M. F. Mubarak, J. l. A. Manan, and S. Yahya, "Mutual Attestation Using TPM for Trusted RFID Protocol," in Network Applications Protocols and Services (NETAPPS), 2010 Second International Conference on, 2010, pp. 153-158.

[7] B. Toiruul and K. Lee, "An Advanced Mutual-Authentication Algorithm Using AES for RFID Systems," IJCSNS International Journal of Computer Science and Network Security, vol. 6 No.9B, 2006.

[8] Y. Eun Jun, "Improvement of the securing RFID systems conforming to EPC Class 1 Generation 2 standard," Expert Systems with Applications, vol. 39, pp. 1589-1594, 2012.

[9] H. Li, P. Yin, X. Wang, and L. Pang, "A Novel Hash-based RFID Mutual Authentication Protocol," in Computational Intelligence and Security (CIS), 2011 Seventh International Conference on, 2011, pp. 774-778.

[10] T. C. Yeh, Y. J. Wang, T. C. Kuo, and S. S. Wang, "Securing RFID systems conforming to EPC Class 1 Generation 2 standard," Expert Systems with Applications, vol. 37, pp. 7678-7683, 2010.

[11] L. Yanfei, "An Efficient RFID Authentication Protocol for Low-Cost Tags," in Embedded and Ubiquitous Computing, 2008. EUC '08. IEEE/IFIP International Conference on, 2008, pp. 180-185.

[12] H. Y. Chien and C. H. Chen, "Mutual authentication protocol for RFID conforming to EPC Class 1 Generation 2 standards," Computer Standards &amp; Interfaces, vol. 29, pp. 254-259, 2007.

[13] National Institute of Standards and Technology (NIST). (2001, 05 Dec 2011). Announcing the ADVANCED ENCRYPTION STANDARD (AES). FIPS 197. Available: http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf

[14] EPCglobal, 2008. EPC Radio-Frequency Identity Protocols Class-1 Generation-2 UHF RFID Protocol for Communication at 860 MHz - 960 MHz Version 1.2.0. Available: http://www.epcglobalus.org

[15] I. Erguler and E. Anarim, "Attacks on an Efficient RFID Authentication Protocol," in Computer and Information Technology (CIT), 2010 IEEE 10th International Conference on, 2010, pp. 1065-1069.

[16] P. P. Lopez, J. C. H. Castro, J. M. E. Tapiador, and A. Ribagorda, "Cryptanalysis of a novel authentication protocol conforming to EPC-C1G2 standard," Comput. Stand. Interfaces, vol. 31, pp. 372-380, 2009.

[17] M. Habibi, M. Gardeshi, and M. R. Alaghband, "Practical Attacks on a RFID Authentication Protocol Conforming to EPC C-1 G-2 Standard," International Journal of UbiComp (IJU), vol. Vol.2, No.1, 2011.

[18] [18] A. Mitrokotsa, M. Rieback, and A. Tanenbaum, "Classifying RFID attacks and defenses," Information Systems Frontiers, vol. 12, pp. 491-505, 2010.

[19] N. PK, P. Kumar, and A. A. Philip. AES Encrypt – Source code in C/C++ for AES Encryption. Available: http://www.hoozi.com/downloads/

[20] T. Good and M. Benaissa, "692-nW Advanced Encryption Standard (AES) on a 0.13-um CMOS," Very Large Scale Integration (VLSI) Systems, IEEE Transactions on, vol. 18, pp. 1753-1757, 2010.

[21] Compaq, Hewlett-Packard, Intel, Lucent, Microsoft, NEC, and Philips. (2000, USB 2.0 Specification. Available: http://www.usb.org/developers/docs/usb_20_1011111.zip

[22] L. Simon, P. Saengudomlert, and U. Ketprom, "Speed Adjustment Algorithm for an RFID Reader and Conveyor Belt System Performing Dynamic Framed Slotted Aloha," in RFID, 2008 IEEE International Conference on, 2008, pp. 199-206.

[23] H. D. Chon, S. Jun, H. Jung, and S. W. An, "Using RFID for Accurate Positioning," Journal of Global Positioning Systems, vol. 3, pp. 32-39, 2004.

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# Automatic Detection of Circular Defects During Ultrasonic Inspection

Thouraya Merazi-Meksen, Malika Boudraa and B.Boudraa
University of Science & Technology H. Boumediene
Faculty of Computer Science and Electronics
Algiers, Algeria
tmeksen@usthb.dz

*Abstract*—**In the non-destructive testing of materials, the use of imagery allows for more than only a qualitative representation of the results. Imagery can aid in automatically carrying out operations of detection, location and sizing of defects present in a structure. The rapidity of the calculators and their graphic performance allow us to use classical tools of image processing today, even though they require a large number of calculations. This work consists of applying the Hough transform to detect inclusions in a material by analyzing an ultrasonic c-scan image. In the first step, this image is binarized using the –6dB method, which reduces the defects to their real sizes. The Hough transform is then applied to the edges to detect the circular forms that characterize the inclusions.**

***Keywords-Non Destructive Testing; Ultrasonics; Imagery; Pattern Recognition.***

## I. Introduction

In non-destructive testing (NDT) of materials, the automatic characterisation of defects allows for faster decision making [1]. With the appearance of computing systems that are able to process recorded signals rapidly and create images, the possibilities of ultrasonic imagery in NDT improve steadily [2,3,4,5]. In addition to aiding in the control and decision-making processes, automation allows for control under dangerous conditions of high-level radiation or in some industrial processes with low or null visibility due to the environment.

The Hough transform is a classical tool for pattern recognition [6]. It allows for the detection of parameterized forms, such as lines, circles, parabolas, etc. Because it requires a large number of calculations, it had been abandoned for ultrasonic imagery applications since its initial development in 1962 by Paul Hough. However, its robustness and its capacity to detect even incomplete edges keep it on the list of the most interesting tools for pattern recognition.

In precedent works, we have used Hough transform to detect cracks defects by analyzing another type of ultrasonic images, called TOFD (Time Of Flight Diffraction). In this case, sets of parabolas appear on the image and their recognition allows the crack detection [7].

This paper describes how the Hough transform can be used on an ultrasonic image produced by non-destructive testing to detect inclusions characterised by circular forms.

The c-scan image used here displays ultrasonic waves reflected by the defect at each position of a transducer moving on the structure. In Section 2, the synthesis of this type of image is described. The application of the –6dB method to the objects detected to reduce them to their real sizes is explained in Section 3 [8]. This operation also allows us to transform the image into a binarized one. Section 4 describes how the Hough transform can be applied to detect and locate circular forms corresponding to inclusions.

## II. IMAGE FORMATION

An automatic scanning system was used to move an ultrasonic focused probe step by step on a rectangular surface of the structure in a controlled manner.

At each position, an ultrasonic signal was emitted, and the reflected signal was detected and stored [9,10].

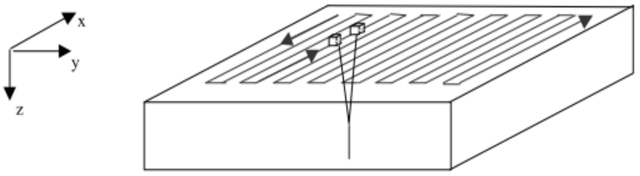


Figure 1. Displacements of the probe.

Figure 2 shows an example of the reached signal. Where appear the defect signal between surface echo and the backwall one.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

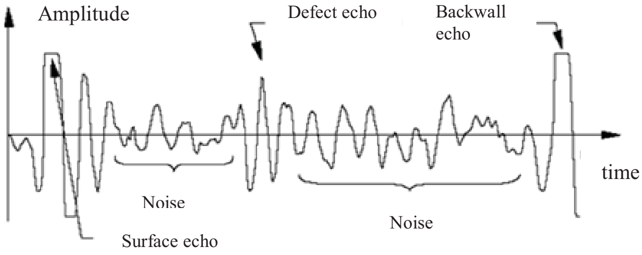


Figure 2. Example of a reflected signal.

The maximum amplitude of the echo corresponds to the value of the pixel at the corresponding coordinates of the probe. Figure 3 shows the results obtained by scanning a square area on the structure.

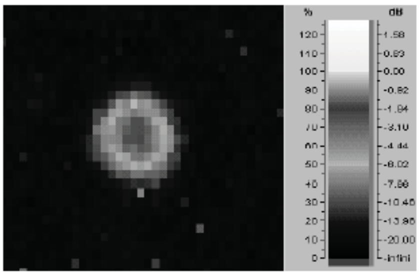


Figure 3. C-scan image containing a circular defect.

To detect small inclusions, the probes must be focused. The image used in this work was formed with measurements taken with focused transducers of 5 MHz on a Ti-Al part with steel inclusions (Figure 4).

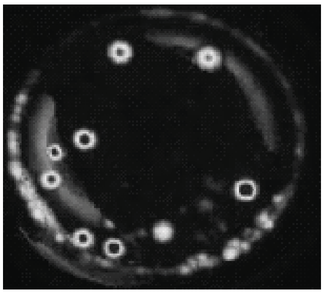


Figure 4. C-scan on a Ti-Al part with steel  inclusion.

To reduce the objects in the c-scan image to their real sizes, we must take into account the surface of the ultrasonic beam. Figure 5 illustrates the principle behind the –6dB method. The probe is moved along the defect, and the distance over which the signal amplitude is above half of the maximum corresponds to the length of the defect.

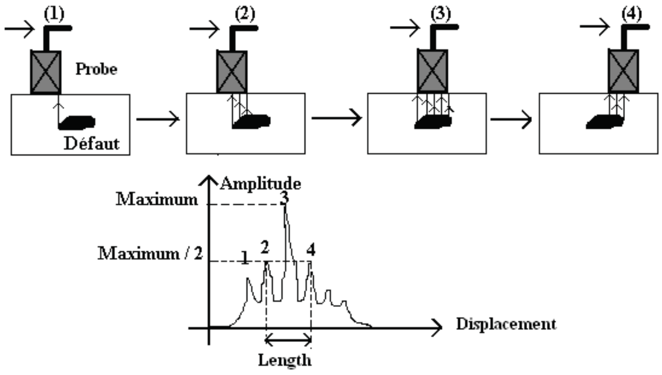
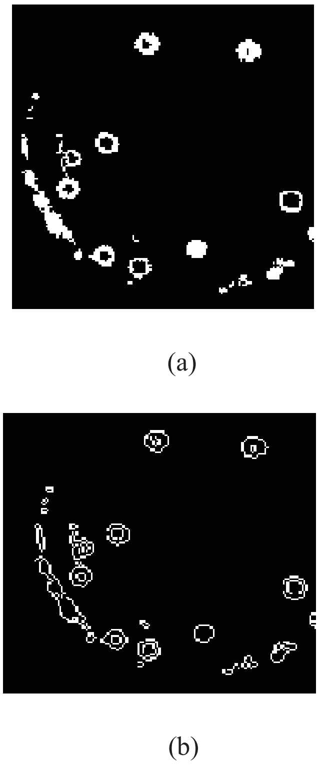


Figure 5. -6dB method for sizing defects.

Applying this method to the image in Figure 4 results in cancelling all the pixels for which the value is lower than half of the maximum signal. The remaining pixels take the value 1 to produce a binary image (Figure 6 (a)). The edges of the defects, which are sufficient for the application of the Hough transform, are then extracted, as shown in Figure 6 (b).



(a)

(b)

Figure 6. (a) Binarized image. (b) Resulting edges.

## III. THE HOUGH TRANSFORM

### *A. Principle of the Hough Transform*

The Hough transform was initially developed for the detection of points aligned on a straight line. More generally, it can detect more complex shapes as long as we know how to define a mathematical model for these shapes. The Hough transform consists of considering all the possible positions and orientations of the shape (for example, all the possible values for the parameters, a and b, of a straight line defined by the equation, y = ax + b, in the x-y plane). Then, for each position and orientation (defined by a particular value of the parameters, a and b), the number of contributing points in the source image is counted. In the decision step, all straight lines defined by the parameters (a,b) such that the number of contributing points in the source image is greater than a threshold are determined.

### *B. Circle Hough Transform*

The Hough transform can be used to determine the parameters of a circle when the number of points that fall on the perimeter are known [11]. A circle with radius r and center (a,b) can be described by the following parametric equations:

$$x = a + r\cos(\theta) \qquad (1)$$

$$y = b + r\sin(\theta) \qquad (2)$$

When the angle θ sweeps through 360 degrees, the points (x,y) trace the perimeter of a circle.

If an image contains many points, some of which fall on perimeters of circles, then the job of the search program is to find parameter triplets (a,b,r) to describe each circle. The fact that the parameter space is 3D makes a direct implementation of the Hough technique more expensive in terms of both computer memory and time. In this work, a graph partitioning separates different populations of pixels [12]. Applied separately on each partition, the Hough transform in this case needs only to detect the co-ordinates (a,b) of the center; the radius is then calculated by taking into account the perimeter of the object .

### *C. Application of the Hough Transforms*

For each pixel $(x_i,y_i)$ of a given population of pixels, the Hough transform associates a circle in the Hough space with the co-ordinates $(x_i,y_i)$. The radius was calculated previously using the length of the edges; the radius is the perimeter divided by 2π.

At the same time, the intersections of the circles are counted in an accumulator, and one is added for each cell crossed.

At the end of this process, the Hough space is analyzed by looking for the points with the maximum number of intersections. Relatives maxima correspond to the circles detected.

## IV. RESULTS

Applied to the image of the edges, as shown in Figure 6, the Hough accumulator is obtained, as shown in Figure 7.

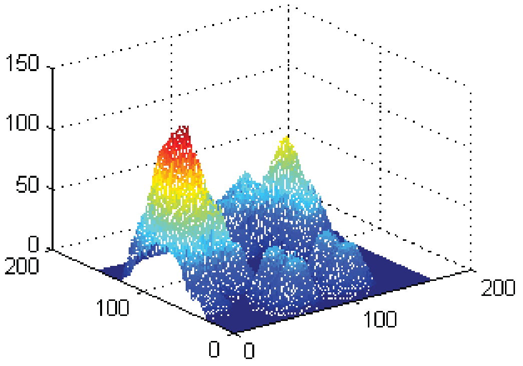


Figure 7. Hough accumulator.

The decision must take into account the highest value cell, which corresponds to the biggest inclusion. Thus, the lowest peaks are eliminated, depending on the precision needed. For this example, a threshold equal to half of the maximum peak was chosen. After detecting the coordinates of the other relative maxima, the corresponding circles were presented on the same image (Figure 8).

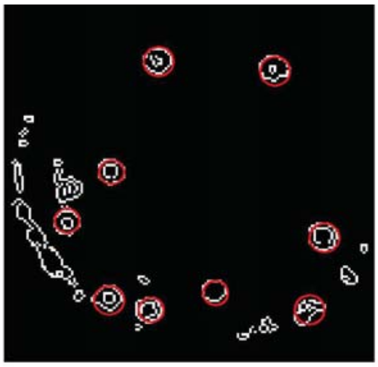


Figure 8. Detected inclusions.

## V. CONCLUSION

In this work, the Hough Transform, a classical tool for pattern recognition, was applied to automatically detect inclusions in a structure, which are characterized by their circular forms.

Graph partitioning, which involves separating different populations of pixels, allowed us to leave the radius out of the parameters to consider. With a reduction to two parameters (the coordinates of the centre), the Hough space shows most of circles corresponding to inclusions. The overall result is promising.

## REFERENCES

[1] B. Eriksson and T. Stepinski, Ultrasonic Characterization of Defects, Part 2. Theoretical Studies, SKI Report 95:21, Swedish Nuclear Power Inspectorate 1995.

[2] M. Krause ; B. Milmann; F. Mielentz et al. " Ultrasonic imaging methods for investigation of post-tensioned concrete structures: A study of interfaces at artificial grouting faults and its verification . Journal of Nondestructive Evaluation", vol. 27, issue 1-3, pp. 67-82 sept 2008.

[3] K. Mayer; Karl-Joerg L.; M. Krause et al . "Characterization of reflector types by phase-sensitive ultrasonic data processing and imaging". Vol. 27, Issue 1-3, pp. 35-45, Sept. 2008.

[4] T.Stepinsky, Data representations for eddy current and ultrasonic Applications. Thesis, Uppsala University, 2000.

[5] C .H Chen . "Advanced Image Processing Methods for Ultrasonic NDE Research ". World Congress of Non Destructive Testing, Proc. WCNDT 2004 , Aug. 30-Sep. 3,2004, pp. 39-43. Montreal, Canada.

[6] H.Maitre, "Uun Panorama de la Transformée de Hough". Revue Traitement du signal, vol. 2, n° 4 pp. 305-317, 1985.

[7] T. Merazi Meksen and B.Boudraa, M. Boudraa , "Automatic crack detection and characterization during ultrasonic inspection". Journal of Nondestructive Evaluation, Vol.29, n°38, pp. 169-174 Sept.2010.

[8] J.Moysan ,G.Corneloup ,"Detection et dimensionnement de défauts en imagerie ultrasonore '', proc. of XIII the colloque sur le traitement du signal et d'images, pp. 181-184, Juan les Pins, GRETSI, 1991

[9] J.Moysan, "Imagerie Ultrasonore pour la Detection Automatique de Défauts en Contrôle non Destructif", thèse de l'institut national des sciences appliquées de Lyon 1992.

[10] P.Bolland, Traitement d'images ultrasonores: Application de la transformée de Hough aux échos de diffraction. Thesis 1999.

[11] D.Ioannou, W.Huda, Circle Recognition through a 2D Hough Transform and radius histograming, Image Vision Compt, 17, 1999, pp. 15–26.

[12] A.Gupta, "Fast and Effective Algorithms for Graph Partitioning and Sparse Matrix", IBM Journal of Research and Development,1996.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# System Identification and Robust Anti-windup Control for a Helium Liquifier

Jing Na, Guang Li, Ryuji Maekawa and Luigi Serio

*Abstract*— **This paper addresses the system identification and advanced control of a helium (He) liquifier supplying cooling power at liquid helium (LHe) temperatures (-269 °C). To study the dynamic response to heat load variation, a He liquifier simulation model is utilized. The main focus is to regulate the discharge pressure of the compressor station to guarantee stable system operation. System identification is first conducted to obtain plant models, and a two degree of freedom $H_\infty$ controller is designed to achieve regulation. Moreover, saturation of the control valve is compensated via an anti-windup technique, which is suitable for regulation problem with disturbance rejection. The effectiveness of the proposed control designs is demonstrated by dynamical simulations in EcosimPro (*EA International*) software.**

*Index Terms*— **Helium liquefier, anti-windup, robust control, disturbance rejection.**

## I. Introduction

ITER [1] that is now under construction at Cadarache, South of France is designed to demonstrate the scientific and technical feasibility of nuclear fusion as a primary source of virtually inexhaustible energy. It is the biggest fusion energy research project, and one of the most challenging and innovative scientific endeavors in the world today [2]. The machine requires high magnetic fields to confine and stabilize the plasma. For such a facility, a cryogenic system will be employed to cool-down and maintain the superconductivity state of the magnets. The ITER cryogenic system [3], [4] will be one of the largest cryogenic systems in the world with a refrigeration capacity of 65 kW equivalent at 4.5 Kelvin.

In cryogenic systems, various components (e.g. heat exchangers, valves, turbines, compressors, etc.) are employed, and the associated controllers are required. To facilitate the control design, several dynamic simulators have been developed, e.g. C-PREST for LHD [5], PROCOS for LHC [6], [7]. Based on dynamic simulators, some advanced control methods have also been presented, for example, internal model control (IMC) [8], nonlinear model predictive control (MPC) [9] and other methods [10], [11]. In these schemes, the discharge pressure of the compressor station (CS) is regulated to guarantee the stable system operation. In [8] and [10], system identification method is also utilized. However, the modeling uncertainties are not explicitly considered in the control design and the potential control valve saturation is not studied.

To address the aforementioned issues, this paper proposes an alternative system identification and advanced control design method. A helium liquefier simulation model originally designed in [12] is employed as the plant to be controlled. However, to facilitate the control design, modifications on the CS control configuration have been done (will be explained in Section II). Moreover, to study the dynamic response of the liquefier, time-varying heat loads are applied in the reservoir. The major focus of this study is to regulate the discharge pressure of CS under dynamic heat loads to sustain the system stability.

The frequency-domain system identification is conducted to obtain the mathematical models, which are supposed to cover system dynamics in a wide frequency band compared to those models from the time-domain identification. Then a $H_\infty$ robust controller [13], [14] is designed to accommodate the modeling uncertainties. Moreover, to recover the control performance in the presence of the control valve saturation, a robust disturbance rejection anti-windup framework (DRAW) recently developed in [15] is employed. The salient feature is that the modeling uncertainties and the disturbance rejection can be incorporated into the AW compensator design and synthesis. Simulation results in a commercial software, EcosimPro [16], are provided to illustrate the performance improvement over alternative approaches. Hence the main contribution of this paper is to provide a systematic robust anti-windup control design procedure for CS pressure regulation of He liquifier with dynamic heat loads, based on frequency system identification and advanced control strategies.

## II. Helium liquifier

Commercial modeling and simulation software, EcosimPro [16], has recently been used for cryogenic simulations. With this software, a cyo-library has been developed by researchers [6], [7], [8] at the European Organization for Nuclear Research (CERN). In this library, various cryogenic component models are constructed based on the mass flowrate and energy balance and the helium properties are obtained from the helium library HEPAK [17].

### A. Helium liquefier model

A helium liquefier designed by CERN [12] is adopted in this study. The liquefier consists of a warm compressor station (CS) and a cold-box (CB) including two expansion turbines (TU), four heat exchangers (HX), a LHe reservoir and connecting pipes and valves. The overall process and instrument diagram (P &ID) is depicted in Fig.1. The CS model and CB model and their associated control structure simulated in EcosimPro are shown in Fig.2 and Fig.3. The entire model contains 1799 differential and algebraic equations (DAEs) with 157 derivatives, 21 boundary variables and 10 control loops.

J. Na, R. Maekawa and L. Serio are with ITER Organization, Route de Vinon sur Verdon, 13115 Saint Paul Lez Durance, France. {Jing.Na, Ryuji.Maekawa,Luigi.Serio}@iter.org

G. Li is with School of Engineering, Computing and Mathematics, University of Exeter, Harrison Building, Streatham Campus, North Park Road, Exeter, EX4 4QF, UK. g.li@exeter.ac.uk

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

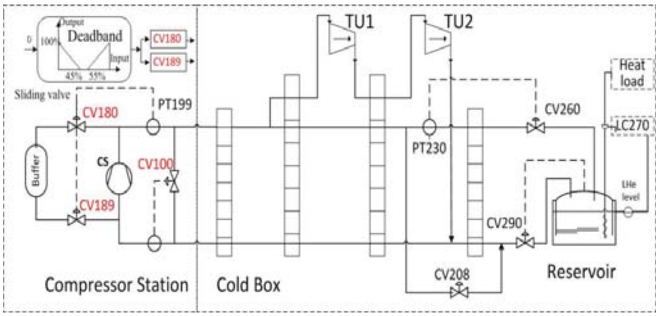


Fig. 1. P&ID of cryogenic system.

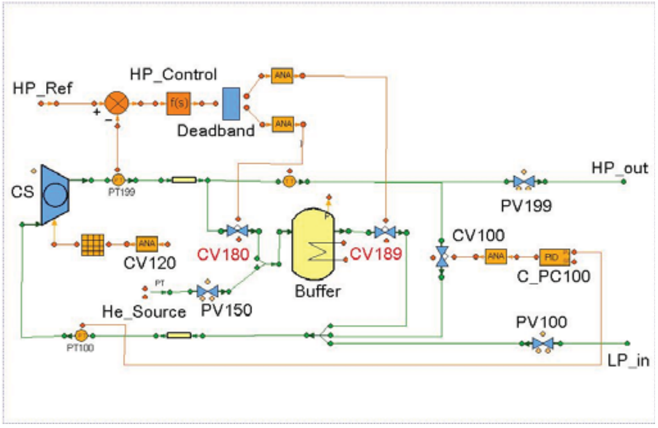


Fig. 2. Compressor station model.

The compressor station compresses helium from 0.1 MPa to 1.3 MPa providing mass flowrate 100 g/s. The discharge pressure or high pressure (HP) is regulated by two antagonist valves: the discharge valve (CV180) that removes the high pressure He into the buffer and a charge valve (CV189) that charges the low pressure from the buffer. It should be pointed out that in the original model proposed in [12], two separated PI controllers with different setpoints are utilized for valves CV180 and CV189. However, in this study the control configuration of CV180 and CV189 is modified to reduce the complexity of control design and to improve the performance. As shown in the left top of Fig.1, the HP is regulated by controlling a split-range valve with a deadband between MV=45%-55%, where the deadband takes the controller output as the input and the corresponding stem positions of the discharge valve CV180 and charge valve CV189 as the outputs. In this case, only one controller (i.e. HP_Control in Fig.2) is required for operating the valves CV180 and CV189 simultaneously. Moreover, the low pressure (LP) is regulated by a bypass valve CV100 connecting HP and LP.

In the cold-box (Fig.3), the inlet valves of turbines (TU1, TU2) are controlled based on their input pressures and the output temperatures. The Joule-Thomson (J-T) valve is controlled taking the inlet pressure as the control variable. The LHe level in the reservoir is regulated by means of an embedded electrical heater with a PI control.

### B. Control problem formulation

To simulate the dynamic response of the liquifier (as shown in Fig.1), the time-varying heat loads are directly applied in the reservoir [10]. In this case, the pressure fluctuations caused by the LHe evaporation (due to the heat loads) may deteriorate the system stability or even trigger instability (e.g. turbine trips) in the worst case. To sustain system stability and the LHe product, as explained in [11], several conditions should be guaranteed: 1) the HP and LP need to be regulated at the *a prior* designed constants (or within fixed intervals) to ensure the reliability of compressor; 2) the LHe level in the reservoir has to stay within an acceptable level to avoid drying or overflow.

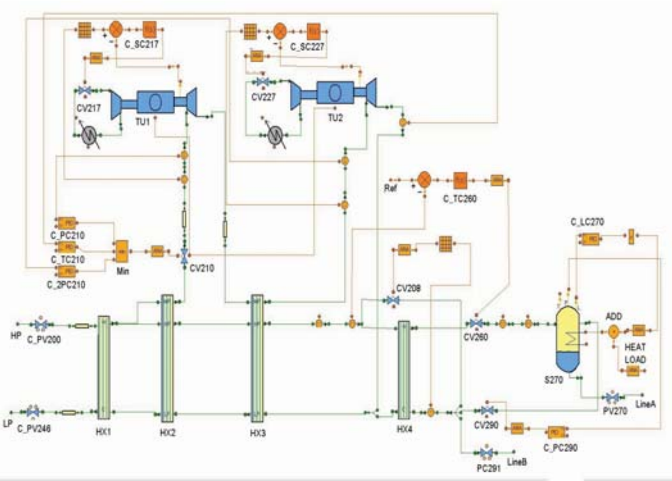


Fig. 3. Cold-box model.

To fulfil the requirement 1), we propose an alternative control design strategy for the liquifier (Fig.1) to regulate the HP of CS under dynamic heat loads. In particular, the control valve saturation is considered via an anit-windup (AW) compensator. Other control loops in the cold box are operated with PI controls guaranteeing also condition 2).

## III. Plant Model Identification

As shown in Fig.1, the HP regulation can be achieved by controlling the sliding valve (CV180/CV189). The heat load applied in the LHe reservoir is taken as the disturbance $d(t)$, the HP of CS is taken as the output $y(t)$, and the position of sliding valve (CV180/CV189) is taken as the input $u(t)$ which is constrained within $[0,100]$ (i.e. valves should be opened between 0% - 100%). Then the overall system dynamics can be described as

$$Y(s) = G_u(s)U(s) + G_d(s)D(s) \qquad (1)$$

where $Y(s), U(s)$ and $D(s)$ are the Laplace transform of $y(t), u(t)$ and $d(t)$, respectively. The transfer function $G_u$ denotes the dynamics from the control input $u$ to output $y$ and $G_d$ denotes the dynamics from the disturbance $d$ to output $y$.

Note there are some nonlinearities (e.g. deadband) in the system (Fig.1). For ease of system identification and control design, linear transfer function (1) is used to represent the main system dynamics and the modeling uncertainties will be considered in the control design in terms of robust control scheme. To obtain models $G_u(s)$ and $G_d(s)$ under wide operation conditions, a frequency identification method based on the correction analysis [18] is adopted. The identification procedure is briefly summarized as follows:

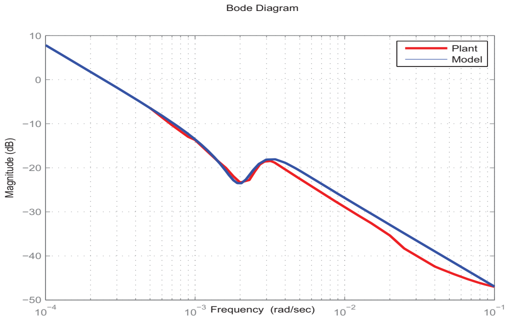


Fig. 4. Process model $G_u(s)$ (2) validation.

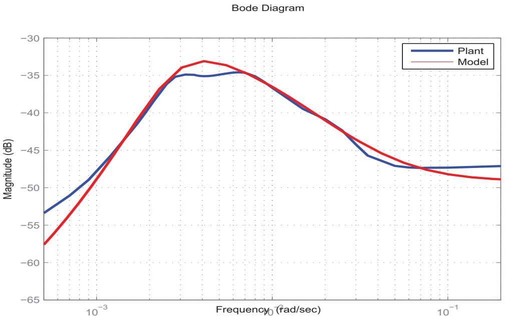


Fig. 5. Disturbance model $G_d(s)$ (3) validation.

1) Apply sinusoidal signals $u(t) = a_1 + b_1 sin(\omega t)$ as the system input and/or $d(t) = a_2 + b_2 sin(\omega t)$ as the disturbance with $a_i > 0, b_i > 0$, and then record the output $y(t)$;

2) Conduct the correlation analysis for data set $(u(t), y(t)), t \in [0, T]$ to deduce the auto-correlation functions $R_u(\omega)$, $R_d(\omega)$, $R_y(\omega)$ and cross-correlation functions $R_{uy}(\omega)$, $R_{dy}(\omega)$. Then calculate the magnitude $H_u(\omega)$, $H_d(\omega)$ and phase $\phi_u(\omega)$, $\phi_d(\omega)$ corresponding to frequency $\omega$ by using $R_u(\omega)$, $R_d(\omega)$, $R_y(\omega)$ and $R_{uy}(\omega)$, $R_{dy}(\omega)$.

3) Change the frequency $\omega$, and repeat calculations 2) for a number of frequencies in the studied frequency band.

4) Plot the magnitude and phase responses $\phi_i(\omega), H_i(\omega), i = u, d$ versus frequency $\omega$.

5) Derive the mathematical models from the plotted Bode diagrams via the curve fitting method.

In this study, the signal $u(t) = 51 + 10sin(\omega t)$ with $\omega$=0.0005∼0.8 rad/s is used as the input and $d(t) = 200 + 80sin(\omega t)$ Watt with $\omega$=0.0001∼0.3 rad/s is employed as the disturbance. 45 group of data sets for $(u(t), y(t))$ and 37 group of data sets for $(d(t), y(t))$ are recorded. Then following the above identification steps, we can derive the process model as

$$G_u(s) = K_u \frac{(s^2 + 2\xi_1 \omega_{n1} s + \omega_{n1}^2)}{s(s^2 + 2\xi_2 \omega_{n2} s + \omega_{n2}^2)} \tag{2}$$

with $\xi_1 = 0.2$, $\omega_{n1} = 0.002$, $\xi_2 = 0.4$, $\omega_{n2} = 0.0027$, $K_u = 4.5 \times 10^{-4}$, and the disturbance model as

$$G_d(s) = K_d \frac{s(1 + \frac{1}{T_1 s})(1 + \frac{1}{T_2 s})}{(s^2 + 2\xi \omega_n s + \omega_n^2)(1 + \frac{1}{1 + T_3 s})} \tag{3}$$

with $K_d = 1.995 \times 10^{-5}$, $T_1 = 0.0008$ ,$T_2 = 0.05$, $T_3 = 0.007$, $\xi = 0.6$, $\omega_n = 0.003$.

For comparison, Fig.4 and Fig.5 provide the magnitude responses versus frequency of process model (2) and disturbance model (3). It is shown that the derived models can capture main dynamics of the liquifier among the studied frequency band, which illustrates the validity of the proposed identification results.

## IV. $H_\infty$ CONTROL AND ANTI-WINDUP COMPENSATION

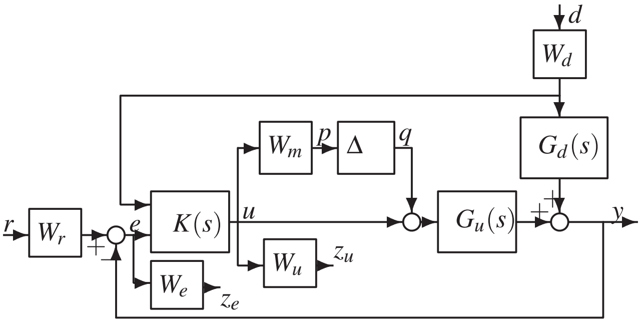


Fig. 6. Block diagram of the $H_\infty$ control for liquifier

In practice, there are usually unavoidable modeling uncertainties based on process model (2) and disturbance model (3) (as shown in Fig. 4 and Fig. 5). To accommodate unmodeled dynamics, $H_\infty$ synthesis method [13], [14] is used to design a 2-degree-of-freedom (DOF) controller $K(s)$. Moreover, since the sliding valve (CV180/CV189) can only be opened between 0-100%, an anti-windup compensator [15] is further incorporated to compensate for the valve saturation.

### A. $H_\infty$ controller design

The scheme for $H_\infty$ controller design of the liquifier is illustrated in Fig. 6. In this case, the plant model is supposed to be an input-multiplicative uncertainty model

$$\tilde{G}_u(s) = G_u(s)(1 + W_m(s)\Delta(s)), \tag{4}$$

where the uncertainty $\|\Delta(s)\|_\infty \leq 1$ and $W_m(s)$ is a weighting function. This uncertainty represents unmodeled dynamics from the control input to output (e.g. difference between the blue line and red line in Fig. 4). The uncertainties from the disturbance to output is not considered in $H_\infty$ controller design, since this uncertainty does not influence the robust stability of the feedback control system. A 2-DOF controller $K(s) = [K_d(s), K_e(s)]$ is designed, where $K_d(s)$ is the feedforward controller and $K_e(s)$ is the feedback controller. The controller $K(s)$ aims to minimize the mapping from $w = [d, r]^T$ to $z = [z_e, z_u]$, represented by $\mathscr{T}_{zw}$. In Fig. 6, the following relations hold

$$\begin{aligned} y &= G_u q + G_d W_d d + G_u u, && \\ u &= K_d d + K_e e, & e &= W_r r - y, \\ z_e &= W_e e, & z_u &= W_u u, \\ p &= W_m u, & q &= \Delta p. \end{aligned}$$

From the above relations, we can derive the following equation for $H_\infty$ control synthesis

$$\left[\begin{array}{c} p \\ \hline z_e \\ z_u \\ \hline d \\ e \end{array}\right] = \left[\begin{array}{cccc} 0 & 0 & 0 & W_m \\ -W_e G_u & -W_e G_d W_d & W_e W_r & -W_e G_u \\ 0 & 0 & 0 & W_u \\ 0 & W_d & 0 & 0 \\ -G_u & -G_d W_d & W_r & -G_u \end{array}\right] \left[\begin{array}{c} q \\ \hline d \\ r \\ \hline u \end{array}\right] \tag{5}$$

Here the uncertainty weighting function is chosen as $W_m = \frac{2s+0.05}{s+1}$, which represents an increase of unmodelled uncertainties from 5% at low frequency to 200% above 1 rad/s. The tracking performance weighting function is $W_e = \frac{0.25(s+0.4)}{s+2\times 10^{-5}}$. The input weighting function is $W_u = \frac{20(s+0.025)}{s+1}$; the magnitude of the weight increases above 0.025 rad/s to limit the closed-loop bandwidth. Those weighting functions should be chosen as a tradeoff between the tracking performance and input magnitude. The disturbance weighting function is chosen as $W_d = \frac{s+1}{s+0.005}$ to describe the high amplitude of the disturbance below 0.005 rad/s. The reference weighting function is $W_r = 1$. Using these weighting functions, an $H_\infty$ controller is synthesized using the commercial routine `hinfsyn` in robust control toolbox of MATLAB®.

### B. Anti-windup (AW) compensator design

In order to recover the performance during the control valve saturation (due to the valve hardware constraints), a separate AW compensator is needed [19], [20]. The AW compensator only provides compensation for the control signal when saturation occurs. In this paper, the AW approach recently proposed in [15] is adopted, which can explicitly incorporate the plant uncertainties and disturbance rejection into synthesis, and thus can make a tradeoff between the robustness and performance. The AW approach proposed in [15] is based on the framework shown in Fig. 7. It is assumed that the model has an additive uncertainty $\Delta_a$ which is related to (4) by

$$W_a = G_u W_m \tag{6}$$

The plant is $G = [G_d, G_u + W_a\Delta]$ and the controller $K$ is the $H_\infty$ controller designed in the last section. Suppose the right co-prime factorization of $G_u(s)$ is $N(s)M^{-1}(s)$, then the AW compensator takes the form of

$$\begin{bmatrix} M(s)E - I \\ N(s) \end{bmatrix} \sim \left[\begin{array}{c|c} A_p + B_pF & B_pE \\ \hline F & E - I \\ C_p + D_pF & D_pE \end{array}\right] \tag{7}$$

Here $G_u \sim (A_p, B_p, C_p, D_p)$ with $A_p \in \mathbb{R}^{n_p \times n_p}$. $F \in \mathbb{R}^{1\times n_p}$ and $E \in \mathbb{R}$ are the variables to be designed by the AW synthesis approach. The input of the AW compensator is the difference between the input and output signals of the saturation operator, and the AW compensator provides two compensation signals, which are added to the input and output signals of the feedback controller. The essential concept of this AW approach to minimize the $\mathscr{L}_2$ gain of

$$\frac{1}{\gamma_d}\left\| \begin{matrix} W_y^{\frac{1}{2}} y_d \\ W_r^{\frac{1}{2}} u \end{matrix} \right\|^2 - \gamma_d \|d\|^2 \leq 0 \tag{8}$$

while incorporating the saturation nonlinearity into synthesis. In (8), the $\mathscr{L}_2$ gains of the mapping from the disturbance $d$ to $y_d$ and the mapping from $d$ to $u$ should be minimized simultaneously. The mapping from $d$ to $y_d$ reflects the performance: a smaller $y_d$ can bring a better performance; the mapping from $d$ to $u$ reflects the robustness: a smaller $u$ leads to a larger margin for robustness. $W_y$ and $W_r$ are weighting matrices, which are positive diagonal and used to tradeoff the gains of the two mappings.

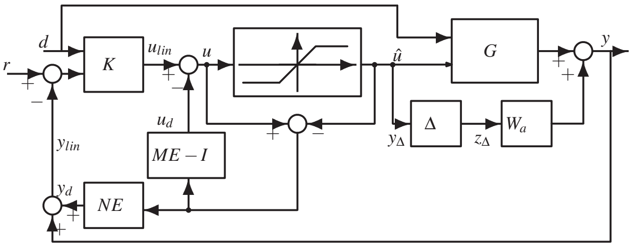


Fig. 7. Anti-windup framework with additive uncertainty

For completeness, the main steps for this AW compensation synthesis procedure is presented (see [15] for more theoretical background and [21] for a real application case.)

1) Compute a disturbance filter

$$P_d := (I - K_y G_u)^{-1}([K_d, w_d] + K_e[G_d, 0])$$

whose state space realization is $(A_d, B_d, C_d, D_d)$ with $A_d \in \mathbb{R}^{n_d \times n_d}$.

2) Given the matrix variable $P = P^T > 0$, solve $\gamma_d^* := \min \gamma_d > 0$ subject to LMI-1 in (9) and LMI-2 in (10) to yield $P^*$ and $\gamma_d^*$.

$$\mathbf{LMI-1}: \quad \begin{bmatrix} PA_o + A_o^T P + PW_A + W_A^T P & W_C + PW_B \\ W_C^T + W_B^T P & W_D \end{bmatrix} < 0 \tag{9}$$

with

$$A_o = \begin{bmatrix} A_p & 0 \\ 0 & A_d \end{bmatrix} \qquad W_A = \begin{bmatrix} 0 & B_pC_d \\ 0 & 0 \end{bmatrix}$$

$$W_B = \begin{bmatrix} B_pD_d & 0 & -B_p \\ B_d & 0 & 0 \end{bmatrix} \qquad W_C = \begin{bmatrix} 0 & C_p^T & 0 \\ 0 & C_{\tilde{d}}^T D_p^T & 0 \end{bmatrix}$$

$$W_D = \begin{bmatrix} -\gamma_d I_{n_w} & D_d^T D_p^T & 0 \\ D_pD_d & -\gamma_d I_{n_y} W_y^{-1} & -D_p \\ 0 & -D_p^T & -\Gamma \end{bmatrix}$$

and

$$\mathbf{LMI-2}: \quad \begin{bmatrix} A_d^T P_{22} + P_{22}A_d & P_{22}B_d & C_d^T \\ B_d^T P_{22} & -\gamma_d I_{n_w} & D_d^T \\ C_d & D_d & -\gamma_d W_r^{-1} \end{bmatrix} < 0 \tag{10}$$

Here the variables are the scalar $\gamma_d > 0$, diagonal matrix $\Gamma = \text{diag}(\gamma_1, \ldots, \gamma_{n_u}) > 0$, and $P$ is a symmetric positive definite matrix with a structure

$$P := \begin{bmatrix} P_{11} & P_{12} \\ P_{12}^T & P_{22} \end{bmatrix} \in \mathbb{R}^{n_p + n_d} \tag{11}$$

3) Substituting $P^*$ and $\gamma_d^*$ with some chosen diagonal positive definite $W$, solve the LMI:

$$\Psi + H^T \Lambda G + G^T \Lambda^T H < 0 \tag{12}$$

for $\Lambda$, with $\Lambda := \begin{bmatrix} F & E \end{bmatrix}$ and

$$\Psi = \begin{bmatrix} A_o^T P + PA_o & PB_o + C_{do}^T \tilde{W} & C_{po}^T \\ B_o^T P + \tilde{W} C_{do} & \tilde{W} D_{do} + D_{do}^T \tilde{W} - \gamma_d \tilde{I}_{n_w} & 0 \\ C_{po} & 0 & -\gamma_d I_{n_y} \end{bmatrix},$$

$$H = \left[\begin{array}{cc|cc|c} B_p^T & 0 & -1 & 0 & D_p^T \end{array}\right] \text{diag}(P, \tilde{W}, I),$$

$$G = \left[\begin{array}{cc|cc|c} I_{n_p} & 0 & 0 & 0_{n_p\times 1} & 0 \\ 0 & 0 & 1 & 0 & 0 \end{array}\right],$$

where $\tilde{W} = \begin{bmatrix} W & 0 \\ 0 & 1 \end{bmatrix}$, and

$$B_o = \begin{bmatrix} 0_{n_p\times 1} & 0_{n_p\times 1} \\ 0_{n_d\times 1} & B_d \end{bmatrix} \qquad C_{do} = \begin{bmatrix} 0_{1\times n_p} & C_d \\ 0_{1\times n_p} & 0 \end{bmatrix}$$

$$D_{do} = \begin{bmatrix} 0 & D_d \\ 0 & 0 \end{bmatrix} \qquad C_{po} = \begin{bmatrix} C_p & 0 \end{bmatrix}$$

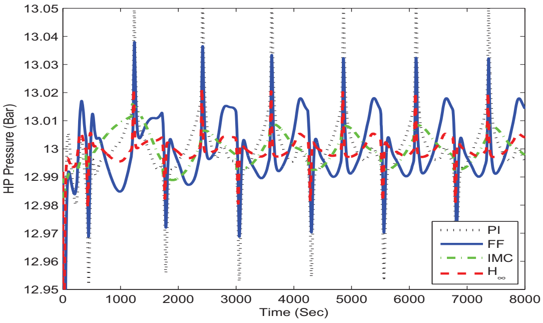


Fig. 8. HP regulation steady-sate performance.

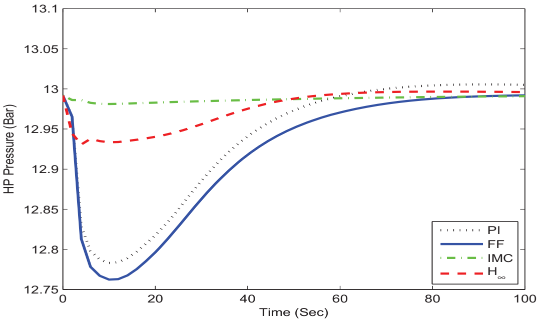


Fig. 9. HP regulation transient performance.

Note that the possible algebraic loop issue (in case $E \neq 1$) can be resolved using the idea as in [15].

## V. Simulation and Discussion

### A. Simulation in EcosimPro

To verify the proposed control method, real-time simulation results in the software EcosimPro are given. Several cases are simulated:

- **Time-varying disturbance**: The simulation is first conducted with a time-varying heat load $d(t) = 80\sin(0.005t) + 200$ applied in the reservoir. Several control methods, PI control (PI), feedforward control (FF), internal model control (IMC) and the proposed $H_\infty$ control, are tested. Comparative simulation results are provided in Fig.8 and Fig.9. It is shown that the IMC can achieve best steady-state performance (i.e. smallest regulation error) and transient performance (i.e. fastest convergence speed). It is reasonable since the plant model and its inverse model are all directly utilized in the control implementation. Among other approaches, the $H_\infty$ control gives smaller transient overshoot as well as the steady-state regulation error. The PI control has similar transient to FF control but gives larger steady-state error.
- **Turbine trip**: To evaluate the compensation for the windup effect, simulations are also performed with a turbine trip between 550-650Sec as [8]. Comparative regulation results are depicted in Fig.10, for which the IMC controller allows disturbances rejection with a faster recovery speed. However, the IMC imposes aggressive control behavior, i.e. with valve stem position fluctuations, which is shown in Fig.11. On the other hand, the proposed $H_\infty$ control with AW gives smaller overshoot and steady-state regulation error and smoother transient performance compared to FF and PI. For more details, the corresponding controller outputs are also given in Fig.12, which shows that the proposed $H_\infty$ control with the AW compensation almost avoids the saturation issue, i.e. the control output is within [0,100].

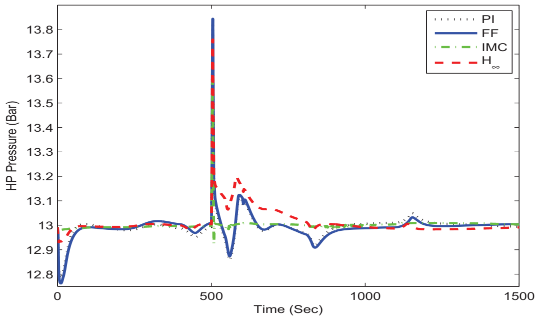


Fig. 10. HP regulation performance with a turbine trip.

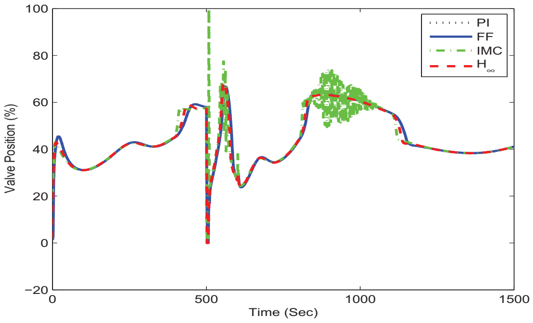


Fig. 11. Sliding valve position with a turbine trip.

### B. Discussion

From above simulations, it can be seen that the PI, FF, IMC and the proposed $H_\infty$ control with AW have their own characteristics:

**PI control**: It has a simple control structure for implementation, and no plant model is used. Although the disturbance rejection can be improved with high gains, the oscillated transient is obtained or even the stability may be lost in the worst case. Moreover, the integral windup will degrade the control performance in the presence of control constraints.

**FF control**: The disturbance model and the plant model inverse are superimposed in the feedback control as an extra compensation for the disturbance. Consequently, the improved steady-state regulation can be expected. The problem is that the disturbances are assumed to be precisely measurable, and the plant models are assumed to be exactly known. Moreover, the robustness to modeling uncertainties

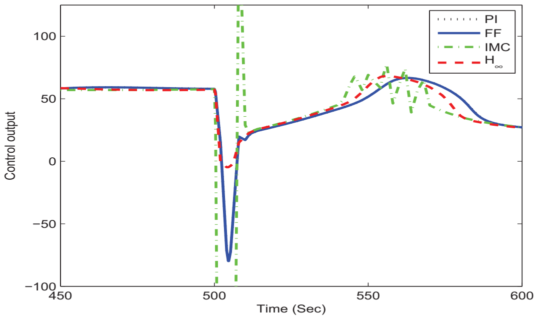


Fig. 12. Control output with a turbine trip.

are not considered.

**IMC control**: The plant model and its inverse are employed. Note in this paper the inverse model with a low pass filter is used as the feedback control, such that the disturbance rejection is guaranteed without any information on the disturbance measurement and its model. The robustness can be theoretically studied. Best output regulation performance has been achieved among the tested controllers. The complexity of the control design and implementation is moderate. However, the utilization of the model inverse as the feedback control action may introduce high gain and thus cause unexpected fluctuations and oscillation in the control actions (see Fig.11). This may be further remedied by redesigning the feedback control.

$H_\infty$ **control with AW**: The modeling uncertainties can be explicitly considered in the $H_\infty$ control synthesis, and 2-DOF feedback control also allows for the disturbance rejection with guaranteed robustness and performance. The valve control saturation is specifically compensated, which can further improve the transient performance when saturation occurs. However, a control with relatively high order may be derived following this synthesis, and the subsequent order reduction is necessary in the practical applications.

## VI. Conclusions

System identification and advanced robust control is studied for a helium liquifier to regulate the discharge pressure of compressor station. Time-varying heat loads are applied in the reservoir to simulate dynamic responses. An anti-windup framework is adopted to compensate for the control valve saturation.

Comparative simulations conducted in the software EcosimPro reveal that the proposed robust AW control scheme works well in the presence of control valve saturation. It gives designers more freedom to tradeoff the robustness and performance (i.e. the modeling uncertainties and disturbance rejection are explicitly considered in the control synthesis). However, the control order reduction needs to be studied before applying this control scheme to actual system. On the other hand, among the tested controllers (PI, FF, IMC and $H_\infty$), IMC can achieve good regulation and disturbance rejection performance in terms of transient and steady-state. Nevertheless, the possible high gain control due to the model inverse may trigger the valve saturation and thus result in aggressive control actions, which may be overcome by redesigning the feedback control part. Moreover, the frequency-domain identification approach provides fairly precise models that cover system dynamics among a wide frequency band.

## Acknowledgment

The authors would like to thank Dr. Benjamin Bradu and Mr. William Booth in CERN for sharing the cryogenic models and also for their support in the collaborative work. The authors also appreciate all staffs in the ITER cryogenic system section for their supports.

## Disclaimer

This proceeding was prepared as an account of work by and for the ITER Organization which Members are the People's Republic of China, the European Atomic Energy Community, the Republic of India, Japan, the Republic of Korea, the Russian Federation, and the United States of America. The views and opinions expressed herein do not necessarily reflect those of the ITER Organization, its Members or any agency thereof. Dissemination of the information in this proceeding is governed by the applicable terms of the ITER Joint Implementation Agreement.

## References

[1] R. Aymar, P. Barabaschi, and Y. Shimomura. The ITER design. *Plasma physics and controlled fusion*, 44:519–565, 2002.

[2] L. Serio et al. Challenges for cryogenics at ITER. In *AIP Conference Proceedings*, volume 1218, pages 651–633, 2010.

[3] V. Kalinin, E. Tada, F. Millet, and N. Shatil. ITER cryogenic system. *Fusion engineering and Design*, 81(23-24):2589–2595, 2006.

[4] M. Kalinin V. Henry D. Sanmarti M. Serio, L. Chalifour and B. Sarkar. conceptual design of the cryogenic system for ITER. In *Proceedings of the 22nd International Cryogenic Engineering Conference*, 2008.

[5] R. Maekawa, K. Ooba, M. Nobutoki, and T. Mito. Dynamic simulation of the helium refrigerator/liquefier for LHD. *Cryogenics*, 45(3):199–211, 2005.

[6] B. Bradu, P. Gayet, and S.I. Niculescu. A process and control simulator for large scale cryogenic plants. *Control Engineering Practice*, 17(12):1388–1397, 2009.

[7] B. Bradu, P. Gayet, and S.I. Niculescu. Modeling, simulation and control of large scale cryogenic systems. In *Proceedings of the 17th IFAC World Congress*, pages 13265–13270, 2008.

[8] B. Bradu, P. Gayet, and S.I. Niculescu. Control optimization of a LHC 18 kw cryoplant warm compression station using dynamic simulations. In *AIP Conference Proceedings*, volume 1218, pages 1619–1627, 2010.

[9] E. Blanco, C. de Prada, S. Cristea, and J. Casas. Nonlinear predictive control in the LHC accelerator. *Control Engineering Practice*, 17(10):1136–1147, 2009.

[10] R. Maekawa, S. Takami, and M. Nobutoki. Adaptation of advanced control to the helium liquefier with C-PREST. In *Proceedings of the 22nd International Cryogenic Engineering Conference*, pages 243–248, 2008.

[11] F. Clavel, M. Alamir, P. Bonnay, A. Barraud, G. Bornard, and C. Deschildre. Multivariable control architecture for a cryogenic test facility under high pulsed loads: Model derivation, control design and experimental validation. *Journal of Process Control*, 21(7):1030–1039, 2011.

[12] B. Bradu. Tutorial: Modeling of cryogenic systems with Ecosimpro. Technical report, CERN. Geneva, Switzerland, 2011.

[13] K. Zhou, J. C. Doyle, and K. Glover. *Robust and optimal control*. Prentice Hall, Upper Saddle River, N.J., 1996.

[14] S. Skogestad and I. Postlethwaite. *Multivariable feedback control analysis and design*. John & Sons, Ltd, 2005.

[15] G. Li, G. Herrmann, D. P. Stoten, J. Tu, and M. C. Turner. A novel robust disturbance rejection anti-windup framework. *International Journal of Control*, 84(1):123–137, 2011.

[16] EA Internacional. Ecosimpro 4.4: Mathematical algorithms and simulation guide. Technical report, Madrid, Spain., 2007.

[17] Cryodata Inc. Users guide to HEPAK version 3.4. Technical report, Cryodata Inc, Louisville, Colorado, USA, 1999.

[18] L. Lennart. *System identification: theory for the user*. PTR Prentice Hall, Upper Saddle River, NJ, 1999.

[19] A. Zheng, V.K. Mayuresh, and M. Morari. Anti-windup design for internal model control. *International Journal of Control*, 60(5):1015–1024, 1994.

[20] P. F. Weston and I. Postlethwaite. Linear conditioning for systems containing saturating actuators. *Automatica*, 36:1347–1354, 2000.

[21] G. Li, G. Herrmann, D. P. Stoten, J. Tu, and M. C. Turner. Application of robust antiwindup techniques to dynamically substructured systems. *IEEE/ASME Transactions on Mechatronics*, Accept for publication, 2011.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Trajectory Tracking of Exothermic Batch Reactor using NIR Spectroscopy

**Olufemi Osunnuyi* Ognjen Marjanovic* Jian Wan* Barry Lennox***

* *School of Electrical and Electronic Engineering, The University of Manchester*

**Abstract:** The control of exothermic chemical batch reactors has received much attention in literature over the years for their increasing importance in manufacturing industries and also the unique quality control challenges that they provide. However, most of the control schemes proposed to deal with these challenges make use of models that implicitly control the product quality. For example it is assumed that a control scheme that successfully regulates the reactor temperature along an *apriori* calculated optimal profile should imply satisfactory quality trajectory control. In this paper it is shown that this assumption is not robust enough to deal with some kinds of disturbances that may occur during the batch. It is also shown that product quality control can be greatly improved by proposing a new control scheme that makes use of NIR spectroscopic measurements as feedback information for a quality control system. The results of two controllers using this scheme are compared with a more widely used implicit control strategy in two test cases with unmeasurable system disturbances.

***Keywords: Batch process, Near Infrared Spectroscopy***

## I. INTRODUCTION

Batch and semi-batch processes are gaining increasing importance in manufacturing industries. Due to the flexibility they offer they are prevalent in the polymer, pharmaceutical and specialty chemicals industries where the focus is on the production of low-volume, high-value added products. An example of a batch process is the chemical batch reactor where the batch operation consists of charging the reactor, controlling the reactor temperature to meet some processing criterion, shutting down and then emptying the reactor. There is a steady increase of heat required during the initial phase of the batch to obtain the desired reaction temperature after which appropriate cooling is required to maintain the required temperature for the remainder of the batch [1, 2]. Previously not much attention was paid into how the reactor temperature reached the desired set points. However, due to work done on obtaining optimal reactor profiles so as to increase the yield of the desired product, research work subsequently focused on trajectory tracking control of batch reactors [3-5].

However, challenges occur when trying to implement reliable trajectory tracking control systems in batch processes due to the absence of steady state operation which requires the controller to track a highly transient profile within a short operating regime [6]. Some other unavoidable inherent characteristics include presence of time-varying and nonlinear dynamics, multitude of unmeasured disturbances such as concentrations of various raw materials, and the presence of irreversible behaviour which all inevitably affect the reliability of overall quality control [7]. There has been quite a bit of literature dedicated to improving the quality of control in this regard. For instance, the capability to track the optimal reference profile of the reactor temperature for various kinds of dual-mode controllers was compared in [1]. Also, a multivariate control strategy based on empirical dynamic PCA models for the trajectory tracking of the optimal profile was developed in [8].

All these publications however, have been based on the assumption that the successful regulation of the reactor temperature should result in good control of the quality of the desired product. However, even if the adequate temperature control system is in place and the reactor temperature does follow closely its reference profile, there is no guarantee that the final product will meet its specifications [9]. For example if a disturbance occurred that altered the dynamic relationship between the product quality and the reactor temperature, the production of the product quality will depreciate despite successful regulation of the reference temperature profile. Hence, the control objective should be focused on controlling variables that are more directly related to the product quality.

In a quite a number of process industries, Near-infrared (NIR) spectroscopic measurements are widely used in providing a more direct quantification for the quality of the product. NIR spectroscopy is based on the absorption of electromagnetic radiation at wavelengths in the range 780–2500 nm [10]. This frequency range covers mainly overtones and combinations of the lower-energy fundamental molecular vibrations making the NIR intensities significantly weaker than the fundamental bands from which they originate [11]. Thus, the low molar absorptivity of NIR bands permits operation in the reflectance mode and hence the non-destructive recording of spectra of solid samples with minimal or no pre-treatment, thereby substantially increasing the throughput [12, 13]. Uses of NIR spectroscopy are widespread in the pharmaceutical industry to test raw materials, control product quality and monitor processes [12-15]. NIR analytical techniques have also found popularity in the food industry being routinely used for the compositional, functional and sensory analysis of food ingredients, process intermediates and final products [10].

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

Hence, the integrity of the quality of information provided by the NIR spectra can be exploited as essential feedback information used by the control system to enable enhanced quality production.

In this paper a quality control scheme that uses NIR spectroscopic measurements for the control of an exothermic batch reactor is proposed. In addressing the evident problem of dealing with the enormous amount of NIR spectra data two different controllers were designed and both compared with a conventional implicit reactor temperature controller using two different case studies.

## II. METHODLOGY

### A. Introducing NIR Measuring Point

The most widely used method of controlling exothermic batch reactors involves regulating the reactor temperature along an optimal profile which has been calculated offline. The intention being that this optimal profile will help yield the product quality trajectory that maximises the conversion of the desired product. One of the most common forms of this implicit control scheme is the Temperature Cascaded Control (TCC) shown in *Fig 1*. TCC uses cascade control for the control of the reactor temperature. The slave controller controls the jacket temperature, whose setpoint is determined by the reactor temperature controller.

It can be observed in *Fig 1* that the control loop is closed with the reactor temperature and hence will only reject disturbances that occur within the loop ($m_{d1}$) i.e. on $G_1$. An example of this disturbance could be a change in the heat transfer coefficient. However, when dealing with disturbances occurring outside this loop ($m_{d2}$), i.e. on $G_2$, the regulation of the reactor temperature proves incapable of tracking the desired trajectory for product quality. The reason being that the disturbance $m_{d2}$ changes the underlying relationship between $T_R$ and $M_{ABCD}$; hence the reactor temperature trajectory ought to change to yield the desired conversion of *Mc*. The best way of solving this problem is to introduce a new measuring point in the reactor that measures a variable that is directly related to the product quality, thus enabling the incorporation of another control loop using this new variable as feedback as shown in *Fig 2*.
This relationship ensures that the setpoint to the reactor temperature is dynamic, as opposed to the TCC scheme. In this paper the NIR spectroscopic measurements are used as this feedback information.

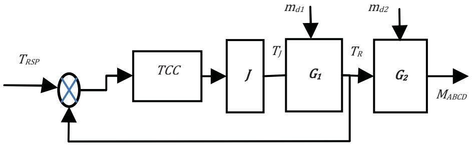


*Figure 1: Block diagram of Temperature Cascaded Control on Batch Process*

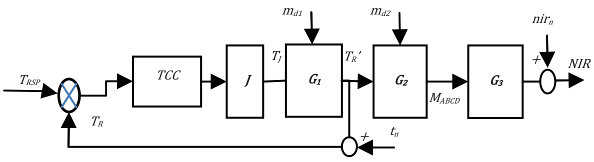


*Figure 2: Block Diagram of Batch Process producing NIR Spectra*

### B. Creation of Artificial NIR Spectra

The measured NIR Spectra is constructed under the assumption of Beer-Lambert Law for low concentration absorption media. This law enables us to obtain the NIR Spectra as the weighted sum of the individual linear mappings of the concentrations through their respective 'pure NIR spectra' [16]. These individual linear spectral mappings represent the NIR spectra of each component. And their sum is the NIR spectra of the entire batch run.
Assuming a matrix $M \in \mathbb{R}^{N\times m}$ which contains $m$ different compounds for $N$ number of samples in a batch run and $\theta \in \mathbb{R}^{A\times m}$ as the pure spectra matrix for these compounds with $A$ number of wave channels, then a complete batch of NIR Spectra ($NIR \in \mathbb{R}^{N\times A}$) can be calculated according to:

$$NIR(i,:) = \frac{M(i,:)}{(\sum_j^m M(i,j))} * \theta^T \qquad (1)$$

In order to obtain the pure spectra, an artificial spectra generator was designed to create pure spectra not related to any particular compound but sufficiently shaped to resemble forms of common pure spectra used in industry.
It was observed with non-artificial pure spectra that quite often the 'spectral peaks' were very similar in shape to standard Gaussian distribution. This basic assumption was essential in developing the code that would produce reasonable pure NIR Spectra for the purposes of the case studies considered in this paper.
Therefore in order to use this code in constructing a particular pure spectrum the following would need to be specified: the desired *number of spectral channels* (or wave numbers), the *number of Gaussian distributions* and the *maximum* and *minimum widths* (i.e. the width range) of these 'bell-shaped' Gaussian distributions. With randomly selected wave channels acting as mean centred positions ($\mu$) and with standard deviations ($\sigma$) also randomly chosen from the Gaussian width range, each Gaussian was generated on single a spectrum using the Gaussian function:

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (2)$$

where $x$ represents the individual wave numbers.
The normalised sum of all these spectra incorporates all the individual Gaussians and thus forms a single pure NIR Spectrum. This procedure is repeated for as many compounds are present in each specific simulation. An example of artificially generated pure spectra for four compounds is shown in *Fig 3*.

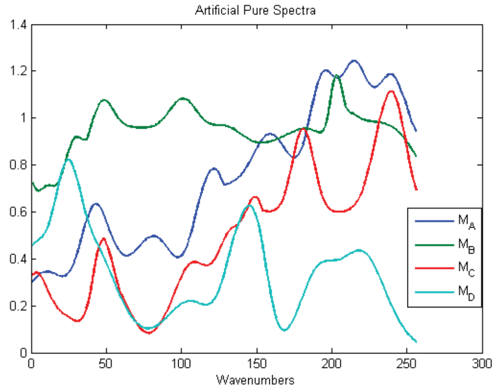


*Figure 3: Sample of Artificially generated pure NIR Spectra*

## C. Quality Control Methodology

### 1) NIR Data Reduction for SISO Control

The control scheme used to regulate the product quality is achieved by closing the loop at the new NIR spectra measuring point as shown in *Fig 4*.

When closing the new control loop a new block $Y$ is included because of the large amount of feedback information that is typical of NIR measurements. In this scheme block $Y$ is used to reduce the number of variables that is fed back to the new quality controller $C_q$. In this paper, in which a SISO simulation is considered for the case study example, block $Y$ represents two different data reduction techniques.

The first method of data reduction is a *wavenumber-selector* in which the most appropriate wavenumber for control is used. The selection of the appropriate wavenumber is very important because a wrong selection will lead to very poor control. One way of selecting the most appropriate wavenumber to control is by selecting the peak amplitude occurring in an NIR sample. However, this approach is based on trial and error. In this paper, the wave number from the NIR reference spectra with the largest variance was used. The reason for this choice is that the highest varying wavenumber will contain the best qualitative information of $G_2$ and also the highest amplitude does not necessarily imply the highest varying wavenumber as shown in *Fig 5*.

The second data reduction technique used was making block $Y$ a Principal Component Analysis (PCA) model. This reduces each fed back NIR spectrum into a specified number of latent PCA variables based on the extraction of the largest possible variance in the data set.

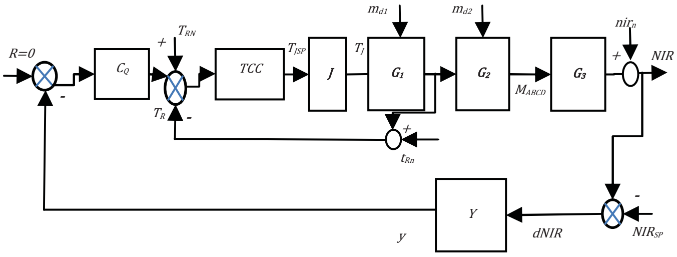


*Figure 4: Quality Controller implemented on Batch Process using NIR Spectral measurements*

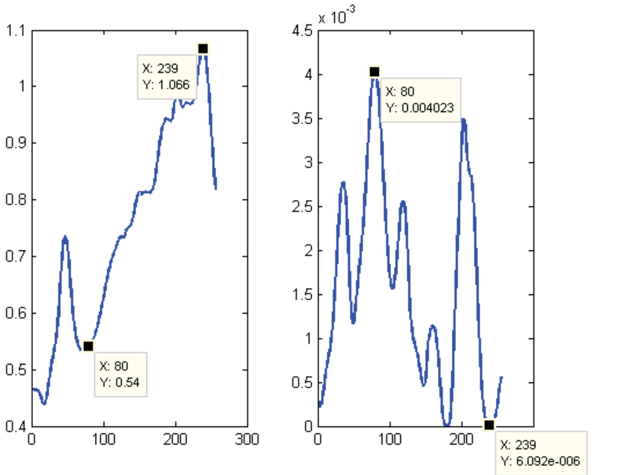


*Figure 5: Single NIR spectrum and variation spectrum of entire batch spectra*

In the case study of this paper, the results of controllers using these two choices of $Y$ were compared with that of TCC.

### 2) Deviation Control

Finally, rather than tracking the selected NIR reference wavenumber or the corresponding PCA score trajectory the control requirement in this quality control loop is to minimise the deviation from the quality variable trajectory. This takes into account the advantage of possessing the nominal reactor temperature trajectory which is subsequently added to the setpoint of the quality controller. Quite often the Achilles heel of trajectory tracking control is the wide operating range that the process variables cover during the batch.

By subtracting this dominant transient from the NIR spectrum before feedback the control requirement is in effect reduced to a less complex setpoint tracking and causes the speed of the controller's response to increase. In other words the batch control problem becomes similar to a continuous control problem.

# III. CASE STUDY

## A. Specification of Model

As a benchmark case to test the control scheme shown in *Fig 4* on, modifications were made to the nonlinear exothermic batch reactor model used by Cott and Macchietto [2] (schematically shown in *Fig 6*) to incorporate NIR Spectral measurements. This model has also been used by Aziz [1] and all of the model parameter values are the same as used in the latter paper except for the time constant of the jacket temperature $T_J$ ($\tau_J$) which is now 5mins as opposed to 3mins and the rate constant 1 for reaction 1 ($k_1^1$) was changed from 20.9057 to 20.8057. The sample time is 30s.

The control objective in the initial papers that have used this model is to control the reactor temperature by adjusting the setpoint to the inlet jacket temperature.

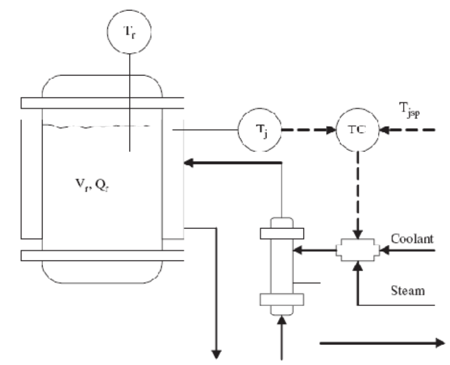


*Fig 6: Schematic of a chemical batch reactor* [17]

The reactions taking place are as follows:

$$A + B \xrightarrow{R_1} C \tag{3}$$

$$A + C \xrightarrow{R_2} D \tag{4}$$

Where *A* and *B* are the raw materials, *C* is the desired product, *D* is the waste product and $R_i$ is the reaction rate for reaction *i*. The dynamic equations of the process that form block $G_2$ are:

$$\frac{dM_A}{dt} = -R_1 - R_2 \tag{5}$$

$$\frac{dM_B}{dt} = -R_1 \tag{6}$$

$$\frac{dM_C}{dt} = +R_1 - R_2 \tag{7}$$

$$\frac{dM_D}{dt} = +R_2 \tag{8}$$

$$R_1 = k_1 M_A M_B \tag{9}$$

$$R_2 = k_1 M_A M_C \tag{10}$$

$$k_1 = e^{\left(k_1^1 - \frac{k_1^2}{(T_R + 273.15)}\right)} \tag{11}$$

$$k_2 = e^{\left(k_2^1 - \frac{k_2^2}{(T_R + 273.15)}\right)} \tag{12}$$

Where $M_i$ (*kmol*) is the number of moles for component *i*, $k_i^j$ is rate constant *j* for reaction *i* and block $G_3$ which converts raw materials and products into a combined NIR Spectra is as given in equation (1).

The reader is referred to the original publications for a more complete description of the model which include the description of block $G_1$.

## *B. Test Case Description*

The evaluation of the performance of the three controllers will be carried out on two different test cases. The controllers to be compared are:

a) *TCC* (PID)

b) *Wn-PID* (i.e. $C_q$ being a PID controlling a single wave number)

c) *Sc-PID* (i.e. $C_q$ being a PID controlling the first PCA score only).

The two simulated test cases are to evaluate the controllers' ability to reject disturbances occurring at $G_1$ ($m_{d1}$) and $G_2$ ($m_{d2}$) respectively:

TEST 1: A drop in the heat transfer coefficient, which occurs due to deterioration of the thermal jacket over a period of time and causes a change in the operating conditions. This will test the robustness of the controllers in reacting to a change in an unmeasured parameter.

TEST 2: A 2% increase in the rate constant 1 for reaction 1($k_1^1$). This causes a disturbance in the reaction rate of reaction 1 and therefore fundamentally changes the model relationship between the reactor temperature and the reactants (and hence products) for which the controllers were nominally designed for.

## *C. Controller Assessment*

Performance Index (PI): Under most industrial applications constructing a *PI* using the measurement of $M_C$ is not feasible due to the extremely high cost of purchasing quality analysers. However, because a simulation example is being used it was convenient assessing the performance of the controllers using $M_C$ and at the same time easily adapting this *PI* to the NIR spectra as well.

The MAE (mean absolute error) of the $M_C$ of each controller from the nominal trajectory was calculated as follows:

$$MAE_C = \Sigma_i^N \frac{|M_C(i) - \overline{M}_C(i)|}{N} \tag{13}$$

The same was calculated for all the wavenumbers of the NIR spectra. And to combine these metrics into a singular measure ($MAE_{NIR}$) the square root of the sum of the squared scaled mean absolute errors was used.

## *D. Results and discussion*

TEST 1: *Fig 7* and *Fig 8* show the responses of all three controllers to a change in the heat transfer coefficient. When the heat transfer coefficient decreases there is less flow of heat from the inlet jacket to the reactor which results in a rapid rise in the reactor temperature. The controllers are required to respond quickly to this change to avoid a large overshoot of the reactor temperature which would cause the off-specification production of the desired product. In both *Fig 7* and *Fig 8* it is observed that the quality controllers responded quicker to this change than the TCC. The extra control loop enables additional product quality-related feedback information to the jacket temperature and a quicker adjustment than that of the TCC.

Also, of the two quality controllers *Fig 7* and *Fig 8* show that the Sc-PID reacts more rapidly to the disturbance when compared to the Wn-PID. This is due to Sc-PID using information that more comprehensively represents the system dynamics of the reactor than the Wn-PID. This advantage is clearly represented in *Fig 7b* where the output of Sc-PID signals a further and quicker lowering of the jacket temperature than the Wn-PID. The effect of this as seen in *Fig 8a* is a lower jacket temperature drop particularly at the end of the heating phase of the reactor, resulting in a lower reactor temperature overshoot and better tracking of the optimal profile. Overall, the effect of tracking the desired product trajectory is shown in *Fig 8b* to be significantly better when using Sc-PID and the performance indices in *Table 1* show considerably lower values for the quality controllers –with Sc-PID being the lowest –than that of the implicit TCC.

TEST 2: *Fig 9* and *Fig 10* give the responses of all three controllers to a disturbance in the reaction rate due to an increase $k_1^1$ responsible for producing *C*. With this disturbance occurring at $G_2$ in *Fig 4*, the TCC lacks any essential feedback data informing it to adjust the reactor temperature to suit the new dynamic relationship. Therefore as shown in *Fig 9a* the TCC successfully tracks the now sub-optimal profile which results in very poor tracking of the desired product trajectory and a selected wavenumber (80) as show in *Fig 10a* and *10b*. Using an implicit product quality control scheme (such as TCC) essentially loses the batch in such conditions. Whereas the product quality-related information supplied by the NIR spectroscopic measurements enables the quality controllers to create a new optimal trajectory for the reactor temperature as seen in *Fig 9a*. By reducing the overall temperature of the reactor in the presence of the disturbance there is a closer tracking of the quality variables as seen in *Fig 10a* and *Fig 10b*.

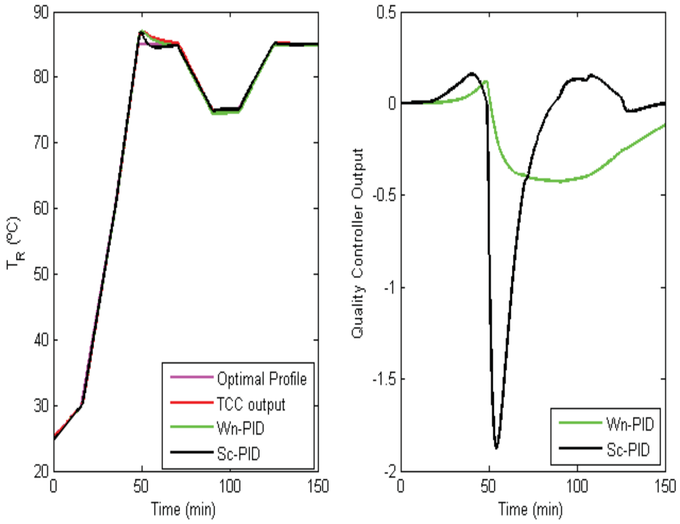


*Figure 7: a) Responses to change in heat transfer coefficient in Reactor Temperature and b) Quality controller outputs*

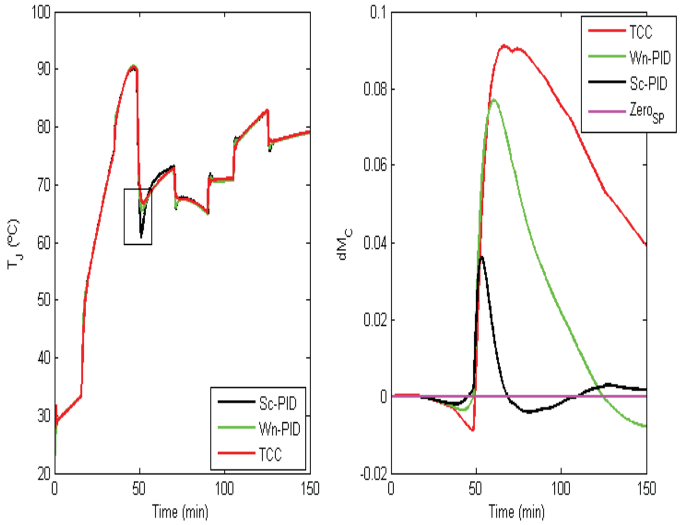


*Figure 8: a) Responses to change in heat transfer coefficient in Jacket Temperature and b) Deviation from nominal trajectory in $M_C$*

However, a much better performance of the Sc-PID in comparison to the Wn-PID is again noticed. There is a much quicker and settled response to the presence of this disturbance from the Sc-PID when observing the quality controller outputs of *Fig 9b*. The reason for this once again is that the first PCA score, which reflects the entire NIR Spectra, contains a lot more system information than any single wave number even when it is the highest varying wavenumber used.

These performances are again reflected in the *MAE* values of *Table 1*. The quality controllers far outperform the TCC, with the Sc-PID also proving to be better than the Wn-PID. It is worth noting that when the same test was carried out with Wn-PID to control the peak wavenumber (i.e. $Wn_{239}$) as opposed to the highest varying wavenumber (i.e. $Wn_{80}$), the performance index values returned were $MAE_C = 0.7216$ and $MAE_{NIR} = 0.4809$ for $Wn_{239}$. This very poor performance highlights the need to ensure the appropriate selection of a wavenumber.

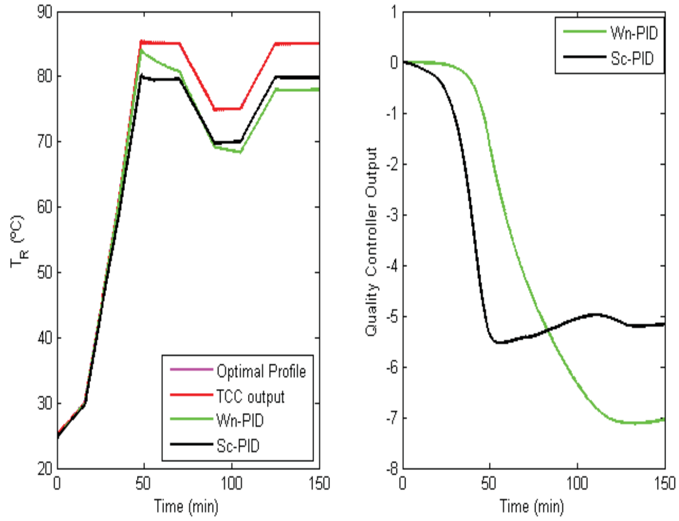


*Figure 9: a) Responses to reaction rate disturbance in reactor temperature and b) Quality controller outputs*

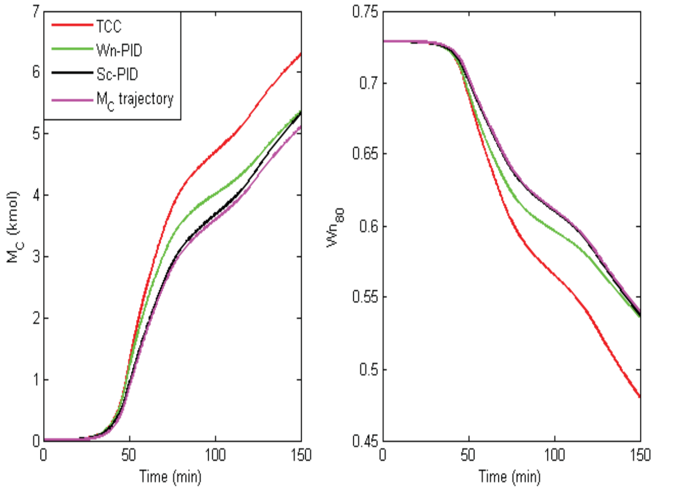


*Figure 10: a) Responses to reaction rate disturbance in $M_C$ and b) Wavenumber 80*

## IV. CONCLUSION

This paper has investigated the use of NIR spectroscopic measurements as feedback for trajectory tracking of the product quality of exothermic batch reactors. This provides a more direct form of tracking the desired product quality as opposed to the implicit form of quality control quite often employed.

Instead of the control objective being the trajectory control of the reactor temperature, the objective was modified to minimising the deviation of a product quality related variable from its nominal trajectory. This scheme not only enables the controller to respond to various disturbances affecting different components of the reactor dynamics, but also greatly increases the speed of the controller's response as well.

These were shown by comparing the results of a standard Temperature cascade control (TCC) scheme with two controllers that utilise NIR spectra as feedback information. The first of these two controllers is Wn-PID, which minimises the deviation of the highest varying wavenumber of the nominal NIR batch from its setpoint trajectory, and the second being Sc-PID, which minimises the deviation of the first PCA score from its nominal trajectory. The results of the two tests carried out reveal that the quality controllers far outperform the TCC. However, the Sc-PID consistently outperformed Wn-PID owing to the fact that the first PCA score contains significantly more information than any single wave number.

| Disturbance Type | Controller | $MAE_C$ | $MAE_{NIR}$ |
|---|---|---|---|
| HTCC | TCC | *0.0459* | *0.0256* |
| | Wn-PID | *0.0197* | *0.0097* |
| | Sc-PID | *0.0038* | *0.0020* |
| RR-disturbance | TCC | *0.7103* | *0.4009* |
| | Wn-PID | *0.2802* | *0.1258* |
| | Sc-PID | *0.0778* | *0.0318* |

*Table 1: Table of Mean absolute errors (in $M_C$ and NIR) for the controllers*

## V. REFERENCES

1. Aziz, N., M.A. Hussain, and I.M. Mujtaba, *Performance of different types of controllers in tracking optimal temperature profiles in batch reactors.* Computers and Chemical Engineering, 2000. **24**: p. 1069-1075.
2. Cott, B.J. and S. Macchietto, *Temperature control of exothermic batch reactors using generic model control.* Industrial & engineering chemistry research, 1989. **28**(8): p. 1177-1184.
3. Aziz, N. and M.A. Hussain, *Optimal control of batch reactors.*, in *IChemE advance in process control conference*. 1998.
4. Logsdon, J.S. and L.T. Biegler, *A relaxed reduced space SQP strategy for dynamic optimization problems.* Computers & chemical engineering, 1993. **17**(4): p. 367-372.
5. Luus, R., *Optimal control of batch reactors by iterative dynamic programming.* Journal of process control, 1994. **4**(4): p. 218-226.
6. Shinskey, F.G., *Process control systems: application, design, and tuning*. 1996: McGraw-Hill.
7. Bonvin, D., *Control and optimization of batch processes.* Control Systems Magazine, IEEE, 2006. **26**(6): p. 34-45.
8. Flores-Cerrillo, J. and J.F. MacGregor, *Latent variable MPC for trajectory tracking in batch processes.* Journal of process control, 2005. **15**(6): p. 651-663.
9. Lin, H., et al. *Application of Near-infrared Spectroscopy in Batch Process Control*. in *International Symposium on Advanced Control of Chemical Processes, Istanbul, Turkey*. 2009.
10. Osborne, B.G., *Near-infrared spectroscopy in food analysis*. 1986: Wiley Online Library.
11. Bakeev, K.A., ed. *Process Analytical Technology: Spectroscopic Tools and Implementation Strategies for the Chemical and Pharmaceutical Industries*. 2005, Wiley Blackwell: Oxford.
12. Blanco, M., et al., *Near-infrared spectroscopy in the pharmaceutical industry.* ANALYST-LONDON-SOCIETY OF PUBLIC ANALYSTS THEN ROYAL SOCIETY OF CHEMISTRY-, 1998. **123**: p. 135-150.
13. Reich, G., *Near-infrared spectroscopy and imaging: Basic principles and pharmaceutical applications.* Advanced Drug Delivery Reviews, 2005. **57**(8): p. 1109-1143.
14. Burns, D.A. and E.W. Ciurczak, *Handbook of near-infrared analysis.* Vol. 35. 2008: CRC.
15. Luypaert, J., D.L. Massart, and Y. Vander Heyden, *Near-infrared spectroscopy applications in pharmaceutical analysis.* Talanta, 2007. **72**(3): p. 865-883.
16. Chen, J. and X.Z. Wang, *A new approach to near-infrared spectral data analysis using independent component analysis.* Journal of Chemical Information and Computer Sciences, 2001. **41**(4): p. 992-1001.
17. Cho, W., T.F. Edgar, and J. Lee, *Iterative learning dual-mode control of exothermic batch reactors.* Control Engineering Practice, 2008. **16**(10): p. 1244-1249.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Decentralized Nonlinear Control of 300MWe Circulating Fluidized Boiler Power Unit

Xue Yali, Li Donghai, Zhang Yuqiong,Gao Qirui
Thermal Engineering Department,
Stake Key Lab of Power Systems,
Tsinghua University,
Beijing 100084, China
xueyali@tsinghua.edu.cn

Wang Jihong
School of Engineering
University of Warwick
Coventry CV4 7AL, UK
Jihong.Wang@warwick.ac.uk

Sun Zhixin
School of Energy and Power Engineering
Xi'an Jiaotong University
Xi'an 710049, China
zxsun0626@gmail.com

*Abstract* —**The dynamics of large scale circulating fluidized bed (CFB) boiler power plant shows the behaves of large time delay, complex coupling and nonlinearities, which brings difficulties for controller design to improve the plant performance. This paper firstly introduces a well-established and validated nonlinear model of a 300MWe CFB power plant, and then the model is simplified through analysis and approximation which could show clearly the reason of difficulties associated with controller design. It is found that the coupling and nonlinearity mainly come from the opening of main steam governing valve under various load conditions, and it is hard to completely decouple the multivariable system. To avoid the difficulties of decoupling, a decentralized PID control system is finally proposed to control the unit power and main steam pressure after nonlinear model transformation. The simulation results show that the control system has improved dynamic performance in load reference tracking for a wide range of load changes, and also good disturbance rejection to the coal quality variation. The analysis and control in this paper provide the first step to move forward for future control quality improvement and advanced control strategy study.**

***Keywords- CFB control; nonlinearity; decentralized control; PID controller***

## I. INTRODUCTION

As one of the promising clean coal utilization technologies, the circulating fluidized bed (CFB) boiler has gained rapid development in recent years [1]. Compared with traditional pulverized coal-fired (PC) power plant, CFB power plant offers the distinct advantages such as more fuel flexibility, lower pollution emission, wider operation range and higher combustion efficiency. Meanwhile, the wider operation range and complicated combustion process brings more nonlinearities into the unit dynamics. Since the heat release of coal combustion goes through repetitious circulating, the time delay and inertia from fuel disturbance of CFB boiler to the unit load fluctuation is much longer than those of PC power plants. And similar to the PC power plant, both of the fuel feed flow rate and governor valve position of steam turbine have prominent influences on the main steam pressure and unit power output simultaneously. Thus, the control of a CFB boiler power plant is a more challenging task than that of PC power plant, especially for its coordinated control system.

There are many advanced control strategies reported on the coordinated control of thermal power plants, such as active disturbance rejection control [2-3], nonlinear control based on approximate or exact feedback linearization [4-5], gain scheduling control [6], intelligent algorithm-based optimization control [7], etc. Most advanced control strategy simulation shows satisfactory control performance, but in fact it is hard to be implemented in the real power plant due to its complicated control design procedure or inconvenient tuning method. In addition, their applications to CFB power plants are rarely reported.

PID control is the most widely used control strategy in real control engineering of thermal power plants, usually merged with some amelioration elements such as unit load feed-forward compensations and nonlinear elements if necessary. In recent years, direct energy balance (DEB) control is very popular in coordinated control of thermal power plants in China. By integrating gain scheduling principle to eliminate the static gain nonlinearity [8], it can balance the turbine energy demand and boiler heat supply, and consequently improve the control performance under load variations. But the parameter tuning of DEB is of great dependence on engineers' experience and it is time-consuming to obtain final energy balance [2].

Since there is currently no well-accepted coordinate control strategy for CFB power plants, it is important to test any new control strategy in simulation using a full-scale verified nonlinear dynamic CFB power plant model before it can be implemented in practice. Some nonlinear models have been developed for CFB power plant in the last twenty years, and they are valuable in understanding the complex mechanisms of CFB process. However, those models have given very limited contributions to control system design or control engineering practice. The main reason may be that the modeling motivation is not for control study so that its dynamics has not been validated in the full load operation range to reveal the characteristic of nonlinearity, coupling, large time delay and high order inertia quantitatively. In addition, large scale CFB power plant has just accomplished its demonstration several

This work is funded by Tsinghua-Shenhua Strategic Cooperation Basic Research Project.

years before, so its modeling and control problem is still under investigation.

In this paper, a well-established nonlinear model of 300MWe CFB power plant in [9] is introduced. The model was built up to facilitate the coordinated control field adjustment. After reasonable model structure selection and model parameters identification, the nonlinear model is validated by using various load operation data in a real 300MWe CFB power plant. Thus it can be used to explore the coordinated control design of large scale CFB power plants.

The organization of this paper is as follows. In section II, we give an introduction to the nonlinear model of CFB power plant. In section III, the nonlinearity of CFB model is analyzed, followed with model approximation and reconstruction. In Section IV, the decentralized nonlinear control system is designed, and simulation results are presented in Section V to show its control ability on load reference tracking and coal quality disturbance rejection. Finally, the conclusion is given in section VI.

## II. A Nonlinear model of 300 MWe CFB power unit

Based on a nonlinear boiler-turbine model structure proposed in [10], a nonlinear boiler-turbine unit model for a 300MWe CFB power plant is developed [9] in Simulink to facilitate field commissioning of coordinated control system. The model structure is shown in Fig.1, and Tab. 1 gives the symbol annotations.

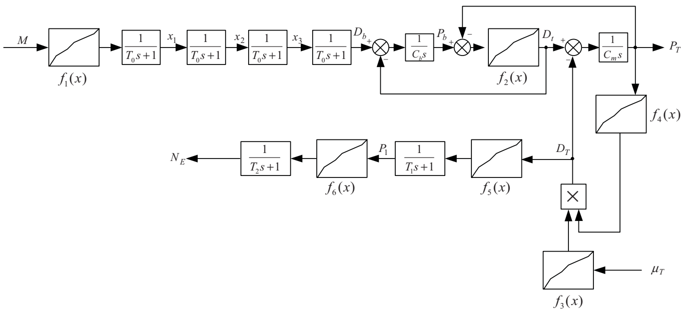


Figure 1. Nonlinear model structure of 300MWe CFB power plant

TABLE I. Symbol Annotation

| Symbol | Annotation |
|---|---|
| $M$ | Fuel flow rate |
| $\mu_T$ | Opening of steam turbine governing valve |
| $P_T$ | Main steam pressure |
| $N_E$ | Generator output power |
| $D_b$ | Heat produced in boiler |
| $P_b$ | Drum pressure |
| $D_t$ | Steam flow rate generated by boiler |
| $D_T$ | Inlet steam flow rate of turbine |
| $P_1$ | First stage pressure of turbine |
| $T_0,T_1,T_2$ | Time constants |
| $C_b$ | Heat storage coefficient |
| $C_m$ | Volume storage coefficient |
| $f_i(x)$,<br>$i=1,\cdots,6$ | Nonlinear mapping functions |
| $x_1,x_2,x_3$ | Intermediate state variables |

In Fig.1, nonlinear mapping function $f_i(x), i=1,\cdots,6$ is used to depict the static gain nonlinearity with load condition changes. They are identified by using the full-scope operating data of a real 300MWe CFB power plant [9]. The time constants $T_0,T_1,T_2$ and coefficients $C_b$ , $C_m$ are also identified from operating data of the same power plant in transient processes. Their values can be found in Appendix Tab. II and III.

Denote the system state $X$ and output variable $Y$ as:

$$\begin{aligned} X &= [x_1,x_2,x_3,D_b,P_b,P_T,P_1,N_E]^{\mathrm{T}} \\ &= [x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8]^{\mathrm{T}} \end{aligned} , \quad (1)$$

$$Y=[P_T,N_E]^{\mathrm{T}}=[y_1,y_2]^{\mathrm{T}} , \quad (2)$$

then the CFB model shown in Fig. 1 can be expressed as follows:

$$\begin{cases} \dot{x}_1 = \dfrac{1}{T_0}[-x_1+f_1(M)] \\ \dot{x}_2 = \dfrac{1}{T_0}(x_1-x_2) \\ \dot{x}_3 = \dfrac{1}{T_0}(x_2-x_3) \\ \dot{x}_4 = \dfrac{1}{T_0}(x_3-x_4) \\ \dot{x}_5 = \dfrac{1}{C_b}[x_4-f_2(x_5-x_6)] \\ \dot{x}_6 = \dfrac{1}{C_m}[f_2(x_5-x_6)-f_4(x_6)\cdot f_3(\mu_T)] \\ \dot{x}_7 = \dfrac{1}{T_1}[f_5(f_4(x_6)\cdot f_3(\mu_T))-x_7] \\ \dot{x}_8 = \dfrac{1}{T_2}[f_6(x_7)-x_8] \end{cases} , \quad (3)$$

$$Y=[x_6,x_8]^{\mathrm{T}} \quad (4)$$

Obviously, it is a nonlinear model due to the involvement of six nonlinear gain functions $f_i(x), i=1,\cdots,6$ .

To ensure that the power units can response quickly to grid load demand while maintaining the key parameters within their permitted limits, the coordinated system of a CFB boiler power plant should manipulate the control variable $M$ and $\mu_T$ to meet the following specifications:

(1) For a variable load reference, the unit load should follow load reference closely, and the main steam pressure fluctuation should be as small as possible.

(2) For a fixed load reference, the internal disturbance from combustion system should have little influence on the main steam pressure and unit load.

## III. MODEL APPROXIMATION

By plotting all the nonlinear mapping functions $f_i(x), i = 1,\cdots,6$, it is noticed that the nonlinearities associated with $f_i(x), i = 1,2,5,6$ are very close to linear relationship, so they can be linearized as:

$$f_i(x) \approx k_i x, i = 1,2,5,6\,, \tag{5}$$

where $k_i, i = 1,2,5,6$ are constant coefficients. In this way, the nonlinear model (3)(4) can be approximated as:

$$\dot{X} = \begin{bmatrix} -\frac{1}{T_0} & 0 & & \cdots & & & & 0 \\ \frac{1}{T_0} & -\frac{1}{T_0} & 0 & & & & & \\ 0 & \frac{1}{T_0} & -\frac{1}{T_0} & 0 & & & & \\ 0 & 0 & \frac{1}{T_0} & -\frac{1}{T_0} & 0 & & & \vdots \\ & & 0 & \frac{1}{C_b} & -\frac{k_2}{C_b} & \frac{k_2}{C_b} & & \\ \vdots & & & \ddots & \frac{k_2}{C_m} & -\frac{k_2}{C_m} & 0 & \\ & & & & 0 & 0 & -\frac{1}{T_1} & 0 \\ 0 & 0 & & \cdots & & 0 & \frac{k_6}{T_2} & -\frac{1}{T_2} \end{bmatrix} X + \begin{bmatrix} \frac{k_1 M}{T_0} \\ 0 \\ 0 \\ 0 \\ 0 \\ -\frac{f_4(x_6) f_3(\mu_T)}{C_m} \\ \frac{k_5}{T_1} f_4(x_6) f_3(\mu_T) \\ 0 \end{bmatrix} \tag{6}$$

$$Y = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} X\,. \tag{7}$$

The simplified model is still a nonlinear model, and now the nonlinearity mainly comes from $f_4(P_T) f_3(\mu_T)$, especially the opening of governor valve $\mu_T$ for the fixed pressure operation.

We select a new control input $D_T$ instead of $\mu_T$ as:

$$D_T = f_4(x_3) f_3(\mu_T)\,, \tag{8}$$

and substitute (8) into (6), then we have

$$\dot{X} = \begin{bmatrix} -\frac{1}{T_0} & 0 & & \cdots & & & & 0 \\ \frac{1}{T_0} & -\frac{1}{T_0} & 0 & & & & & \\ 0 & \frac{1}{T_0} & -\frac{1}{T_0} & 0 & & & & \\ 0 & 0 & \frac{1}{T_0} & -\frac{1}{T_0} & 0 & & & \vdots \\ & & 0 & \frac{1}{C_b} & -\frac{k_2}{C_b} & \frac{k_2}{C_b} & & \\ \vdots & & & \ddots & \frac{k_2}{C_m} & -\frac{k_2}{C_m} & 0 & \\ & & & & 0 & 0 & -\frac{1}{T_1} & 0 \\ 0 & 0 & & \cdots & & 0 & \frac{k_6}{T_2} & -\frac{1}{T_2} \end{bmatrix} X + \begin{bmatrix} \frac{k_1 M}{T_0} \\ 0 \\ 0 \\ 0 \\ 0 \\ -\frac{D_T}{C_m} \\ \frac{k_5}{T_1} D_T \\ 0 \end{bmatrix}. \tag{9}$$

Thus the nonlinear model (6)(7) can be rewritten in a linear state space representation as follows:

$$\begin{cases} X = AX + BU \\ Y = CX \end{cases}, \tag{10}$$

where,

$$A = \begin{bmatrix} -\frac{1}{T_0} & 0 & 0 & & \cdots & & 0 & 0 \\ \frac{1}{T_0} & -\frac{1}{T_0} & 0 & & & & & 0 \\ 0 & \frac{1}{T_0} & -\frac{1}{T_0} & 0 & & & & \\ 0 & 0 & \frac{1}{T_0} & -\frac{1}{T_0} & 0 & & & \vdots \\ & & 0 & \frac{1}{C_b} & -\frac{k_2}{C_b} & \frac{k_2}{C_b} & & \\ \vdots & & & \ddots & \frac{k_2}{C_m} & -\frac{k_2}{C_m} & 0 & 0 \\ & & & & 0 & 0 & -\frac{1}{T_1} & 0 \\ 0 & 0 & & \cdots & & 0 & \frac{k_6}{T_2} & -\frac{1}{T_2} \end{bmatrix} \tag{11}$$

$$B = \begin{bmatrix} \frac{k_1}{T_0} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -\frac{1}{C_m} & \frac{k_5}{T_1} & 0 \end{bmatrix}^{\mathrm{T}} \tag{12}$$

$$C = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{13}$$

$$U = [M \quad D_T]^{\mathrm{T}} \tag{14}$$

## IV. CONTROL SYSTEM DESIGN

Both of the controllability matrix and observability matrix of system (10-14) have full rank, so the system is completely controllable and observable, a full-state feedback controller can be designed to move the initial states to any specified condition in a finite time interval without considering input constraints. Considering the motivation of control system implementation in the real CFB power plants, only PID controllers are adopted in this paper instead.

For linear system (10-14), a two-input-two-output transfer matrix can be derived as follows:

$$\begin{bmatrix} P_T(s) \\ N_E(s) \end{bmatrix} = \begin{bmatrix} g_{11}(s) & g_{12}(s) \\ g_{21}(s) & g_{22}(s) \end{bmatrix} \begin{bmatrix} M(s) \\ D_T(s) \end{bmatrix} \tag{15}$$

where,

$$\begin{aligned} g_{11}(s) &= \frac{k_1 k_2}{s(1+T_0 s)^4 (C_m C_b s + k_2 C_m + k_2 C_b)} \\ g_{12}(s) &= \frac{-(C_b s + k_2)}{s(C_m C_b s + k_2 C_m + k_2 C_b)} \\ g_{21}(s) &= 0 \\ g_{22}(s) &= \frac{k_5 k_6}{(1+T_1 s)(1+T_2 s)} \end{aligned} \tag{16}$$

It has a half-decoupled structure, so a local controller can be tuned separately for the second loop $g_{22}(s)$. Here a PID controller is employed for the sake of simplification:

$$c_2(s) = \frac{k_{p2}}{k_5 k_6} \frac{(1+T_1 s)(1+T_2 s)}{s(1+T_2 s / T_f)}, \tag{17}$$

so that the open-loop transfer function of second loop approximates as

$$g_{2o}(s) = g_{22}(s) c_2(s) \approx \frac{k_{p2}}{s} \tag{18}$$

For the first loop, it is hard to compensate completely the disturbance from $D_T$ to $P_T$ owning to the realization difficulty of decoupling transfer function as well as input constraints. An approximately decoupling controller, such as P or PD controller can be designed to obtain partly decoupling if necessary. Here, a PI controller is designed for the first loop to explore the ability of decentralized control structure:

$$c_1(s) = k_{p1} + \frac{k_{i1}}{s}, \tag{19}$$

The above two controller parameters are tuned as:

$$k_{p1} = 50,\ k_{i1} = 0.1,\ k_{p2} = 1,\ T_f = 10. \tag{20}$$

The closed-loop control structure of CFB coordinated control system is shown in Fig. 2. Since the first loop output $P_T$ is used to derive the second loop control action $\mu_T$, so it is actually a partial decoupled control system even if it is represented in a simple decentralized structure.

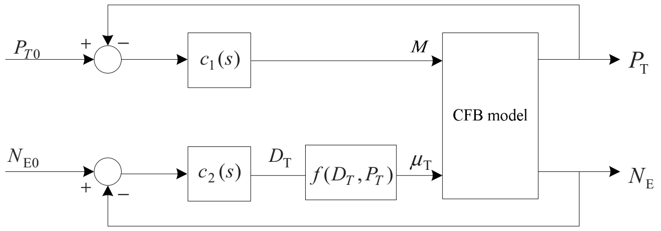


Figure 2. Control structure of CFB unit control

From the nonlinear mapping functions in Appendix, it is clear that nonlinear function $f_3(x), f_4(x)$ are both monotone increasing functions with constraints, their inverse functions exist in region of interest. Then it is realizable to derive the real control action $\mu_T$ from $D_T$ and $P_T$ as follows:

$$\mu_T = f(D_T, P_T) = f_3^{-1}[D_T / f_4(P_T)] \tag{21}$$

## V. Simulation Results

### A. Change load reference

Starting close-loop simulation from 210 MWe (70% load), at 1000 s the load reference increases to 250 MWe with a rate of 0.3 MWe/s, and at 5000 s the load reference decreases to 170 MWe with a rate of 0.3 MWe/s. The dynamic responses of power output and main steam pressure are shown in Fig. 3(a)-(b). It can be found that the load output can rapidly follow its reference closely, while keeps the disturbance on main steam pressure within a small range of ±1 MPa. The dynamics of fuel flow rate and valve opening are also given in Fig. 3(c)-(d), they behave relatively smooth.

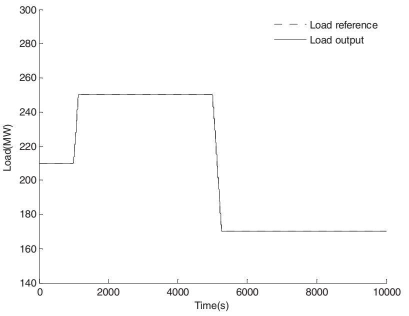


Figure 3(a) Load output response

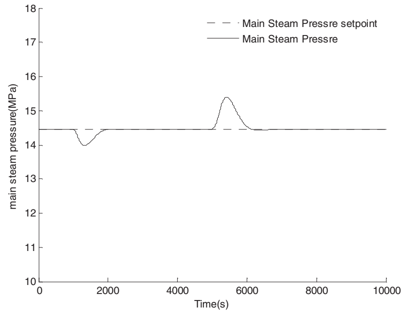


Figure 3(b) Main steam pressure response

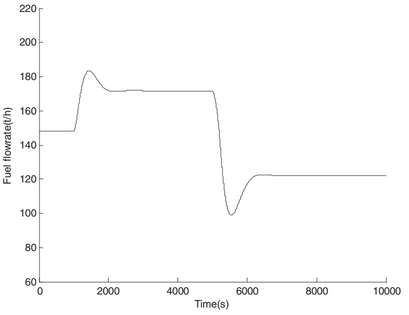


Figure 3(c) Fuel Flow rate response

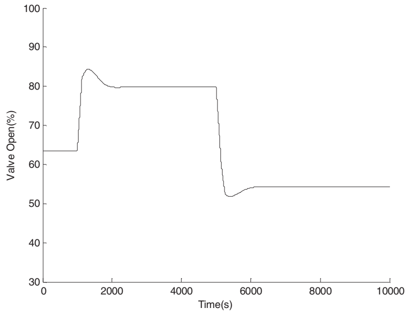


Figure 3(d) Valve Opening response

A simulation of larger load reference change is conducted, which is not shown in this paper for length limits. The simulation results indicates that the coordinated control system shows similarly good load following capacity without retuning the controller parameters for the situation that the manipulation of fuel flow rate and valve opening do not reach their limits.

### *B. Dynamics at fuel flow rate disturbances*

The coal quality variation will influence the operation of power plant greatly by bringing uncertainty to control system. Here, the coal quality variation is treated as a fuel flow rate disturbance. Firstly a step fuel flow rate disturbance of -50 t/h at 1000 s is conducted, then followed with a step fuel flow rate disturbance of 100 t/h at 5000 s, Figure 4(a)-(d) illustrate the dynamic responses of unit load, main steam pressure, fuel flow rate and valve opening.

It is obviously that the coal quality disturbance has no influence on load output owning to the rapid opposite direction manipulation of valve opening. Since the disturbance is added directly to the fuel flow rate, the main steam pressure shows obvious departure from setpoint value and has a positive correlation with the amplitude of fuel flow rate disturbance.

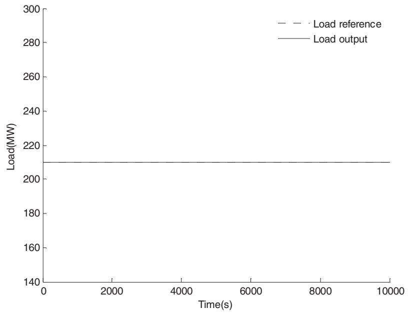


Figure 4(a) Load response at fuel flow disturbance

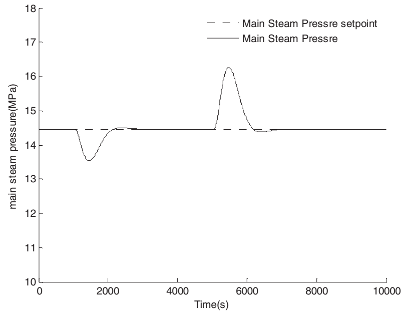


Figure 4(b) Main steam pressure response at fuel flow disturbance

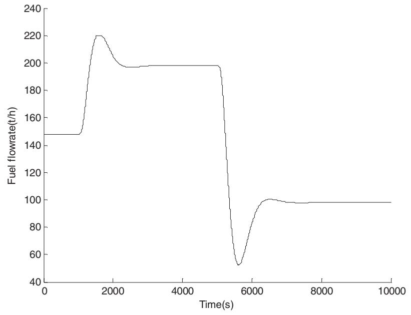


Figure 4(c) Fuel Flow disturbance

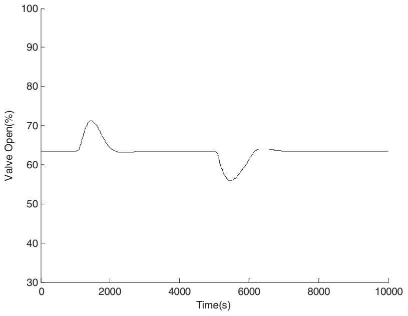


Figure 4(d) Valve Opening at fuel flow disturbance

## VI. CONCLUSION

In this paper, a well-established and validated nonlinear model of a 300MWe CFB power plant is introduced and analyzed to reveal its control difficulty. It is found that the nonlinearity firstly comes from static gain changing with load, then from the opening of steam turbine governing valve. After some model approximation, it is still hard to fully decouple the two-input-two-output system. Then a decentralized PID control is designed to control the unit load and the main steam pressure by defining a new control variable and transforming the nonlinear model into a linear formula. Simulation results show that the control system has good performance in load reference tracking for a wide load change in operation conditions, and also good disturbance rejection for the coal quality variation. The analysis and control in this paper is the first step to move forward for future control quality improvement and advanced control strategy study.

## REFERENCES

[1] Yao Qiang. Clean Coal Technology. Beijing, China: Chemical Industry Press, 2006.

[2] Huang Huanpao, Wu Liqiang, Han Jingqing, Gao Feng, Lin Yongjun, "A study of active disturbance rejection control on unit coordinated control system in thermal power plant", Proceedings of the CSEE, vol. 24, no. 10, 2004, pp. 168-173.

[3] Liu Xiang, Jiang Xuezhi, Li Donghai, Wan Jingfang, Xue Yali, "Coordinated auto disturbance rejection cotrol for boiler-turbine unit", Control Theory and Applications, vol. 1, suppl. 2001, pp. 149-152.

[4] Yu Daren, Xu Zhiqiang, "Nonlinear coordinated control of drum boiler power unit based on feedback linearization", IEEE Transactions on Energy Conversion, vol. 20, no. 1, 2005, pp. 204-210.

[5] T.Yu, K.W.Chan, J.P.Tong, B.Zhou and D.H.Li,"Coordinated robust nonlinear Boiler-Turbine-Generator control systems via approximate dynamic feedback linearization", Journal of Process Control, vol 20, no. 4, 2010, pp.365-374.

[6] Chen Pang Chia, S. Shamma Jeff, "Gain-scheduled L1-optimal control for boiler-turbine dynamics with actuator saturation", Journal of Process Control, vol.14, no. 3, 2004, pp. 263-277.

[7] Xue Yali, Li Donghai, Lu Chongde."Optimization of PID controllers of a boiler-turbine coordinated control system based on a Genetic Algorithm", Journal of Engineering for Thermal Energy & Power, vol. 21, no. 1, 2006 , pp. 80-83 ,87.

[8] Yu Daren, Xu Zhiqiang, Weng Yiwu, Li Yuehua. "A new understanding of DEB – gain scheduling control", Journal of Engineering for Thermal Energy & Power, Vol. 14, no. 83, 1999, pp. 379-396.

[9] Lu Yong, Liu Youkuan, Du Chaobo and Li Ping, "Optimization of the coordinated controller parameters for a nonlinear model based on a 300MW CFB boiler generating unit," Yunnan Electric Power, vol. 34, 2006, pp. 18-21.

[10] Han Zhongxu, Qi Xiaohong, Liu Min, Zhou Guang, "Mathematical model of controlled object in coordinated control system for 300MW boiler-turbine unit in Yaomeng Power Plant", Power System Technology, vol. 30, no. 1, 2006, pp. 47-50.

## Appendix I

TABLE II. NONLINEAR MAPPING FUNCTION PARAMETERS FOR A 300MWE CFB MODEL[9]

| Function | Parameter | Value | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $f_1(x)$ | $M$(t/h) | 0 | 40 | 60 | 80 | 100 | 120 | 140 | 160 | 180 | 200 | 220 |
| | $D_b$(t/h) | 0 | 127.3 | 202.9 | 295.8 | 405.5 | 496.1 | 583.2 | 664.3 | 763.6 | 858.4 | 915.1 |
| $f_2(x)$ | $P_b - P_T$(MPa) | 0 | 0.15 | 0.3 | 0.45 | 0.6 | 0.75 | 0.9 | 1.05 | 1.2 | 1.35 | 1.5 |
| | $D_t$(t/h) | 0 | 160.7 | 280.4 | 410.1 | 509.4 | 578.5 | 641.1 | 720.1 | 796.3 | 859.7 | 907.5 |
| $f_3(x)$ | $\mu_T$(%) | 0 | 30 | 60 | 82 | 90 | | | | | | |
| | $y$ | 0 | 0.2 | 0.9 | 1.12 | 1.2 | | | | | | |
| $f_4(x)$ | $P_T$(MPa) | 0 | 3.4 | 5.1 | 6.8 | 8.5 | 10.2 | 11.9 | 13.6 | 15.3 | 17 | |
| | $D_T$(t/h) | 0 | 285.3 | 340.0 | 394.2 | 446.9 | 495.2 | 544.6 | 610.8 | 705.13 | 811.6 | |
| $f_5(x)$ | $D_T$(t/h) | 0 | 50 | 150 | 250 | 350 | 450 | 550 | 650 | 750 | 850 | 1000 |
| | $P_1$(MPa) | 0 | 0.508 | 1.805 | 3.093 | 4.461 | 5.93 | 7.273 | 8.617 | 9.945 | 11.28 | 12.815 |
| $f_6(x)$ | $P_1$(MPa) | 0 | 2.0 | 3.0 | 4.5 | 6.0 | 7.0 | 8.5 | 10.0 | 11.5 | 13.0 | |
| | $N_E$(MW) | 0 | 38.1 | 66.3 | 107.9 | 151.3 | 179.8 | 218.9 | 261.9 | 300.4 | 328.4 | |

TABLE III. TIME CONSTANTS AND COEFFICENTS OF A 300MWE CFB MODEL[9]

| Parameter | $T_0$ | $T_1$, | $T_2$ | $C_b$ | $C_m$ |
|---|---|---|---|---|---|
| Value | 23.57 | 1.77 | 0.75 | 35726.1 | 14101.5 |

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# Developing virtual laboratories for introductory control

J.A. Rossiter and Y. B. Shokouhi
Automatic Control and Systems Eng., University of Sheffield, UK.
e-mail: j.a.rossiter@sheffield.ac.uk

*Abstract*—**This paper focuses on student access to learning opportunities and in particular those where students can learn by trial and error, such as in laboratories. It is recognised that regular student access to real equipment is a challenge for many institutions and thus alternatives are required, such as remote access laboratories. However, even remote laboratories are non-trivial to make available and thus this paper focuses on virtual laboratories. It demonstrates how these can be formulated very efficiently, can be highly accessible and critically, enhance the student learning experience. Several examples of virtual laboratories are discussed.**

**Keywords:** Web accessible laboratories, independent learning, authentic learning

## I. Introduction

There has been a sizeable body of work in recent years focussed on the laboratory experiences of students within engineering and a reassertion of the long standing view that laboratories form a key component of the student learning experience Abdulwahed (2010). This view is also strongly made by accreditation bodes Council (2011). Nevertheless, it is recognised that laboratories are expensive (Hofstein and Lunuetta, 2004) and indeed not necessarily efficient learning activities. Consequently Universities must seek a balance between the benefits of students interacting with equipment and the corresponding expense and inefficiency Lindsay and Good (2005); Ma and Nickerson (2006).

### *A. Remote or web accessible laboratories*

Many Universities have bought large scale into the concept of remote laboratories, e.g. (RELOAD, 2010; Qiao et al., 2010; LILA, 2010; Nagy and Agachi, 2004; Trevelyan, 2004). These enable departments to overcome many of the constraints associated to putting students into a laboratory room: typically there are restrictions on the number of duplicate equipment sets which means running the same activity numerous times in order to allow the entire cohort to participate and thus puts corresponding pressures on timetables. Consequently, most undergraduate students may only gain access to equipment about once a fortnight, with the exception perhaps of their final year research project.

Remote laboratories overcome barriers such as the timetable as the laboratory is then available 24/7. In principle, these laboratories can also be much cheaper as duplicate sets are not required and moreover, there is a not a requirement to find space for students to access the equipment (many universities now deploy space charges). With the right interface, especially with a suitable webcam, it is clear to students using a remote laboratory that it is real equipment and the data they are receiving is authentic.

Nevertheless, remote laboratories also have significant failings. Where the activity has a relatively slow timescale, there is still a need to allocate students specific access times. Even when the activity has a fast timescale, students may still need access for 5-10 min to complete their tests and this would be a severe irritation to students in a queue for access. With large class sizes, it is apparent that remote and web accessible need not imply there is good accessibility, which in turn could lead to student frustration.

### *B. Virtual Laboratories*

One alternative to remote laboratories is a virtual laboratory (Foss et al., 2006; Guzman et al., 2006; Khan and Vlacic, 2006), that is one which emulates real equipment and has the appearance of being authentic, despite being in fact just a simulation. Of course these have limitations (Engum et al., 2003; Magin and Kanapathipillai, 2000) because they are not the real thing, but nevertheless they can be highly authentic if done well (Goodwin, 2010). Moreover, they can form an invaluable component of an overall student activity set Abdulwahed (2010); Callaghan et al. (2008) because they provide activities which emulate much more closely than paper exercises the actual equipment. More specifically, virtual laboratories can form an invaluable preparation for access to real equipment as they can encourage students to think through the key concepts and tests that are required, and thus enable much more efficient use of equipment. Recent work in Southampton is also exploring how good quality video and animation could similarly improve student preparation and this has equally been denoted a virtual experiment (Memoli, 2011).

The main advantage of a virtual laboratory is that the access is much improved over remote laboratories; in principle all students can access simultaneously (unless there are license restrictions on the associated software). This means students have fewer obstacles to engagement and learning through trial and error in an pseudo-authentic scenario.

### *C. Summary*

This paper focuses on the role and development of virtual laboratories. The role is largely to support student learning and provide an accessible pseudo-authentic experience which helps students relate lecture content to real life scenarios, and

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

thus improve insight and understanding. However a secondary and equally valid role can be to facilitate preparation for a real experiment. The virtual laboratory can emulate activities and concepts required for the actual laboratory (Abdulwahed, 2010; Memoli, 2011) and thus enable students to prepare effectively.

The second contribution is to discuss the actual laboratories developed and give some evaluation from students on their views about these laboratories. The focus is on activities which support the learning of fundamental control engineering concepts.

## II. Supporting student learning through authentic activities

The priority for the author's department was to develop activities to support large cross faculty modules in modules related to control. Given the large size of the cohorts, access to equipment is very difficult in practice and thus remote activities were essential to give students access to more authentic scenarios. The topics of most interest within this paper are:

1) In year 1 students learn about modelling and system behaviours with most focus on 1st and 2nd order differential equation step responses. Some experimental activities were wanted to reinforce the concepts covered in lectures.
2) Also in year 1, students are introduced the concepts of feedback and PI control. There was a desire for activities that allow students to experiment with the PI parameters in both an emulated environment and on real equipment.

The developments follow the TRILAB concept to some extent.

- Students are introduced to the theory in lectures
- Students access a remote laboratory to test same ideas on real equipment and also to understand the differences.
- Students have virtual laboratories to practise in a pseudo-authentic environment.

Chronologically, the first two bullet points were developed first and thus this section will discuss the theoretical back ground and the remote equipment. The next section will focus on the third bullet point, which in fact will ultimately become the 2nd activity to help students prepare and thus is a key pedagogical element in the overall learning experience.

### A. *Creating a remote laboratory*

This paragraph is a summary of key points and is discussed in more detail in (J.A.Rossiter et al., 2011). The development of a web accessible laboratory is surprisingly easy and will be summarised in the following steps.

1) Connect up the hardware to the computer with a compatible I/O card. The authors found National Instruments cards easy to link into LabVIEW thus saving time.
2) Develop and test your LabVIEW virtual instrument (vi) or programmes' for communicating and controlling the experiment. The Front Panel window of the LabVIEW will be displayed to the user and it can be designed in a user friendly manner.
3) Once the experiment is working under a local computer it is a one click operation to use the web publishing tool to generate a web link for the vi and publish this link in your website. Students can then control the equipment via a web interface as if sitting next to the equipment.
4) There are some minor requirements on plug-ins for the browser to display correctly, e.g. Vision Development Module Run Time Engine and LabVIEW 2009 Run Time Engine.

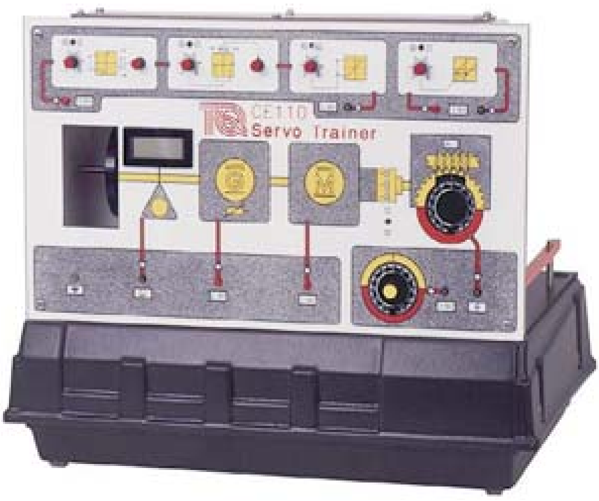


Figure 1. The DC servo equipment

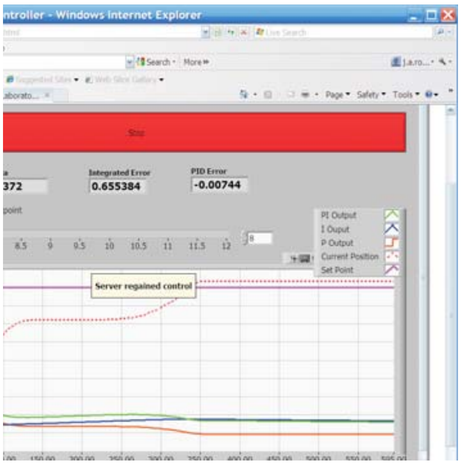
### B. *Activity 1 and equipment (J.A.Rossiter et al., 2011)*

The first activity is focussed on reinforcing student understanding of first order dynamics. In lectures students are taught to derive and analyse first order models and thus to understand the links between model parameters and behaviour as well as analogies between different systems. A laboratory can reinforce this by demonstrating:

1) Real systems do indeed have responses that are closely modelled by a first order response.
2) The system model parameters can be estimated reliably from measured data.
3) Real responses differ slightly from ideal behaviour.

The main parameters in a first order model are gain and time constant. Consequently the laboratory activity is split into three parts: (i) estimate the gain; (ii) estimate the time constant and (iii) compare the estimated model response with the actual behaviour.

The equipment selected is a simple DC Servo motor kit (see figure(1)) with analogue inputs and outputs. It consists of 5 different units in addition to nonlinearities. The students can see the axle rotation and the display showing angular velocity. Within the experiment the only input used is the input voltage as the response from voltage to angular velocity is approximately first order.



ntroller - Windows Internet Explorer

Integrated Error 0.655384

PID Error -0.00744

Figure 2. Interface for the stig equipment

The remote laboratory interface has three separate tabs. In tab one students modify the input voltage directly to estimate the steady state gain, positive dead-zone and negative dead-zone of the DC servo motor. Gain is given as the gradient of the input/output curve. In tab two, the time constant of the system is estimated using responses to a square wave input. In tab three, students enter their estimated gain and time constant and produce an exact step response to compare with the plots from tab two.

### C. Activity 2 and equipment

The second activity (internally denoted as 'the stig') is designed to reinforce student understanding of PI control and basic control concepts. The underlying objectives are for students to investigate:

1) The impact of changing gain with no integral.
2) The effect of changing integral with no proportional.
3) The potential of using proportional and integral together.

The equipment consists a cart on 2 metre long rails. The cart is moved by a motor. The underlying dynamics are such that in the future the same equipment will be suitable for experiments looking at 2nd order modelling and dynamics (e.g. under damped responses) as well as an introduction to feedback.

The remote interface is very simple in form. Students are able to choose a proportional term, an integral term and a set point. The system begins moving and the final interface is a button which allows them to stop the experiment at a point of their choosing. The graphical display (figure 2) shows the output position, the input signal, the output of the integral term and the output of the proportional term for the entire runtime.

The interesting point to note here is that there is clearly some stiction in the system so although, with patience, the system will progress close to the desired steady-state, it never quite gets there because the change in input needs to be large enough to get the system moving again and then it tends to jump. However, stiction aside, the expected behaviour does ensue so the experiment gives students a good insight into reality and the relevance of the theory covered in lectures. Moreover, it allows them to see the role of the integral term (blue line) and proportional term (orange line) in the overall input signal (green line); this is particular important for understanding the real system effects such as stiction because students can see the cart stalls even when the input (and integral term) is changing.

### D. Laboratory design: pedagogy and learning outcomes

Access to the equipment itself is not, in general, enough to support student learning. The authors have experienced significant frustrations due to software and hardware crashes which limit student access until the crash is noted and rectified manually. Although some crashes can lead to an automatic reboot J.A.Rossiter et al. (2011), this is not the case for all.

A second weakness of remote laboratories is that only one student can access these at a time. With small cohorts this may not be an issue, but with cohort sizes of 100 plus, the likelihood is that students will have similar free periods and will all try to access simultaneously. Consequently they may have to wait a substantial period to come first in the queue, and their position in the queue will not be obvious without substantial increase in complexity of coding at the server end. This weakness will limit their ability to learn by a large number of trial and error experiments, something staff may wish to encourage.

Consequently, the next section looks at how students can spend time focussing on learning concepts and thus require less time on the equipment to validate the authenticity of their learning.

## III. Virtual Laboratories to support preparation for remote laboratories

The key aim of the virtual laboratories is to provide maximum accessibility for students to practise. One could argue therefore that web interfaces such as (Khan and Vlacic, 2006; Guzman et al., 2006) are ideal. However, the downside of such laboratories is the skill and time required to develop them, as well as the need for a maintenance of an appropriate server. Consequently, the authors decided to follow a route which minimised the staff skill and time requirement, as this is pragmatic and increases the potential for more staff to participate.

In summary, it was decided to use MATLAB/SIMULINK as the base for virtual laboratories.

1) The relevant files can be distributed easily for students to use anytime and anywhere.
2) MATLAB is available on the University network and thus students have excellent access to the software. Many students also purchase a student version for home use or can get remote access to the University software with the relevant 'access code'.
3) The software is well understood by most staff and widely used, thus making resources easier to produce and share.

4) The author's department also has a well established server system which students can access remotely to run MATLAB files.
5) The GUIDE tool allows for relatively straightforward production of GUIs which make interaction easy and intuitive for students.

The virtual laboratories are designed as far as possible to emulate the physical laboratories so students go through the same steps and focus on the same concepts. This will allow them to become familiar with the key observations expected before accessing the equipment and moreover they are more likely to notice the key differences between the theory and the practice.

In terms of staff effort, each GUI took about half a day to create which is much quicker than requirements with alternative software choices.

### A. Virtual Modelling laboratory

This laboratory was focussed on 1st order modelling. To allow some non-linearity and add realism, the simulation is based on a simulink model which has simple first order dynamics but with some dead zones and measurement noise also added. The virtual laboratory was produced as single GUI (figure 3) which embodied all 3 activities of the experimental equivalent and hence it required:

- Three axes, one for each activity.
- Buttons for changing the input voltage and saving data.
- Boxes to enter the estimates for gain and time constant.

The top right axis shows the steady-state vs the input, marked by crosses, for different inputs and also a best estimate of the slope; clearly the slope is an estimate of gain. The input voltage is selected by a slider, a button requests a simulation with this value and another button confirms the data should be entered into the plot (so a student need not save all values). The little circle in the top middle emulates the spinning of the servo and rotates in real time on the GUI, as well as showing the steady-state speed. The reader will note some stiction is included in the simulation so there is no movement for small voltage inputs. For completeness, figure 4 shows the equivalent interface on the actual equipment. This has separate axis for positive and negative input voltages but otherwise is seen to have equivalent functionality: the webcam is used for students to read the speed, there is box to enter this reading and another button to *add data* to the axes - this data is also displayed in numeric form. Students are also encouraged to identify the dead zone and enter the observation into the relevant boxes.

The axis in the bottom left shows the responses to a square wave, with the same input amplitude as for the first figure. Some noise is added to encourage students to think about real issues. This display can be used to estimate the time constant.

The axis in the bottom right is used to simulate a model based on the gain and time constant estimates. Students must enter their estimates into the boxes provided.

The main objective of the GUI was to allow students to go through the same conceptual steps required for the remote laboratory. That is practise changing the input voltage, reading the steady-state output and then adding this data to the axis (hence the need to a deliberate button press). The remote laboratory will produce a similar figure containing the crosses, although in that case students need to estimate the slope themselves and then enter into a box. The bottom axis gives a very similar plot to that students would see on the real equipment, with a square wave response. From this, in both cases, the activity required is an estimate of time constant. Finally, the third activity in both cases is to use the gain and phase estimates to from an ideal 1st order model and simulate the step response. The intention is that students would therefore find engagement with the remote laboratory straightforward as well as being clear on the key learning outcomes, understanding first order responses.

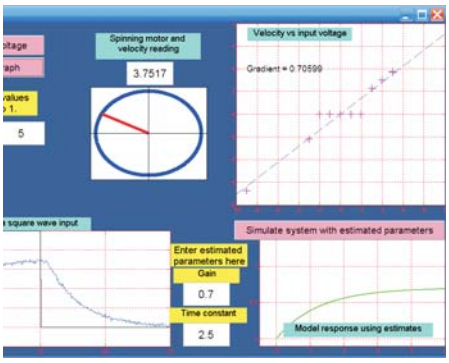
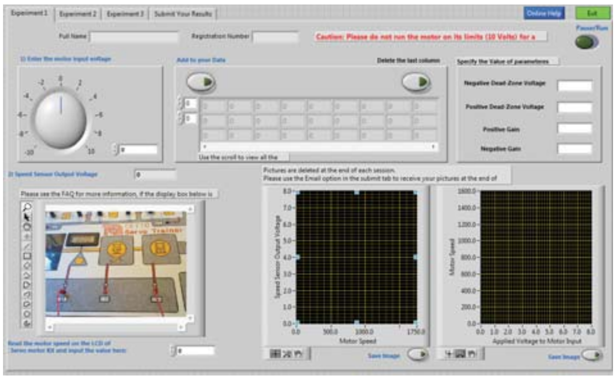

oltage
aph
values
1.
5
Spinning motor and velocity reading
3.7517
Velocity vs input voltage
Gradient = 0.70599
square wave input
Simulate system with estimated parameters
Enter estimated parameters here
Gain
0.7
Time constant
2.5
Model response using estimates

Figure 3. MATLAB GUI for 1st order modelling.



Caution: Please do not run the motor on its limits (10 Volts) for a

Use the scroll to view all the

Please see the FAQ for more information, if the display box below is

Pictures are deleted at the end of each session.
Please use the Email option in the submit tab to receive your pictures at the end of

Read the motor speed on the LCD of Servo motor Kit and input the value here

Figure 4. Interface for activity 1 on remote laboratory.

### B. Virtual PI laboratory

This laboratory is based on position control of a cart (carrying a passenger) along a track. The requirement was for students to explore the impact of changing the PI parameters on performance. Hence the chosen GUI was chosen to be very simple in form as shown in Figure 5. There are simple sliders for entering the choice of proportional and integral terms. The top axis shows the target and output position curves (in figure 5 there is a steady-state offset as the integral is zero). The bottom axis is an animation and students see the passenger

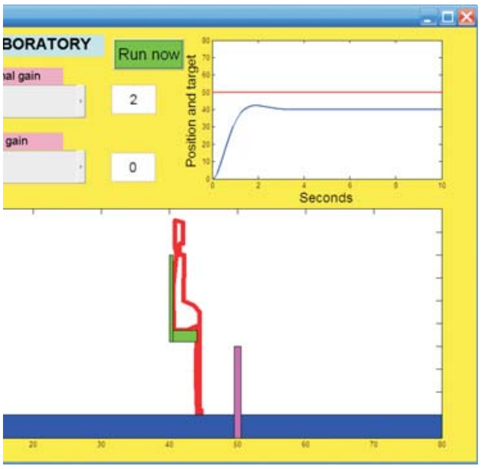


Figure 5. MATLAB GUI for PI control of position.

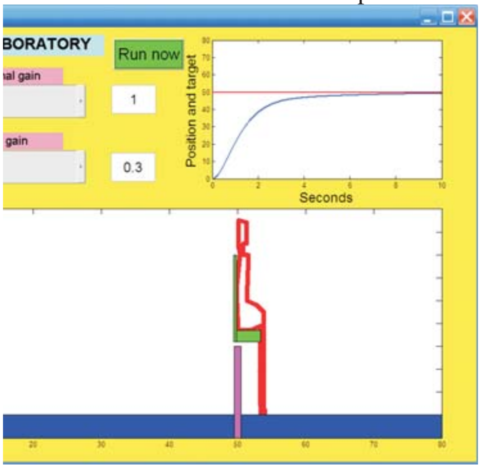


Figure 6. MATLAB GUI for PI control of position.

moving but ultimately stopping short of the target. Figure 6 shows a simulation with a non-zero integral where clearly the offset is removed.

For completeness, figure 2 shows the associated hardware laboratory interface, accessible via the web. Again, students need only set the PI parameters and while the same basic observations will follow, it is clear that the behaviour is far from ideal; this should give students something challenging to ponder.

## IV. Student evaluation

It is known that remote laboratories can be beneficial and hence the main focus of the evaluation here is on the efficacy of the virtual laboratories for enhancing the overall student learning experience. Students were asked a number of questions and the responses are summarised in Tables 1 and 2. Table 1 is a smaller student group who had the virtual laboratories paired with a hardware laboratory. Table 2 is a much larger group (over 200) who had the virtual laboratories solely for supporting learning, but not for assessment.

It is interesting to note that the second group were less positive overall, but this is probably a reflection of their engagement being formative rather than summative and hence many of this group will not have used the virtual laboratories effectively, if at all (many students only put in effort if 'it counts'). Indeed a question on an issue not related to this paper showed that only about 50% of the class had engaged with a key formative resource.

TABLE I
Student evaluation of virtual laboratories (ACS108).

| Question | Strongly agree or agree | Neither agree or disagree | Disagree |
|---|---|---|---|
| Virtual Laboratories (MATLAB GUIs) were easy to use and access | 89% | 7% | 4% |
| Virtual Laboratories helped me prepare for the remote laboratories | 75% | 21% | 4% |
| I felt more confident using the remote laboratories having first gone through the virtual laboratories | 50% | 38% | 12% |
| It was useful to see the differences between a simulation (ideal model of virtual lab) and the responses on real equipment. | 89% | 7% | 4% |
| The modelling virtual laboratory helped me understand the key parameters of gain and time constant | 75% | 18% | 7% |
| The STIG virtual laboratory helped me understand the role and impact of the key parameters of P and I | 82% | 14% | 4% |
| I think the department should produce more virtual laboratories to support preparation activities for laboratories. | 92% | 4% | 4% |
| I think the department should produce more virtual laboratories to support learning of key concepts. | 96% | 4% | 0% |

Where engagement was summative, that is for the ACS108 students, it is clear that the resources were useful for the majority. The relatively poor response on preparation for the actual remote laboratories is more likely a reflection of the poor reliability of the remote laboratories so that accessibility was poor and thus many students failed to do the hardware laboratory; this latter issue is an ongoing priority for technical staff.

## V. Conclusions

This paper has looked at the provision of laboratory activities within engineering curricula and proposed that the role of the virtual (and remote) laboratory has much more potential than is being exploited in most institutions. Virtual laboratories have the advantage of being accessible 24/7 and also allow parallel access by a large number of students, sometimes the whole cohort.

This paper has illustrated two simple uses of virtual laboratories. The most basic use is as a formative learning exercise, to allow students to practise with key concepts and thus to improve their understanding. A second and more integrated use links the virtual laboratories with real equipment and

TABLE II
STUDENT EVALUATION OF VIRTUAL LABORATORIES (ACS124).

| Question | Strongly agree or agree | Neither agree or disagree | Disagree |
|---|---|---|---|
| The virtual laboratories helped me understand the role of modelling and simulation in design. | 84% | 7% | 9% |
| Virtual Laboratories helped me prepare for the remote laboratories | 75% | 21% | 4% |
| The virtual laboratories helped me understand the role and impact of the feedback parameters P and I. | 49% | 35% | 16% |
| I think the department should produce more virtual laboratories to support the learning of key concepts. | 72% | 23% | 5% |

summative assessment. Virtual laboratories can be used to emulate the activities, concepts and questions students will face in an actual laboratory and thus provide a tool for preparation so they get more out of valuable time on the equipment. The combination of real and virtual laboratories also draws students' attention to the differences between theory and practice. Student evaluation has reinforced the efficacy of the approach.

The final contribution of the paper is to discuss practical issues of developing virtual laboratories. This paper has proposed the use of MATLAB/SIMULINK GUIs. The creation of GUIs with very similar interfaces and inputs to the actual hardware is a relatively straightforward coding exercise using the GUIDE tool, especially as most systems and control engineering staff have some proficiency with MATLAB. This has the advantage that virtual laboratories can be created relatively quickly. A second advantage of this proposal is that many Universities provide site licenses and thus student access is straightforward.

## REFERENCES

Abdulwahed, M. (2010). *Ph.D Thesis, Towards enhancing laboratory education by the development and evaluation of the trilab concept*. Univ. Loughborough.

Callaghan, M., Jim, H., Martin, M., and Maguire, L. (2008). Intelligent user support in autonomous remote experimentation environments. *IEEE Trans. on Industrial Electronics*, 55(6), 2355–2367.

Council, E. (2011). Uk-spec. *http://www.engc.org.uk/professional-qualifications/standards/uk-spec*, 55(6), 2355–2367.

Engum, S., Jeffries, P., and Fisher, L. (2003). Intravenous catheter training system: Computer-based education versus traditional learning methods. *The American Journal of Surgery,*, 186(1), 67–74.

Foss, B., Solbjrg, O., Eikaas, T., and Jakobsen, F. (2006). Game playing in vocational training and engineering education. *Proc. ACE*.

Goodwin, G. (2010). Virtual laboratories for control systems design. *http://www.virtual-laboratories.com/ (last checked 1/9/10)*.

Guzman, J., Astrom, K., Dormido, S., Hagglund, T., and Y., P. (2006). Interactive learning modules for pid control. *Proc. ACE*.

Hofstein, A. and Lunuetta, V. (2004). The laboratory in science education: Foundations for the twenty-first century. *Laboratory of Science Education*, 88(1), 28–54.

J.A.Rossiter, Baradaranshokouhi, Y., Lilley, I., and Bacon, C. (2011). Developing web accessible laboratories for introductory systems and control using student projects. *IFAC world congress*.

Khan, A. and Vlacic, L. (2006). Teaching control: benefits of animated tutorials from viewpoint of control students. *Proc. ACE*.

LILA (2010). Library of labs. *http://www.lila-project.org/ (last checked 1/9/10)*.

Lindsay, E. and Good, M. (2005). Effects of of laboratory access modes upon learning outcomes. *IEEE Trans. on Education*, 48(4), 619–631.

Ma, J. and Nickerson, J. (2006). Hands-on, simulated, and remote laboratories: A comparative literature review. *ACM Computer Survey*, 38(3), 1–24.

Magin, D. and Kanapathipillai, S. (2000). Engineering students understanding of the role of experimentation. *European Jour. of Eng. Education*, 25(4), 351–358.

Memoli, P. (2011). Virtual experiments, http://www.edshare.soton.ac.uk/6589/1/preloader-diode.html. *Project funded by HESTEM*.

Nagy, Z. and Agachi, S. (2004). Internet-based interactive remote laboratory for educational experiments. *AIChE Annual Meeting*.

Qiao, Y., Liu, G., Zheng, G., and Luo, C. (2010). Design and realization of networked control experiments in a web-based laboratory. *Proc. UKACC*.

RELOAD (2010). Real labs operated at a distance. *http://www.engsc.ac.uk/mini-projects/reload-real-labs-operated-at-distance (last checked 1/9/10)*.

Trevelyan, J. (2004). Lessons learned from 10 years experience with remote labs. *Int. Conf. Eng. Educ.*

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Using MATLAB GUIs to improve the learning of frequency response methods

R.J.Mitchell
School of Systems Engineering
University of Reading
Reading, UK
r.j.mitchell@reading.ac.uk

*Abstract*—**This paper describes two MATLAB GUIs that have been designed to improve student learning of frequency response methods. One GUI is focused on plotting using asymptotes and one on systems identification. The paper includes positive student feedback on how they felt these GUIs helped their understanding, and useful suggestions on how they can be improved.**

***Keywords- Education, Frequency Response, Bode diagrams***

## I. INTRODUCTION

A recent survey of Control curricula[1] in the UK reaffirmed that Frequency Response methods were an integral part of the degree, typically taught as part of the second course in control. In the past, as shown by books such as Atkinson[2], much emphasis was given to the calculations for plotting and designing using Bode and Nyquist diagrams. Nowadays, programs such as MATLAB can do much of the work for students, but there is the danger that students do not fully understand what is happening. In [1] specific views are given:

"Today however the scenario is different. Students need to have an understanding of how to sketch Bode and Nyquist plots and root-loci, but need to have less concern for computing the numerical details; software can be used."

"I think there is great value in teaching the theory of Bode diagrams and Nyquist plots. This provides the basis of understanding the plant dynamics and the effects of closed loop control and the controller settings. However, although I think students should understand the mechanics of calculating the frequency response of a system, the use of MATLAB should enable the calculations to be done quickly so that time can be devoted to controller design. It would also allow higher order systems to be investigated."

"Computer assessment using MATLAB looks interesting … though still have a hankering after asymptotic Bode sketches as a first start - confirm with MATLAB once or twice then leave it to computer."

This year, after a teaching review, a System Identification and Control module was formed, for which the author gave ten lectures and two assignments on frequency response methods. These covered plotting Bode and Nyquist diagrams, and the design of controllers and identification of linear systems using frequency domain data.

This paper describes the assignments set, to students taking Cybernetics, Robotics and Electronic Engineering degrees at Reading, including the use of two MATLAB GUIs developed to assist students to learn about the plotting and identification, and the associated programming of controllers.

These GUIs involve the use of asymptotes, as they help the understanding of Bode diagrams, for which the author has developed extra asymptotes to help as regards phase plots. These extra asymptotes are also described.

The paper is organized as follows. In the next section, some comments are made as regards frequency response, asymptotes and identification using Bode plots. Section III describes the GUI used for plotting asymptotes and Section IV the GUI for identification. Section V contains details of the assignments, and in section VI student feedback on the GUIs is given. Section VII has the author's reflections on the feedback, and then concluding remarks are made.

## II. ON FREQUENCY RESPONSE

The frequency response of a system is the variation of its gain and phase as the angular frequency ω of its sinusoidal input is varied. The results can be plotted on Bode diagrams, where log(gain) and phase are plotted separately against log(ω), or on a Nyquist diagram, where the gain and phase provide the polar coordinates of the locus. In this paper, Bode plots are considered, which are used to plot the 'open loop' transfer function (that round the feedback loop), to assess stability, design controllers and for identification.

Here, linear systems are modeled as a series of single or quadratic poles or zeros, operating around corner frequencies. Knowing these corner frequencies, it is possible to calculate values so as to sketch the Bode plots. Similarly, the corner frequencies can be estimated from the Bode plots of a system. In both processes one considers what happens at low frequencies and then at each corner frequency in turn, until the highest corner frequency has been processed. It is thus appropriate to develop computer based systems for both the plotting of Bode plots and system identification from Bode plots.

In Thorne[3], laboratory practicals are described in which students learn to identity systems comprising multiple quadratic poles only. Thorne also believes that identification helps students in the understanding of Bode plots.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

Bode plots can be approximated by straight lines, the asymptotes, which can be derived from Bode's fundamental work[4]. For a stable minimum phase system with transfer function P(jω), its phase at $\omega_c$ is given by (1): $u = \ln(\omega/\omega_c)$

$$\angle P(j\omega_c) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{d|P(j\omega)|}{du} \ln\coth\frac{|u|}{2}\, du \qquad (1)$$

By assuming that the gain is constant around $\omega_c$, the derivative tern can be brought outside the integral. The remaining integral evaluates as ½ $\pi^2$. Changing the derivative variable from u to ω, (1) can be rewritten as:

$$\angle P(j\omega_c) \approx \omega \frac{d|P(j\omega)|}{d\omega}\Big|_{\omega_c} * \frac{\pi}{2} \qquad (2)$$

Hence the Bode gain plot can be approximated as straight lines, between the corner frequencies of P(jω), whose gradients are integers, and the phase plot can be approximated as constant integer multiples of ½ π radians (or $90^O$).

Sketches of gain plots start and end on the first and last asymptotes, and can be close to all the asymptotes, depending on the proximity of the system corner frequencies. This is less true of the phase, due to step changes of the asymptotes at the corner frequencies. As such, books like Dorf and Bishop[5] or Nise[6] suggest that these phase asymptotes should be joined by 'diagonal' asymptotes between one tenth of and ten times the associated corner frequency, that is over a range of 100 rad/s. The author has developed a better approach, where the range for such diagonal asymptotes is set so that their slope equals that of the phase plot at the corner frequency. For a single pole or zero, this range is $e^{\pi} \approx 23$rad/s. For a quadratic pole or zero with damping ratio ζ, the range is $e^{\zeta\pi} = 4.8$ rad/s if ζ = 0.5 or 2.2 rad/s if ζ = 0.25. These values and the approach are justified in the appendix, and used in the GUIs.

Consider the following system which illustrates these extra asymptotes, and the approach for plotting and identification. The system being analysed is

$$\frac{50(1 + s/2)}{(1 + s/0.07)(1 + s/0.4)(s^2/20^2 + 0.6s/20 + 1)} \qquad (3)$$

By inspection the corner frequencies are 0.07, 0.4, 2 and 20 rad/s. The gain asymptote has value 50 and slope 0 until 0.07 rad/s, the slope is -1 until 0.4 rad/s, -2 until 2 rad/s, -1 until 20 rad/s and -3 thereafter. In these ranges, the phase asymptotes have values 0, -90, -180, -90 and -270 degrees. The range of the extra phase asymptotes around each corner frequency is 23 rad/s except at 20 rad/s where it is 2.6 rad/s. These can be seen in Fig 1, where the actual Bode plot is also superimposed. The gain plot is easy to sketch from the asymptotes. The extra phase asymptotes make it easier to sketch the phase plot.

As regards identification, the shape of the phase plot here could be used. As the phase moves from $0^O$ past $-90^O$ it is apparent that there are two low frequency poles. The phase then rises, suggesting a zero, and then the phase decreases more rapidly towards $-270^O$ suggesting a quadratic pole: the steepness of the plot at this point, which is consistent with the diagonal asymptote, suggests that the pole is underdamped.

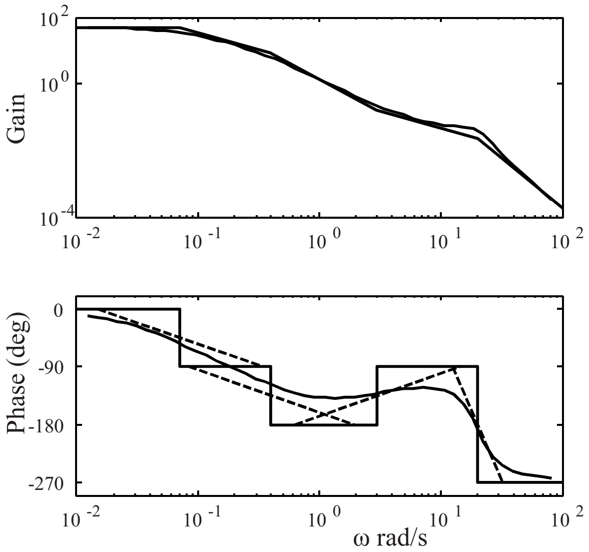


Figure 1 Bode plot of example system including asymptotes

In this example, the corner frequencies are well spaced, making sketching and identification easier as each pole or zero has a relatively small impact on the others. In the GUI described later, it is shown how closer poles and zeroes are addressed, for instance, by identifying two such poles by a quadratic pole, or a pole near a zero by a lead-lag element. First, however, the GUI for plotting is described.

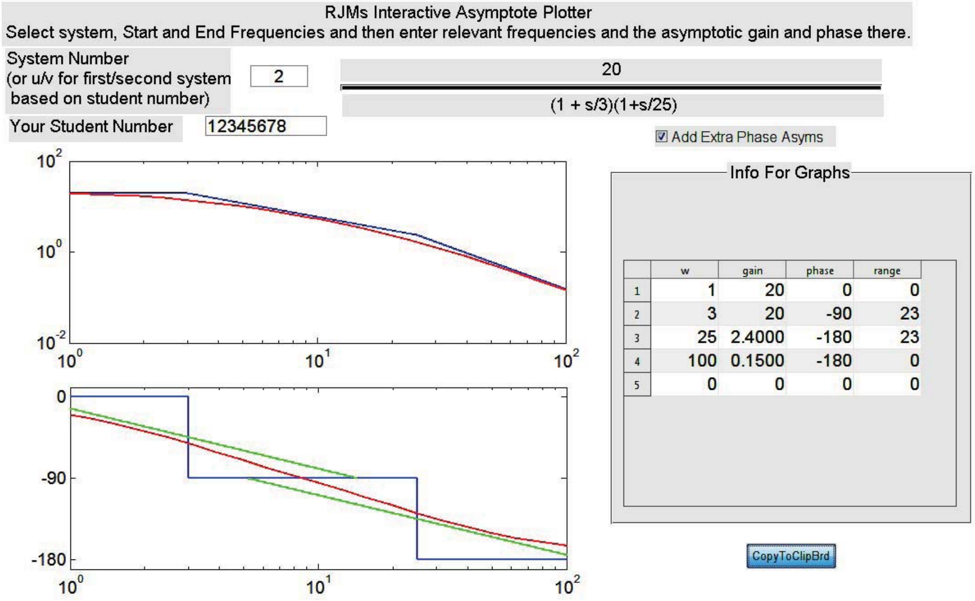
## III. GUI FOR PLOTTING

The first GUI is aimed at helping students to understand the frequency response, by drawing the asymptotes for the Bode gain and phase plots. In the lectures various transfer functions are discussed and the corner frequencies identified. It is stated that the slope of the gain asymptotes are integers and that the phase asymptotes constant values being that slope multiplied by $90^O$. It is explained that, for example, when the corner frequency is that of a single pole, the slope decreases by 1, and hence there is thereafter an extra $90^O$ phase lag. In the past, the emphasis has been on sketching gain asymptotes approximately using these integer slopes. However, in the author's experience, even some of the better students did not appreciate what was happening: this is consistent with[1].

The GUI was therefore designed to require the students to inspect a transfer function and enter each corner frequency in turn and relevant information about the gain and phase.

Initially it was planned that the students would enter the slope of the gain plot and the asymptotic phase immediately after each corner frequency. However, as phase is the gain slope times $90^O$ it was perhaps redundant to require both. The author thought that students would learn more if they had to enter at the corner frequency the asymptotic gain there as well as the asymptotic phase. Thus if the gain was G at corner frequency $\omega_1$, after which the slope was s, then at the next corner frequency $\omega_2$ they would have to calculate the gain as

$$G * \left(\frac{\omega_2}{\omega_1}\right)^s \qquad (4)$$



RJMs Interactive Asymptote Plotter

Select system, Start and End Frequencies and then enter relevant frequencies and the asymptotic gain and phase there.

System Number (or u/v for first/second system based on student number) 2

$$\frac{20}{(1 + s/3)(1+s/25)}$$

Your Student Number 12345678

☑ Add Extra Phase Asyms

Info For Graphs

| | w | gain | phase | range |
|---|---|---|---|---|
| 1 | 1 | 20 | 0 | 0 |
| 2 | 3 | 20 | -90 | 23 |
| 3 | 25 | 2.4000 | -180 | 23 |
| 4 | 100 | 0.1500 | -180 | 0 |
| 5 | 0 | 0 | 0 | 0 |

Figure 2 Plotter GUI after student has used system

Overall, when given a transfer function, the student first has to enter a suitable range of frequencies for the plot, being before the first and after the last corner frequencies. The student then enters for this first frequency, each corner frequency in turn, and the last frequency, the gain of the asymptote there, and the asymptotic phase after that frequency. If the extra diagonal asymptotes are used, the student has to enter the range of the phase asymptote around that frequency.

As each frequency is added, the asymptotes up to and including this frequency are shown. The student can move to the next corner frequency, or 'undo' the last entry.

At the end, the actual Bode plots are superimposed on top of the asymptotes, and the student can assess how well they have entered the data. As an example, suppose the system is

$$\frac{20}{(1 + s/3)(1 + s/25)} \tag{5}$$

The corner frequencies are 3 and 25 rad/s, and so a suitable range of frequencies is 1 to 100 rad/s. The student should then enter the following (the last column being needed for the extra phase asymptotes)

| ω | Gain | Phase | Range |
|---|---|---|---|
| 1 | 20 | 0 | 0 |
| 3 | 20 | -90 | 23 |
| 25 | 2.4 | -180 | 23 |
| 100 | 0.15 | -180 | 0 |

Figure 2 shows the GUI after these data have been input.

The GUI has a variety of systems built into it which the student can be tested upon. In addition, as the GUI is used as part of marked coursework, potential plagiarism needs to be addressed. Thus the GUI can also be invoked with systems whose structure is fixed, but whose corner frequencies are calculated from the student's unique student number.

## IV. GUI FOR IDENTIFICATION

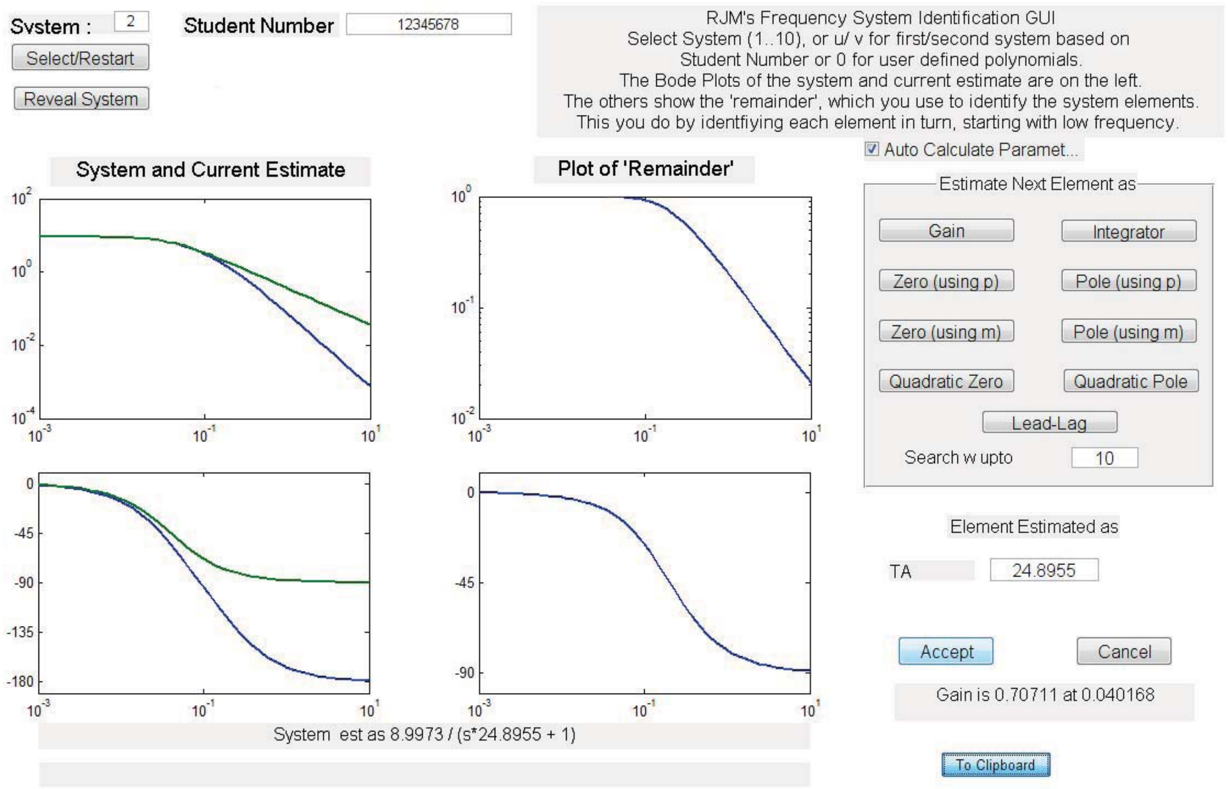
For identification, the process is straightforward, with the result being built up as a series of elements starting at low frequencies, and continuing until after the last corner frequency is found. The user selects that the next element is a constant gain, an integrator, a single pole or zero, a quadratic pole or zero, or a lead-lag element.

For any system to be identified, three arrays are provided, having a series of angular frequencies and the gain and the phase at each of these frequencies. In the identification process, two more gain and phase arrays are calculated: one set has the gain and phase of the identified model, and one has the gain and phase of the model that is still to be identified. The former set is initialised with gains of 1 and phases of 0; the latter set is initialised with the original gain and phase arrays provided.

Once an element of the system model is estimated, the gain and phase of that element are calculated at each of the angular frequencies. The 'already identified' gain array is multiplied by the gain found, and its phase array has the new phase values added. The 'to-be-identified' gain array is divided by the gain found and its phase array has the new phase values subtracted.

Two Bode plots are provided. One shows the frequency response data provided, superimposed on to which is the estimate of the system so far. The other has the response yet to be identified. The user interacting with the system should see the identified system following the original plot at lower frequencies. Deviations of the to-be-identified plot from its low frequency response are used to identify the next element.

As used this year, in the GUI the user specifies the type of element, and the system automatically estimates the parameter(s) associated with that element, such as gains or time constants, which the user can accept, modify or reject. In the final section of the paper consideration is given as to how the user could be required to calculate these parameters.



Svstem : 2

Student Number 12345678

RJM's Frequency System Identification GUI
Select System (1..10), or u/ v for first/second system based on
Student Number or 0 for user defined polynomials.
The Bode Plots of the system and current estimate are on the left.
The others show the 'remainder', which you use to identify the system elements.
This you do by identfiying each element in turn, starting with low frequency.

☑ Auto Calculate Paramet...

Search w upto 10

Element Estimated as

TA 24.8955

Gain is 0.70711 at 0.040168

System est as 8.9973 / (s*24.8955 + 1)

Figure 3 Part way through identification of two pole system

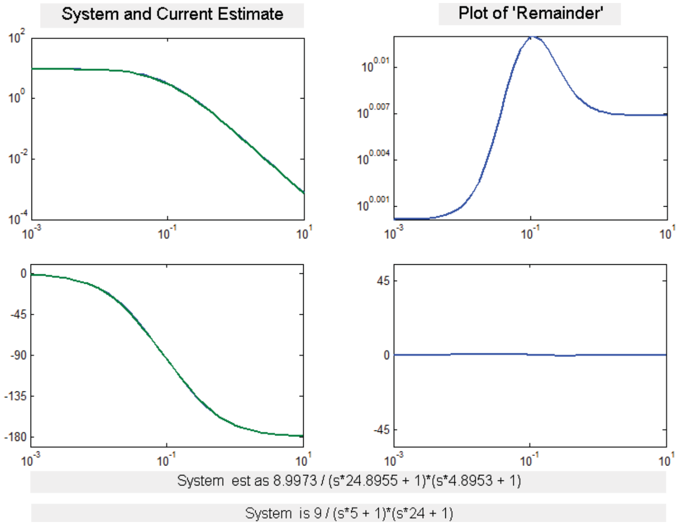
When the user selects the type of the next element, the plots are updated and the user is shown the value of any estimated parameter. The user can then modify that value, perhaps on the basis that a subsequent element is affecting this one. The user can then accept the element and value, or cancel, which could happen if the wrong type of element had been selected.

The system estimates the gain of an element from the first value in the gain array. For a pole or zero, the user specifies whether its time constant is found using the gain or phase array. If the former, the time constant is the reciprocal of the angular frequency where the gain is a factor 0.707 away from the low frequency gain; if the latter the frequency where the phase is $\pm 45^{O}$ is used. For quadratic poles or zeros, the frequency where the phase is $\pm 90^{O}$ gives the estimate of $\omega_n$ and this value and the gain there give an estimate of $\zeta$. If $\zeta > 1$, the result is shown as two time constants. This allows the user to identify two close poles or two close zeros as one element. When a pole and a zero are close, a lead-lag element can be identified: the frequency where the phase is a local minimum or maximum, and the value of that phase are used to find the time constants.

Fig 3 shows the GUI part way through the identification of a system with two poles. Here the gain was first estimated, as 8.9973, the actual value is 9. Then the time constant of the first pole was found, estimated at 24.895: it should have been 25. The graphs on the left show that this model accurately follows the actual plot at low frequencies. The graphs on the right indicate that a further pole needs to be identified.

At the end, the student presses the reveal button and the actual transfer function is given. The student can then see how well the process has worked by comparing the actual and estimated transfer functions and Bode plots.

Fig 4 shows part of the GUI at the end of identifying the two pole system. The second pole should have had a time constant of 5, but the estimate was 4.65. The remainder plots are not quite horizontal lines showing that there have been small (but expected) errors in the estimates of the parameters.



System est as 8.9973 / (s*24.8955 + 1)*(s*4.8953 + 1)

System is 9 / (s*5 + 1)*(s*24 + 1)

Figure 4 At the end of identification of two pole system

## V. Assignments

These GUIs were used in two assignments. The first was aimed at familiarizing the student with frequency response plots. It comprised tasks set after lectures 2, 3 and 4, including the use of the GUI to plot various systems, and to design a simple controller and assess its performance. Formative feedback only was provided for this part.

The first task was to call the m file provided which generated a system based on the student's username: the numerator and denominator polynomials were displayed symbolically, as shown at the top of Fig 2. Students then noted the corner frequencies of that system, and used MATLAB to calculate the gain and phase at those frequencies. In the second task, students used the GUI to plot a few different systems, including one based on their username. Finally, the students used MATLAB to design a proportional controller to achieve a phase margin specification, plot the step response and calculate key values such as overshoot and time to peak.

Students were provided with a word document into which results were to be pasted, with questions asking the student to reflect on their answers, provide the lecturer with feedback on the GUI and say whether they felt it useful to teach asymptotes. Built into this document were options for formative assessment of the work – such as the section was missing, wrong, partially correct or correct. The marker could quickly circle the appropriate assessment and, with more detailed comments, provided rapid and relevant formative feedback on this part of the assignment, which are known to be beneficial [7], [8].

The second assignment included more plotting, of systems with second order poles, the use of the identification GUI, and the design of Phase Lead, P+I and PID controllers using frequency domain data. For the design, students were provided with an m file with functions provided for defining systems, for plotting graphs, and empty functions for controller design which the students had to complete. Again results, and code, were pasted into a word document, but this now had a marking scheme instead of comments like 'wrong'. Students were also asked to comment on the GUI.

## VI. Evaluation

The 20 students on the module generally did the first assignment well: the two who did not cited difficulties understanding the GUI. Most answered the questions asked.

The first question asked was: 'Has the GUI helped/hindered you to understand asymptotes/Bode plots". Almost all students said it had. Some specific quotes are given below:

"At first I felt the GUI was hindering me as I could not work out how to get the correct asymptotes, however the GUI started to encourage me to research and learn the theory … Looking back on the learning process I feel it was very useful for understanding the concepts of asymptotes and Bode plots".

"The GUI has vastly improved my understanding of Bode plots, mainly due to it being a focused exercise towards improving sketching Bodes that includes being able to check answers against the actual plots of the system".

The second question asked if it would have been better if they had been asked to specify the gain slopes rather than the value of the asymptotic gain at the corner frequencies. Many students said that would have been easier, but they would not have learnt as much. Some specific quotes are given below.

"I have previously used the method 'slope specification' when plotting by hand so it is useful to have another method."

"I feel if I had learnt both and tried both on a similar style GUI I would have learnt far more. It would have been useful to practice both methods so that we could work out which method suited us individually."

"I don't think that I would have learnt so much if I had to only enter the slopes"

The third question was whether it was useful to teach asymptotes . All who expressed a preference believed it was.

"It is very useful teaching about asymptotes because the process of plotting them enables one to understand how the various terms of a transfer function relate to the system's gain, phase and corner frequencies. Without the manual plotting of asymptotes I don't believe that my understanding of Bode plots … would be nearly as in depth"

"[Definitely] as it offers a basis for Bode plots without having to do thousands of calculations. … It helped me a lot in understanding Bode plots."

Students were also asked to make suggestions for improving the GUI. On-line help was requested, and there were specific comments on the entry of data at each corner frequency in the plotting GUI. These will be used next year.

For the second GUI students were asked if it helped/ hindered understanding. Again almost all students felt it did.

"I found the GUI extremely helpful as I was able to see what is happening with the frequency response of the system when different terms to the transfer function are added. [Now] I definitely feel more confident about Bode plots"

"The System Id GUI has helped most of all in understanding the estimation of the composition of transfer functions …. The values of each of the elements of the transfer function, which are generally estimable via the analysis of the angular frequency at which changes in the bode plot occur, were an element … that the author had not fully appreciated in the past. Being able to, at a glance, agree or disagree with the value presented in the estimation was helpful"

## VII. Reflection

The students believe that both GUIs have helped them understand the material, and they agreed that teaching asymptotes was beneficial. Although the GUIs were demonstrated in lectures, some difficulties were found, so more help will be given next year as regards the use of both GUIs.

It was interesting to see the comment that it would be more useful to have initially been asked to specify the slope of the gain after the corner frequencies and in later examples to give the value of the gain asymptote at the corner frequency. The first method could be used in familiarization.

Reflecting on this the author realized that for the identification GUI students specified an element type and the system automatically calculated the associated element parameter(s), whereas the plotting GUI expected the student to do relevant calculations. This was inconsistent.

Hence for next year, for the plotting GUI, initially users will enter slopes to specify gain, and later the asymptotic gain at the corner frequencies. Also, the identification GUI will have two modes, one where the parameters are found automatically, and one where a hint is given and the user has to then calculate the parameter. For instance, for a lead-lag element the hint will give data where the phase is a local maximum/minimum using which the student will calculate the two time constants.

## VIII. Conclusion

Two useful GUIs have been produced which the students believe have helped their understanding of frequency response methods. Student feedback is also influencing the development of the GUIs so that next year there will be a better combination of automatic and student-based calculations.

## References

[1] J.A. Rossiter, D. Giaouris, R. Mitchell., and P. Mckenna, 'Typical control curricula and using software for teaching/assessment: a UK perspective'. Proc 17th IFAC World Congress, Seoul, Korea, 2008, pp 10331-6

[2] P. Atkinson. 'Feedback Control Theory for Engineers', Heineman Educational books, 1977.

[3] Robert D. Throne. "Frequency Domain System Identification of One, Two, and Three Degree of Freedom Systems in an Introductory Controls Class", Proc Frequency Domain System Identification of One, Two, and Three Degree of Freedom Systems in an Introductory Controls Class, Paper 2005-493, 2005.

[4] H.W. Bode, 'Network Analysis and Feedback Amplifier Design'. Van Nostrand, 1945.

[5] R.C. Dorf and R. H. Bishop: 'Modern Control Systems', (Pearson, 11th edn. 2008)

[6] N.S. Nise.: 'Control Systems Engineering', (John Wiley & Sons, 5th edn. 2008).

[7] Gibbs, G. & Simpson, C. (2004) Conditions under which assessment supports students' learning *Learning and Teaching in Higher Education,* **1** (1), 3-31

[8] Glover C. and Brown, E. Written Feedback for Students: too much, too detailed or too incomprehensible to be effective? http://www.bioscience.heacademy.ac.uk/journal/vol7/beej-7-3.pdf

## Appendix

This appendix contains the derivation of the range of frequencies around the corner frequency for the extra 'diagonal' phase asymptotes. Conventionally this range is recommended as 100 rad/s [5], and Fig 5 shows these, and the actual phase, for both a single pole and a quadratic pole with damping ratio 0.5.

For a single pole, this approach is quite good, as the area of the region on one side of the asymptote is approximately equal to that on the other. However for a quadratic pole the areas are quite different. A better approach, advocated here, is for the slope of the diagonal asymptote to equal that of the slope of a pole or zero at the corner frequency.

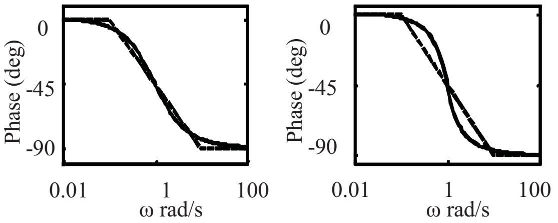


Figure 5. Phase and Asymptotes for Single and Quadratic Pole

As regards, finding the diagonal asymptote, let r be the range over which the diagonal asymptote spans, centred on the corner frequency, $\omega_{CF}$, of a single or a quadratic pole or zero, which changes the phase by nπ. If φ is the phase of such an element, plotted against log(ω), the slope of the asymptote is set by the slope of the element evaluated at $\omega_{CF}$, so

$$\frac{n\pi}{\log(r)} = \left.\frac{d\varphi}{d\log(\omega)}\right|_{\omega_{CF}} = \left.\frac{\omega}{\log(e)}\frac{d\varphi}{d\omega}\right|_{\omega_{CF}} \tag{6}$$

This can be rearranged to give

$$r = e^{n\pi \Big/ \left.\omega\frac{d\varphi}{d\omega}\right|_{\omega_{CF}}} \tag{7}$$

For a single pole or zero, with time constant T, n = ± 0.5 and $\varphi = \pm\tan^{-1}(\omega T)$,

$$\left.\omega\frac{d\varphi}{d\omega}\right|_{\omega_{CF}} = \pm\left.\frac{\omega T}{1+\omega^2 T^2}\right|_{\frac{1}{T}} = \pm\frac{1}{2} \tag{8}$$

$$r = e^{\pi} \tag{9}$$

For a quadratic pole or zero with corner frequency $\omega_n$ and damping ratio ζ, for which n = ±1, it can be shown that

$$\left.\omega\frac{d\varphi}{d\omega}\right|_{\omega_{CF}} = \pm\left.\frac{\omega\, 2\zeta\omega_n\left(\omega_n^2+\omega^2\right)}{\left(\omega_n^2-\omega^2\right)^2+4\zeta^2\omega_n^2\omega^2}\right|_{\omega_n} = \pm\frac{1}{\zeta} \tag{10}$$

$$r = e^{\pi\zeta} \tag{11}$$

Fig 6 shows the phase and these asymptotes for a single and quadratic pole, using these values of r. The phase is easier to sketch than the conventional approach. Instead of crossing the diagonal asymptote and then intersecting with it at the corner frequency, the actual phase moves smoothly between the horizontal asymptote before or after the corner frequency (at some integer * ½ π) and the diagonal asymptote. The gain curve moves between asymptotes in a similar manner.

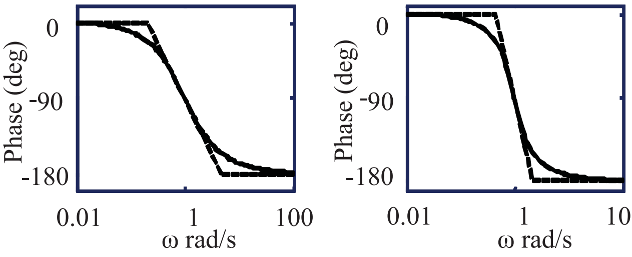


Figure 6 Better Phase Asymptotes for Single and Quadratic Pole

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# Teaching Control Using NI Starter Kit Robot

Payman Shakouri, Member IEEE, (Research student)
Gordana Collier, Member IEEE,
Andrzej Ordys, Senior Member IEEE
School of Mechanical and Automotive Engineering
Kingston University London, UK, SW15 3DW
E-mail: P.Shakouri@kingston.ac.uk

*Abstract*-**Teaching engineering concepts using demonstrations and experiments on real hardware is always engaging and well received by students. This paper provides reference materials (both theoretical and test results), to be used in Control teaching and assessment using a laboratory experiment, with a real-time single board computer based robotic vehicle (National Instruments Robotics Starter Kit). This robotic vehicle is programmed using a graphical programming environment. The Adaptive Cruise Control (ACC) algorithm based on Proportional-Integral (PI) and Proportional – Integral – Derivative (PID) Control are deployed on a field programmable gate array (FPGA), included in the robot's architecture. The robot model (based on a given second order transfer function) is controlled using the same method. The results obtained are compared for the simulation model and a real robot. The performance comparison demonstrates a good correlation between theory and implementation, whilst demonstrating problems and discrepancies introduced by a real system.**

***Keywords: System modelling,* Adaptive Cruise Control (ACC), *PID Controller, Educational robotics, Real-time implementation, Single board computer, FPGA programming.***

## I. INTRODUCTION

Lately, Engineering at Kingston University London has been running a project to develop new Electronics and Control teaching methods. As a result, simulation is not the tool of choice for this task any more. The applied approach to teaching and learning, using latest technology is gaining popularity and interest in the faculty.

This paper proposes an experimental method for teaching control concepts to senior undergraduate engineering students and then assessing the learning by setting a group assignment with the objective to design and implement the software to control the performance of a robotic vehicle with on board Adaptive Cruise Control (ACC), implemented using a PI and a PID loop. Graphical programming language (LabVIEW) was used for code implementation, as it does not deter from the main learning goal by enabling a quick completion of the programming task, allowing students to focus on tuning and testing. The target hardware was a Starter Kit Robot from National Instruments; where two robots were used as 'the leader' and 'the follower'. ACC [1], [2] was used to maintain a set distance between them. These experiments not only illustrate the theoretical concepts but also offer an excellent opportunity for students to programme a real-time, single board computer by directly accessing on-board FPGA, which is likely to open new opportunities for final year project implementation.

The teaching methodology, overall teaching plan (including the formal lectures on control design for distance tracking based on the classical controllers [3], hands-on exercises, and associated assessment task) are presented in Section II. The rest of the paper provides all of the information required by instructors for course delivery, including the solution of the assessment, as follows. Section III describes the hardware and associated system transfer function (given to students), while Section IV deals with the theoretical concepts required. Section V deals with system test, offering all of the results required by the instructor. Section VI discusses the design, tuning and testing of a performance of classical controllers, implemented in ACC. Overall, it verifies the correlation between theoretical expectations and real-life system performance.

## II. UNDERLYING TEACHING STRATEGY

This course is aimed at senior undergraduates, who have moved on from the 'system approach' used in first year to more 'discrete' learning. In year two they study particular concepts, in order to acquire sufficient knowledge in a range of disciplines (programming, electronics and control). They can then amalgamate these concepts and work for the first time with a complex system, applying this knowledge and re-enforcing understanding through implementation.

Students are given a set of formal lectures, covering relevant control concepts as shown in Table I:

Partly sponsored by National Instruments

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

TABLE I

LECTURE SCHEDULE AND MILESTONES

| Week | Milestones |
|---|---|
| 1 | Introduction to modelling of dynamic systems |
| 2 | Use of Laplace transform and transfer functions |
| 3 | Time and feedback domain analysis |
| 4 | Frequency domain analysis |
| 5 | Introduction to PI and PID control including PID loop tuning |
| 6 | ACC and switching rules |

This is followed by a set of tutorials and laboratory exercises as shown in Table II.

TABLE II

LABORATORY SCHEDULE, MILESTONES AND MODE

| Week | Milestones | Mode |
|---|---|---|
| 1 | Introduction to control concepts | Individual |
| 2 | PID tuning using Matlab | Individual |
| 3 | Introduction to LabVIEW and FPGA toolbox | Individual |
| 4 | Introduction to hardware, 'walk-through' exercises (NI single board RIO including FPGA module) | Group (3) |
| 5 - 6 | Implementation and testing (work on group assignment) | Group (3) |

Introduction to the software and hardware is done individually to ensure that every student gets the chance to learn the software and become equipped to work independently. However collaborative learning used in later weeks is highly beneficial when new complex concepts are to be adopted.

Student assessment task aims to:

1. Design an Adaptive cruise control algorithm using MATLAB/SIMULINK using the dynamic model of the robot (given) and obtain optimised solution using PI and PID strategy, in order to perform 'the leader' and 'the follower' action (maintain a certain distance between the leader and follower).
2. Investigate the construction and operation of the real robot, including the incorporated sensors and actuators.
3. Investigate a programming method for a real-time embedded FPGA based system.
4. Apply the Adaptive cruise control strategy developed in 1, to a real robot.
5. Compare system performance in order to discuss the correlation between theory and implementation in terms of PID tuning.
6. Compare the system performance to discuss the correlation between the performance of the robot model and a real robot. Discuss the problems and discrepancies introduced by a real system.

In addition to the written report including the graphs (similar to the ones shown in Sections IV and V), a demo and a viva are also introduced as part of the assessment, to identify the level of learning for individuals and curb plagiarism. Additional benefits of this assessment task are development of 'soft skills' including teamwork, project management and public speaking, with the added benefit of peer support.

## III. HARDWARE CHARACTERISTICS AND SYSTEM TRANSFER FUNCTION

The LabVIEW starter kit robot (Fig. 1) is based on a NI single- board RIO, with Ethernet connectivity. The robot 'follower' uses ultrasonic sensors for velocity and distance tracking of the 'leader' (2cm - 3m sensor range) and two 12V DC brushed motors, (linear velocity of approximately 0.9 [m/s]). The robots can be programmed either by using the high-level LabVIEW robotic starter kit API or by using the LabVIEW FPGA module [4].

The robot model transfer function approximated through the model identification process can be presented as:

$$\frac{v}{v_{sp}} = \frac{1}{0.48s^2 + s + 1} \tag{1}$$

where $v$ and $v_{sp}$ respectively denote the robot's velocity and velocity setpoint. Units are in [m/s].

Fig.1. NI LabVIEW starter kit robot [4]

## IV. CONTROL DESIGN

The Adaptive Cruise Control is based on two controllers; Proportional-Integral-Derivative (PID) control for velocity tracking, and PI control for distance tracking. The distance tracking controller calculates the velocity setpoint using the difference of the current inter-distance between the leading and following robots from the desired distance-headway. The velocity tracking controller computes the velocity setpoint, to maintain the constant desired cruising speed. ACC concept will be discussed in the next section, after tuning of the PID parameters.

### A. PID parameters

The PID controller defined in the FPGA module has the configuration illustrated in Fig. 2. The controller adjusts the speed of the DC motors (in the robot) based on the given set-point speed. The gains of the PID control are set to zero by default; therefore the set-point speed will be directly sent to the DC motor. In this work, we design our controller to calculate the set-point speed which can fulfil the requirement of the ACC system.

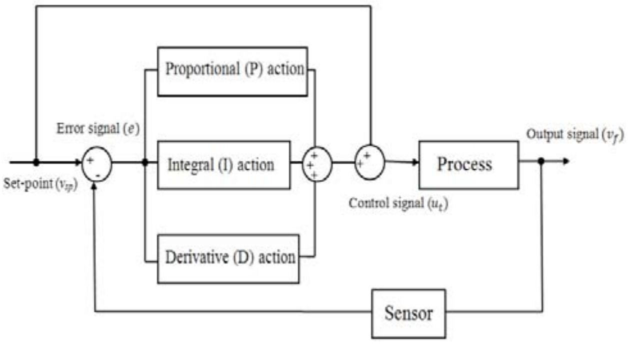
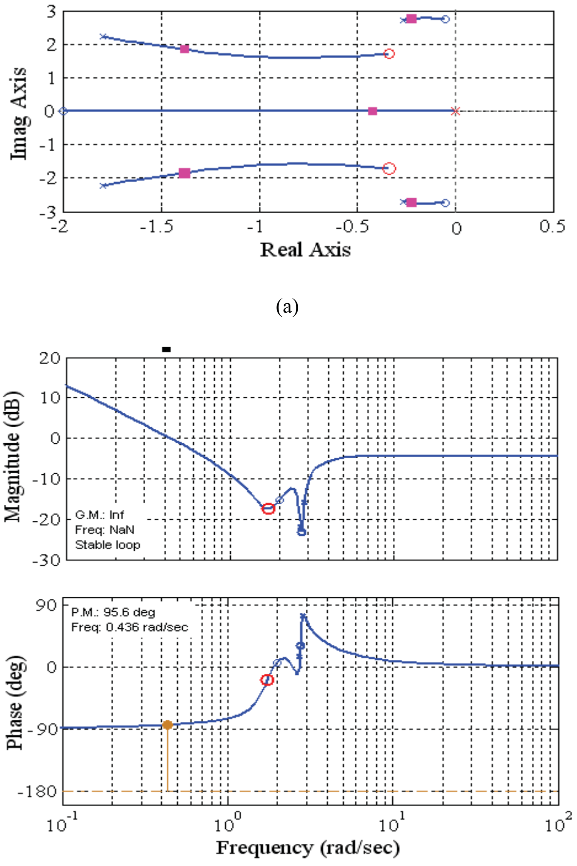


Fig. 2. Configuration of the controller including PID control in the FPFA

Different tuning rules to obtain the Proportional-Integral-Derivative (PID) control gains were introduced in literature [5]; such as, step response (frequency response method), Ziegler-Nichols etc. “Control and Estimation Tools Manager” toolbox in MATLAB/SIMULINK [10] is used. to tune the PI/PID controllers. By using this toolbox, students can easily design the controllers through the interactive plots such as root-locus, Bode, or Nichols, within the SISO design tool. The controller can be graphically tuned by manually moving, adding, or deleting poles and zeros of the model’s tunable blocks, and observing the closed loop response of the system in the analysis plots such as, step response, impulse response, pole/zero response etc. Root locus and open-loop bode diagrams resulted from the tuning process are illustrated in figure 3. These results were obtained using the Control and Estimation toolbox for a single-input-single-output system and selecting singular frequency based tuning for the tuning algorithm. PID controller gains for the three controllers, used for an ACC system, are given in Table III.



(a)

(b)

Fig. 3. (a) Root locus for open loop system, (b) Open-loop bode diagram

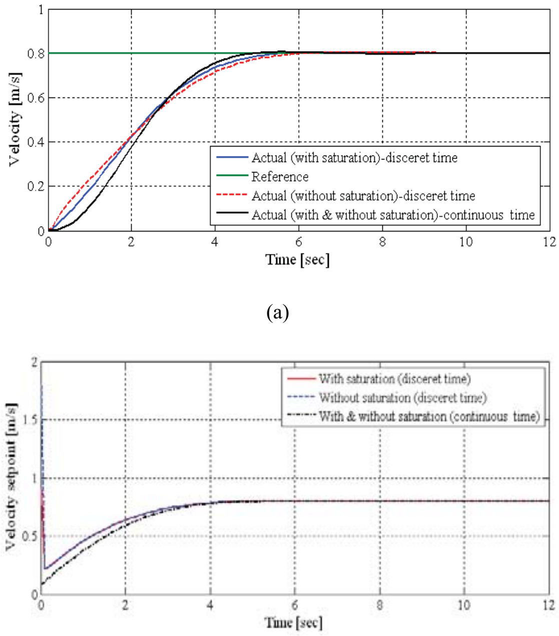
TABLE III

DISCRETE TIME PID CONTROLLERS' GAINS

| | $k_p$ | $k_i$ | $k_d$ |
|---|---|---|---|
| Default PID controller | 0 | 0 | 0 |
| Distance Tracking (ACC) | 0.9 | 0.5 | 0 |
| Velocity Tracking (CC) | 0.2 | 0.48 | 0.2 |

The closed-loop responses of the system given by Eq.1 are presented in Fig. 4. The responses of the closed-loop system in both continuous-time and discrete-time have been compared. The control action (velocity setpoint $v_{sp}$) has to be constrained between [0,0.9], reflecting the physical limitation of the system i.e. motor velocity. The results obtained by saturating the control output have been compared to those without saturation, as illustrated in Fig. 4. Fig. 4. (b) shows that the control action initially violates the upper bound limit if it is not constrained. In most of the system, saturation can degrade the performance, especially when using the PID controller.

This is due to a phenomenon called "integrator windup" which causes a large overshoot in response. To solve this, an anti-windup can be used. Since there was not any serious issue in our design, we did not use anti-windup.



(a)

(b)

Fig.4 (a) the closed-loop responses of the system given by Eq.1 (b) using continuous-time and discrete-time PID controllers, (b) the control actions

## B. ACC switching rules – distance (ACC) and velocity (CC) tracking

ACC systems operate in two different modes depending on the situation in the front - distance tracking or velocity tracking. If the ultrasonic sensor of the follower (ACC equipped) detects any obstacle, or a slower moving robot in front, the controller adjusts the velocity to maintain the clearance inter-distance (desired distance). If the inter-distance measured by the ultrasonic sensor is greater than the desired distance, it will switch to velocity tracking mode, known as cruise control (CC) mode, to track the desired velocity.

The switching logic was devised to implement automated switching between the two modes of operation for the ACC system, i.e. velocity tracking (CC) and distance tracking [3], [6], [7]. The switching rules for transition between CC and ACC modes are illustrated in Table IV.

TABLE IV

LOGICAL RULE FOR SWITCHING BETWEEN CC AND ACC (Shakouri, Ordys and Laila, et al. 2011)

| | $v_f < v_{des}$ | $v_f \geq v_{des}$ |
|---|---|---|
| $d < d_{des}$ | ACC | CC |
| $d \geq d_{des}$ | CC | CC |

A schematic block diagram of this switching algorithm is underlined in close-up of the switching logic CC/ACC in Fig. 5.

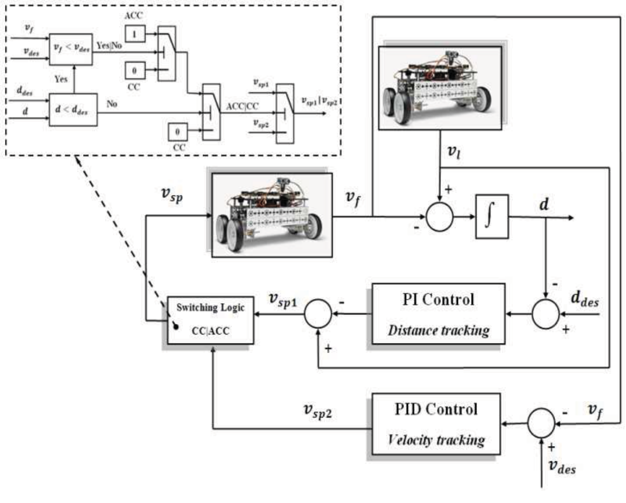


Fig. 5. A schematic block diagram of an ACC system developed in the simulation including a PI control for distance tracking and a PID control velocity tracking.

The desired headway distance $\boldsymbol{d_{des}}$ can be computed using the equation (2), assuming that the follower and leader robot have the same speed ($v_f = v_l$), known as Constant-Time Headway policy [8], [9]:

$$d_{des} = \underbrace{l + d_s}_{d_0} + T_h v_f \qquad (2)$$

*l*- robot length, $d_s$ - additional distance between two robots, $v_f$ - follower robot's velocity and $T_h$ is the constant-time headway (approximates system reaction time) [s].

Distance between real robots is measured by an ultrasonic sensor. In simulation, the distance is determined by taking the integral of their relative velocity:

$$d = \int_0^t (v_f - v_l)dt \qquad (3)$$

The Schematic block diagram from Fig 5, implementing ACC in LabVIEW is shown on Fig. 6.

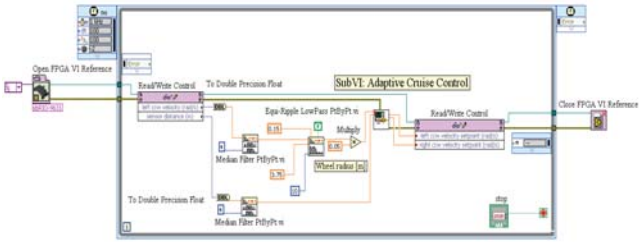
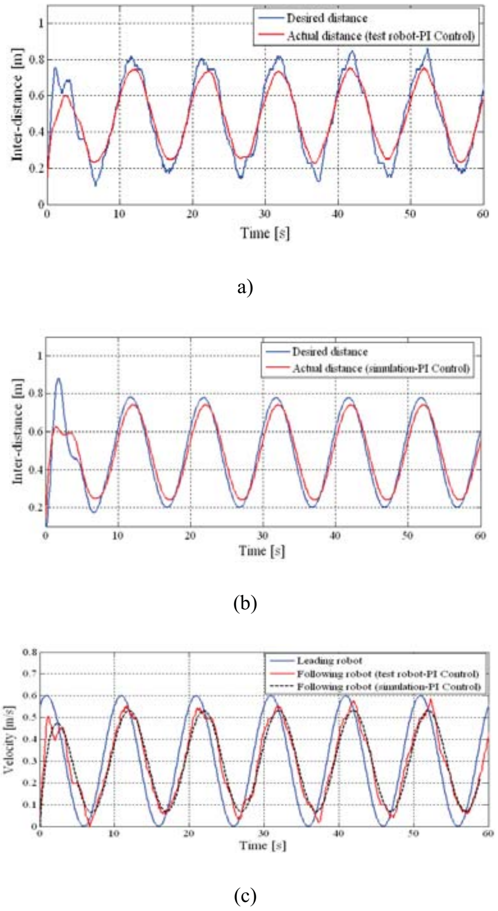


Fig. 6. The configuration of the LabVIEW block functions for implementing the ACC on a robot

## V. Testing and Results

The ACC system was developed and tested using Matlab for PID tuning and LabVIEW for hardware implementation. The programming of the controller has been implemented in LabVIEW MathScript Node. The ACC system was designed and tested in the simulation and implemented on the robot's LabVIEW FPGA module, for comparison. In the initial experiment the leader robot is virtualised, the sinusoidal signal represents the leader robot's velocity and the distance between robots can be calculated from Eq. 8. The final stage of the test used two test robots; and the distance between them is measured by an ultrasonic sensor. The program developed in LabVIEW for real-time implementation is depicted in Fig. 6.

Fig. 7 illustrates the results for distance tracking (ACC) mode for the test executed by using the PID controller. The initial distance between the robots was setup at 0.1 [m]; both robots start from zero velocity. Based on the test, the minimum safe distance $d_o$ and time-constant headway $T_h$ were chosen to be 0.1 [m] and 1.3 [s], respectively. Some slight undulations in the results on the desired distance curve are due to the desired distance being a function of the follower robot velocity; therefore its variation depends on the velocity. However, the ACC system operates very well for distance tracking.



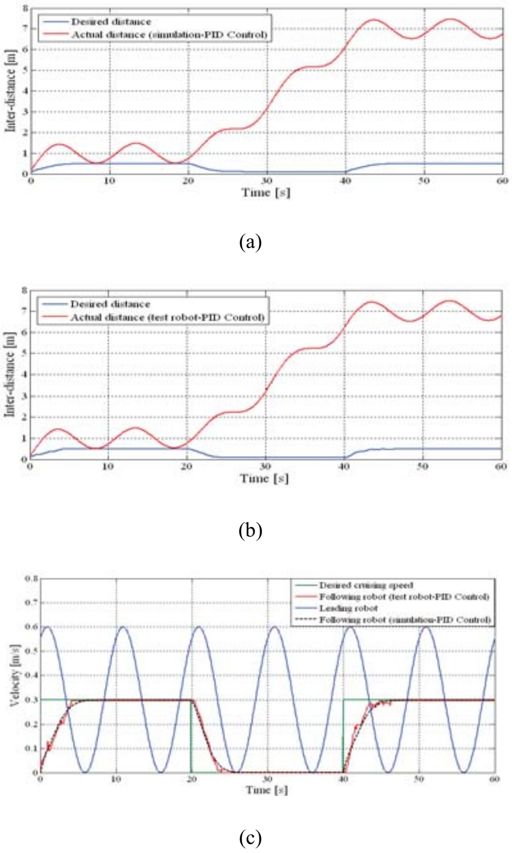
a)

(b)

(c)

Fig. 7. Distance tracking (ACC) mode using PI control- (a) The inter-distance between the robots using test robot, (b) the inter-distance between the robots from simulation, (c) the velocities obtained during distance tracking using test robot and from simulation implementation

Fig. 8 shows the results for velocity tracking (CC) mode accomplished by PID Control with the same initial conditions. However, the desired cruising speed has been selected as a step form; from 0-20 [s], the value of the desired speed is 0.3 [m/s], its value reduces to 0 [m/s] between 20-40 [s] and then it reverts to 0.3 [m/s] from 40 [s] onward. Since the actual distance is far away from the desired distance, the ACC system operates in CC mode, so that the follower robot's velocity tracks the desired cruising speed (see Fig. 8.c), rather than tracking the desired distance. Consequently, it causes the distance between the robots to be increased (see Fig. 8.a-b).



(a)

(b)

(c)

Fig. 8. Velocity tracking (CC) using PID control- (a) The inter-distance between the robots using test robot, (b) the inter-distance between the robots from simulation implementation, (c) the velocities obtained during distance tracking using test robot and from simulation implementation

## VI. CONCLUSIONS

The experiment described here delivers a range of robotics and control concepts, and control strategies, including PI and PID, PID tuning and switching strategies. It requires students to program these strategies and to implement in a real device a target real-time microcontroller board that has been configured as a robot, which is an excellent target system, capturing students' interest and imagination. This paper proposes an assessment brief for the experiment to contain:

1. System modelling and simulation.
2. Implementation in hardware.
3. Testing of the ACC system.

This paper deals with the implementation of only one type of control algorithm and there is scope to implement others e.g. fuzzy logic, or even Model Predictive Control, which can be used for post graduate level teaching. There is also potential for system hardware enhancement/expansion, to include further sensors, for example global positioning systems (GPS), inertial navigation systems (INS), camera etc. (all presently available off the shelf for the specified target board), opening a range of additional opportunities for further control feedback but also experiments and assignments in the field of digital imaging, signal processing, system fusion, robotics, autonomous driving and more.

*The teaching materials described in this paper can be made available on request to the conference participants.*

## REFERENCES

[1] Martinez, J. J., and C. Canudas-de-Wit. *A Safe Longitudinal Control for Adaptive Cruise Control and Stop-and-Go Scenarios.* Vol. 15. IEEE Transactions on Control Systems Technology, 2007.

[2] Xiao, Lingyun, and Feng Gao. "A comprehensive review of the development of adaptive cruise control system." *Vehicle System Dynamics* 48: 10 (April 2010): 1167 — 1192.

[3] Shakouri, P., A. Ordys, D. S. Laila, and M. R. Askari. "Adaptive Cruise Control System: Comparing Gain-Scheduling PI and LQ Controllers." *18th IFAC World Congress.* Milano, Italy, 2011.

*[4] National Instrument.* 15 September 2011. http://zone.ni.com/devzone/cda/tut/p/id/10464.

*[5]* Ogata, K. *Modern Control Engineering.* 3rd. The United State of America: Prentice-Hall, 1997.

[6] Shakouri, P., and A. Ordys. "Application of the State-Dependent Nonlinear Model Predictive Control In Adaptive Cruise Control System." *14th International IEEE Conference on Intelligent Transportation Systems - ITSC 2011.* Washington, DC, USA, at The George Washington University, October 5-7, 2011.

[7] Gerdes, J C, and J K Hedrick. "Vehicle Speed And Spacing Control Via Coordinated Throttle And Brake Actuation." (Control Eng. Practice) 5 (1997): 1607-1614.

[8] van den Bleek, R. A. P. M. *Design of a Hybrid Adaptive Cruise Control Stop-&-Go system.* Master's Thesis, TNO Science & Industry , Technische Universiteit Eindhoven, 2007.

[9] Zhou, W., and S. Zhang. "Analysis of Distance Headways." *Proc. Of the Eastern Asia Society for Transportation Studies.* 2003.

*[10] Control system Toolbox User's Guide.* The MathWorks, Inc., 1999.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Survey and Evaluation of Simulators Suitable for Teaching for Computer Architecture and Organization

## Supporting Undergraduate Students at Sir Syed University of Engineering & Technology

Raza Hasan
Computer Engineering Department
Sir Syed University of Engineering & Technology
Karachi, Pakistan
raza_6@hotmail.com

Salman Mahmood
Computer Engineering Department
Sir Syed University of Engineering & Technology
Karachi, Pakistan
salmanm@ssuet.edu.pk

***Abstract*— Advancement in computer system hardware makes it difficult to meet the demands of teaching computer architecture and organization. Visualization of different architectures enhances the learning process among students by using simulators. This paper attempts to give a survey on the following simulators (1) Electrical Numerical Integrator and Computer (ENIAC) was the first electronic digital universal computer built at Pennsylvania University in 1944-1946., (2) The Visible Virtual Machine (VVM) based on Little Man Computer (LMC) which is general for von Newmann computer architecture, (3) MARS an Education-Oriented MIPS Assembly Language Simulator, (4) Logisim for simulating digital logic circuits and (5) SPIM for MIPS Assembly Language Simulator that are going to be taught in the course Computer Architecture and Organization by the faculty members. Also, evaluate the selected simulators according to the criteria established in the course meetings.**

***Keywords-component Computer Architecture and Organization; Simulation; Evaluation; Electrical Numerical Integrator and Computer; The Visible Virtual Machine; MARS; LOGISIM; SPIM;***

## I. Introduction

Computer-based graphical simulators are widely used in universities to support the teaching of computer architecture [1].These range from relatively simple, visual simulators to advanced, complex simulators for research and product development. One area where software simulators have become almost indispensable is in undergraduate computing courses [2].These simulators are often used to aid student's understanding of complex technologies which are difficult to conceptualize and visualize without the help of graphical animations that the modern simulators can offer [1].

Advancement in computer system hardware makes it difficult to meet the demands of teaching computer architecture and organization. Visualization of different architectures enhances the learning process among students by using simulators. Rather making a comprehensive simulator that can cater the requirement for the course, available simulators can serve the purpose as they can save time resource, ease of use, capacity to learn the concepts to fullest which are delivered in the course, can be accessed anywhere and anytime by the students. Finding the appropriate simulator for the course is a difficult task. Similar study was also conducted [3].

The paper describes the evaluation process followed before the induction of simulators in the course. The paper is organized as follows: Section II gives the teaching strategy followed in the course. Section III describes the criteria for which the simulators were selected. Section IV introduces the evaluation process. Section V describes the questionnaire design. Section VI concludes the results and Section V concludes the paper.

## II. The Teaching Strategy

There was a lack of giving students the exposure of different architectures visualization through graphical representation. Instead, they receive the theoretical aspect of internal structure and components of the computer, how the instruction are being executed and handled by different architectures. This problem can be handled by the use of simulators.

For the batch 2010, currently enrolled in their 5th Semester, the course Computer Architecture and Organization was restructured to two hours of lectures and two hours of laboratory work per week whereas before it was only three hours of lectures per week. This way the gap between theoretical and practical work was removed.

## III. Criteria for Selecting Simulators

As advised by **Chairman** *Computer Engineering Department* **Dr. Syed Misbahuddin** of *Sir Syed University of Engineering & Technology* the simulators that need to be the part of the course should possess certain properties and should be tested by the instructors before it is included in the laboratory work. The characteristics that a simulator should possess are usability, availability and topics covered in the lectures.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

## IV. Evaluation Process

Although simulator imitates a device near to reality but it is necessary to evaluate its effectiveness to which it can be inducted into the course as a teaching aid and to support the learning of the students.

### *A. Methodology*

The evaluation of the simulators was based on the criteria discussed earlier. For evaluation both qualitative and quantitative methods was used. For qualitative method Opinion survey using 5-point Likert scale and having open ended questions as well. For quantitative method describing the data numerically measure of center and location technique was used to interpret the data gathered from the survey.

### *B. Participants*

The participants were composed of eight faculty members teaching the course of Computer Architecture and organization. Among eight, two were Assistant Professors and two were Lecturers who are responsible to give lectures in the lecture room. Two Research Assistant and Junior Lecturer assisting any one of the above in the Laboratory was also included in the survey.

## V. questionnaire Design

The opinion survey questionnaire was given to the faculty members (Sample Size N=8). Each survey was conducted for each simulator making it five surveys consists of four parts. Part A consists of the information about the participant's profile; Part B consists of the information for the simulator usability and the recommendations was taken from []. Part C consists of the information for the simulator availability. Part D consists of the information relevance to the course contents covered by the simulator which was further divided into four sub categories: *Fundamentals of Computer Architecture, Memory System Hierarchy, Communication and Interfacing and Processor System Design.* Values for likert scale in the questionnaire used is as follows 1= Strongly Agree (value = 2) 2= Agree (value = 1) 3= Not Sure/Not Applicable (value = 0) 4= Disagree (value = -1) 5= Strongly Disagree (value = -2)

In the first part, questions 1 to 5 focus on the profile of the respondents. As the respondents type was homogenous so the profile was drawn with close ended questions.

In the second part, questions 6 to 19 focus on the simulator usability. This part of the questionnaire is to achieve the usability of the simulator that needs to be included in the course. To achieve information about the simulator usability the faculty members were asked to evaluate from 1 to 5 their satisfaction and the simplicity about the simulator.

In the third part, questions 20 to 22 focus on the simulator availability. This part of the questionnaire is to find out that the simulator can be used anywhere anytime and on different platforms. To achieve information about the simulator availability the faculty members were asked to evaluate from 1 to 5 their satisfaction.

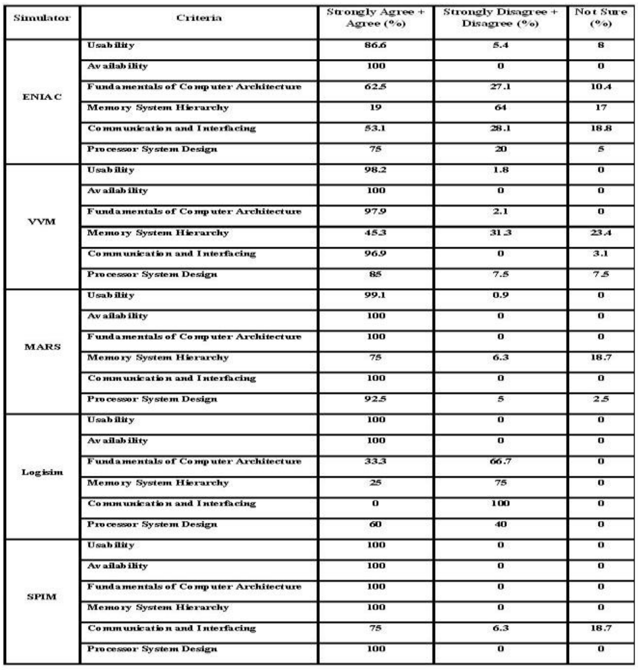
In the fourth part, it is divided in four categories related to the course contents and the topics that the simulator covers. First category *Fundamentals of Computer Architecture*, questions 23 to 28 focus on the fundamental concepts of computer architecture. To achieve information about these, faculty members were asked to evaluate from 1 to 5 their satisfaction. Second category *Memory System Hierarchy*, questions 29 to 36 focus on the performance issues. To achieve information about these faculty members were asked to evaluate from 1 to 5 their satisfaction. Third category *Communication and Interfacing*, questions 37 to 40 focus on the communication between the peripherals and interaction between the components. To achieve information about these faculty members were asked to evaluate from 1 to 5 their satisfaction. Fourth category *Processor System Design*, questions 41 to 45 focus on the substrate of processor logic implementation. To achieve information about these faculty members were asked to evaluate from 1 to 5 their satisfaction.

## VI. Results

Below are the summary of the evaluation based on the opinion survey:

### *A. Qualitative Analysis*

Five opinion surveys were conducted. All before the start of the semester and total of eight faculty members participated in the survey. Figure 1 shows the criteria on which the survey was conducted. The strongly agree/agree and the strongly disagree/disagree results are aggregated and the results are presented below



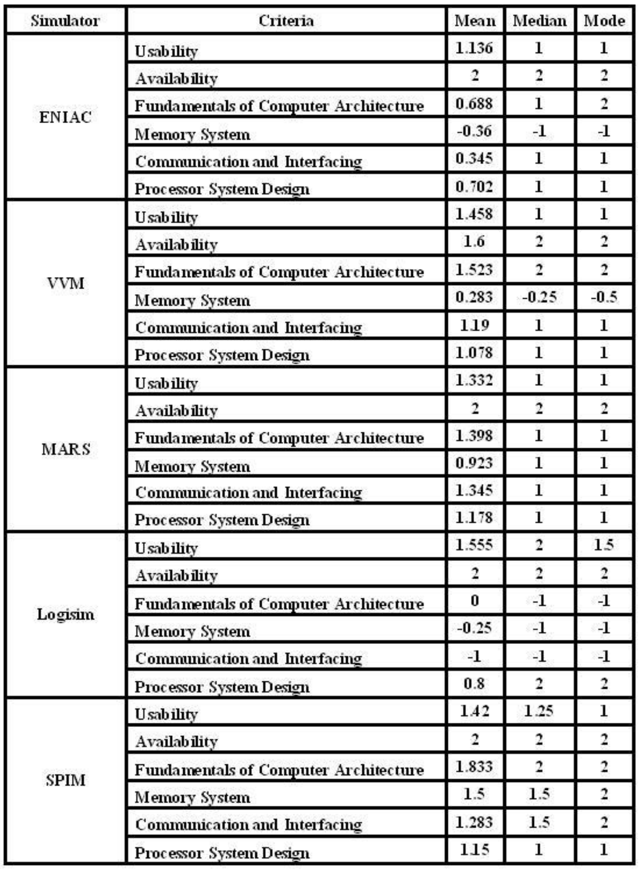
| Simulator | Criteria | Strongly Agree + Agree (%) | Strongly Disagree + Disagree (%) | Not Sure (%) |
|---|---|---|---|---|
| ENIAC | Usability | 86.6 | 5.4 | 8 |
| | Availability | 100 | 0 | 0 |
| | Fundamentals of Computer Architecture | 62.5 | 27.1 | 10.4 |
| | Memory System Hierarchy | 19 | 64 | 17 |
| | Communication and Interfacing | 53.1 | 28.1 | 18.8 |
| | Processor System Design | 75 | 20 | 5 |
| VVM | Usability | 98.2 | 1.8 | 0 |
| | Availability | 100 | 0 | 0 |
| | Fundamentals of Computer Architecture | 97.9 | 2.1 | 0 |
| | Memory System Hierarchy | 45.3 | 31.3 | 23.4 |
| | Communication and Interfacing | 96.9 | 0 | 3.1 |
| | Processor System Design | 85 | 7.5 | 7.5 |
| MARS | Usability | 99.1 | 0.9 | 0 |
| | Availability | 100 | 0 | 0 |
| | Fundamentals of Computer Architecture | 100 | 0 | 0 |
| | Memory System Hierarchy | 75 | 6.3 | 18.7 |
| | Communication and Interfacing | 100 | 0 | 0 |
| | Processor System Design | 92.5 | 5 | 2.5 |
| Logisim | Usability | 100 | 0 | 0 |
| | Availability | 100 | 0 | 0 |
| | Fundamentals of Computer Architecture | 33.3 | 66.7 | 0 |
| | Memory System Hierarchy | 25 | 75 | 0 |
| | Communication and Interfacing | 0 | 100 | 0 |
| | Processor System Design | 60 | 40 | 0 |
| SPIM | Usability | 100 | 0 | 0 |
| | Availability | 100 | 0 | 0 |
| | Fundamentals of Computer Architecture | 100 | 0 | 0 |
| | Memory System Hierarchy | 100 | 0 | 0 |
| | Communication and Interfacing | 75 | 6.3 | 18.7 |
| | Processor System Design | 100 | 0 | 0 |

Figure 1.

The survey indicates that the usability and availability was high for the simulators under examination. Different simulators cover different architectures and depending upon the architecture they cover few concepts high as compared to others. As logisim is a specific for simulating digital logic circuits so it has low percentages in Part D as it covers fewer concepts related to the course as compared to the others. The survey indicates overall, faculty members agreed of the above simulators to be fulfilling the criteria to be included into the course.

### B. *Quantitative Analysis*

Qualitative analysis givens an opinion of the faculty members about the simulators usability, availability and the relevance of contents covered in the course. As the questionnaire consists of ordinal data in likert scale and the result was aggregated in percentage upon the members agreed upon and disagreed upon. To check the discrepancies with the parts of the questionnaire measure of center and location technique was used such as mean, median and mode. Due to the nature of the data collect was ordinal median gives the best measure of the *middle.* Due to the range of scale used for the data collection was between -2 to 2 so the mean values came in negative range. Below are the values



| Simulator | Criteria | Mean | Median | Mode |
|---|---|---|---|---|
| ENIAC | Usability | 1.136 | 1 | 1 |
| | Availability | 2 | 2 | 2 |
| | Fundamentals of Computer Architecture | 0.688 | 1 | 2 |
| | Memory System | -0.36 | -1 | -1 |
| | Communication and Interfacing | 0.345 | 1 | 1 |
| | Processor System Design | 0.702 | 1 | 1 |
| VVM | Usability | 1.458 | 1 | 1 |
| | Availability | 1.6 | 2 | 2 |
| | Fundamentals of Computer Architecture | 1.523 | 2 | 2 |
| | Memory System | 0.283 | -0.25 | -0.5 |
| | Communication and Interfacing | 1.19 | 1 | 1 |
| | Processor System Design | 1.078 | 1 | 1 |
| MARS | Usability | 1.332 | 1 | 1 |
| | Availability | 2 | 2 | 2 |
| | Fundamentals of Computer Architecture | 1.398 | 1 | 1 |
| | Memory System | 0.923 | 1 | 1 |
| | Communication and Interfacing | 1.345 | 1 | 1 |
| | Processor System Design | 1.178 | 1 | 1 |
| Logisim | Usability | 1.555 | 2 | 1.5 |
| | Availability | 2 | 2 | 2 |
| | Fundamentals of Computer Architecture | 0 | -1 | -1 |
| | Memory System | -0.25 | -1 | -1 |
| | Communication and Interfacing | -1 | -1 | -1 |
| | Processor System Design | 0.8 | 2 | 2 |
| SPIM | Usability | 1.42 | 1.25 | 1 |
| | Availability | 2 | 2 | 2 |
| | Fundamentals of Computer Architecture | 1.833 | 2 | 2 |
| | Memory System | 1.5 | 1.5 | 2 |
| | Communication and Interfacing | 1.283 | 1.5 | 2 |
| | Processor System Design | 1.15 | 1 | 1 |

Figure 2.

The values of corresponds correctly to the percentage of the qualitative analysis. It was seen that the aggregate percentage of strongly disagree/ disagree increases made the effect of the quantitative values towards negative. The value of median in decimal was calculated as there are two median values so adding the both values and divided by two. Rest of the percentage of the qualitative analysis correspond the values of median for the quantitative analysis.

## VII. Conclusion & Future Work

Use of simulators in studying Computer Architecture and Organization help the students in understanding the concepts delivered in the classroom. The strategy adapted in the study discussed the usability, availability and the course contents covered by different simulators before inducting them into the laboratory work. Survey was conducted between faculty members and evaluated using qualitative and quantitative analysis.

The evaluation, based on the criteria laid by the faculty members and also the input from the department chairman, as no simulator covers all topics so combination of simulators were surveyed depending upon different architecture and the topics covered in class room. The overall result was satisfactory and it was decided to include all of the simulators in the laboratory work.

In future, survey on simulators would be conducted and see what are the difficulties faced by the students and what students benefited from use of simulators.

Later evaluation will be conducted on the knowledge grasped by students from the use of simulators by comparing the past results of students when there was no laboratory work conducted in comparision to the current result of the students and will see if there is any significant difference in the behavior of the students learning process.

## Acknowledgment

The authors acknowledge the encouragement, support and guidance given by Chairman Computer Engineering Department **Dr. Syed Misbahuddin** to prepare Lectures, Laboratory manual and selecting simulators.

Special thanks to all other faculty members for helping during the research and giving their valuable suggestion and guidance.

## References

[1] B. Chalk, “Evaluation of a cache memory simulator to support the teaching of computer architecture”,3rd LTSN-ICS Annual Conference, August 27-29, 2002, Loughborough, UK. URL: http://myweb.lsbu.ac.uk/~chalkbs/research/research.html

[2] Mustafa, B.; , "Modern computer architecture teaching and learning support: An experience in evaluation," Information Society (i-Society), 2011 International Conference on , vol., no., pp.411-416, 27-29. URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5978481&isnumber=5978433

[3] Nikolic, B. Radivojevic, Z. Djordjevic, J. Milutinovic, V. , "A Survey and Evaluation of Simulators Suitable for Teaching Courses in Computer Architecture and Organization," *Education, IEEE Transactions on* , vol.52, no.4, pp.449-458, Nov. 2009 DOI:10.1109/TE.2008.930097
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4967893&isnumber=5291647.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Backstepping based adaptive sliding mode control for spacecraft attitude maneuvers

Binglong Cong, Xiangdong Liu and Zhen Chen
Key laboratory for Intelligent Control & Decision of Complex Systems, School of Automation
Beijing Institute of Technology, Beijing, China 100081
Email: {cbl, xdliu, chenzhen76}@bit.edu.cn

***Abstract*—This paper aims to address the robust control problem of rigid spacecraft attitude maneuvers in the presence of inertia matrix uncertainty and external disturbance. A backstepping based adaptive sliding mode control (B-ASMC) design is proposed as a solution, where the upper bounds of the parametric uncertainty and disturbance are not required in advance. Compared to current adaptive sliding mode control (ASMC) design, the B-ASMC design has two advantages. Theoretically, the asymptotical stability of the attitude states rather than the sliding function is guaranteed. Practically, the over-adaptation problem in current ASMC design is alleviated and the system performance is improved. Detailed design principle and rigorous closed-loop system stability analysis are provided. A large angle attitude maneuver is employed in the numerical simulation to verify the effectiveness of the proposed algorithm.**

***Index Terms*—attitude maneuver, adaptive sliding mode control, backstepping, over-adaptation.**

## I. Introduction

With the development of aerospace technologies, more and more space missions require that the involved spacecraft implements attitude maneuvers with large angles. Design of an attitude control system for such case poses a challenging problem, including the nonlinear characteristics in the attitude dynamics & kinematics, modeling uncertainty and unexpected external disturbances. Thus, in order to guarantee the control performance, it is necessary to employ nonlinear robust control methods. Sliding mode control (SMC) is a powerful nonlinear control method that is well known for its strong robustness. SMC can provide many good properties, such as insensitivity to model uncertainty, disturbance rejection, and fast dynamic response, which make it a welcome approach for spacecraft attitude control [1]–[4].

According to the equivalent control concept, current SMC algorithms generally consist of two parts, the continuous equivalent control component and the discontinuous switching control component. In order to satisfy the reaching condition, the switching gain should be larger than the upper bounds of the model uncertainty and disturbance. However, those bounds are hard to find in many practical situations. Therefore, conservative design is generally adopted, where the switching gain is selected sufficiently large, such as those in [1]–[4]. Nonetheless, a large switching control component may aggravate the chattering problem which could excite the unmodelled dynamics and may lead to instability.

To eliminate the need of uncertainty and disturbance bounds, adaptive scheme is integrated into SMC design, which is known as the ASMC technique. At the initial stage, it is assumed that the lumped uncertainty was bounded by a linear function of the state-norm. Correspondingly, adaptive laws were designed for the linear function coefficients, as suggested in [5]–[7]. In particular, in [7], an ASMC algorithm was proposed for the attitude stabilization of a rigid spacecraft, where the lumped uncertainty is assumed to be bounded by a linear function of the norms of angular velocity and quaternion. Subsequently, the lumped uncertainty was assumed to be bounded by an unknown constant and consequently a simple adaptive law was proposed for the switching gain calculation in [8]. Subsequent results can be found in many applications such as internal combustion engines [9], induction servomotor [10], planetary gear-type inverted-pendulum [11], etc. However, on the basis of Barbalat lemma, all the ASMC algorithms mentioned above can only guarantee that the sliding function is asymptotically stable but not the system states. And the system performance has not been taken into account.

On the other hand, the backstepping design technique has been widely used to control nonlinear systems with matched or unmatched uncertainties in recent years (see [12] and references therein). The key feature of backstepping design is that it stabilizes the system states through a step-by-step recursive process. Once the final step is completed, the stability of the entire system is guaranteed naturally. However, conventional backstepping design mainly assume that the lumped uncertainty is constant or slowly changing. When the derivative of the lumped uncertainty cannot be regarded as zero, backstepping design with integral adaptive laws are no longer applicable. Recently, there has been continuous efforts to combine the backstepping technique with the SMC method, such as [13]–[15]. Unfortunately, a prior knowledge of the lumped uncertainty bound is required.

Considering the characteristics of both the ASMC method and the backstepping technique, it is natural to combine those two design methodologies to preserve their advantages and at the same time overcome their drawbacks mentioned above, which leads to the proposed B-ASMC design. In this paper, we focus on the robust attitude control for a rigid spacecraft, where the inertia matrix uncertainty and external disturbance are considered. Noticing the cascade structure of the attitude control system, the attitude controller is designed in the backstepping framework, where the ASMC algorithm is

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

designed in the final step to deal with the lumped uncertainty. By virtue of the backstepping design procedure, the proposed B-ASMC algorithm can guarantee the asymptotical stability of the closed-loop system not just the sliding function. Moreover, the system performance can be improved by the proposed algorithm.

## II. Preliminaries

### A. Mathematical Model

Consider a thruster control rigid spacecraft, whose attitude dynamics is governed by the following equation:

$$J\dot{\omega}_b + S(\omega_b)J\omega_b = T_b + T_d \tag{1}$$

where $J \in \mathbb{R}^{3\times 3}$ is the spacecraft inertia matrix, $\omega_b \in \mathbb{R}^3$ denotes the angular velocity vector of $\mathcal{F}_B$, the body-fixed frame, with respect to $\mathcal{F}_I$, the inertia frame. $S(\cdot)$ is the skew-symmetric matrix operator, which is operated as follows:

$$S(\alpha)\beta = \alpha \times \beta$$

where $\alpha$ and $\beta$ are the vectors in $\mathbb{R}^3$. $T_b \in \mathbb{R}^3$ is the vector of control torque provided by the thrusters, $T_d \in \mathbb{R}^3$ is the time-varying external disturbance vector, including environmental and non-environmental disturbance torques. Furthermore, the inertia matrix uncertainty is considered. Let $J = \hat{J} + \Delta J$ with $\Delta J$ the uncertainty caused by the change in mass properties and $\hat{J} = \mathrm{diag}(J_1, J_2, J_3)$ the nominal inertia matrix. Then (1) is described as:

$$\hat{J}\dot{\omega}_b + S(\omega_b)\hat{J}\omega_b = T_b + T_d - \Delta J\dot{\omega}_b - S(\omega_b)\Delta J\omega_b \tag{2}$$

According to the structural feature in (2), one can merge all the elements caused by inertia matrix uncertainty and external disturbance as the lumped uncertainty, i.e., let $d = T_d - \Delta J\dot{\omega}_b - S(\omega_b)\Delta J\omega_b$. Correspondingly, the attitude dynamics is rewritten as:

$$\hat{J}\dot{\omega}_b + S(\omega_b)\hat{J}\omega_b = T_b + d \tag{3}$$

From (3), it is clear that the lumped uncertainty is matched to the system. Without loss of generality, it is assumed that $d$ is smooth and satisfies $\|d\|_\infty < d_{\max}$ with $d_{\max}$ the unknown upper bound and $\|\cdot\|_\infty$ the vector infinity-norm.

As for the attitude representation, quaternion and modified Rodrigues parameters (MRPs) are the two most popular parameters. Quaternion is characterized by its global non-singularity. However, the use of quaternion requires an extra parameter, which leads to a non-minimal parameterization [4]. For the attitude maneuver whose principal angle is within $(-2\pi, 2\pi)$, MRPs can provide a nonsingular minimal attitude description. Moreover, by introducing the shadow MRPs and a switching mechanism, MRPs turn out to be a nonsingular, bounded, minimal attitude representation. Therefore, MRPs are utilized in this paper, whose kinematics is:

$$\dot{\sigma}_b = M(\sigma_b)\omega_b \tag{4}$$

where $\sigma_b \in \mathbb{R}^3$ denotes the inertial MRPs vector of $\mathcal{F}_B$ with respect to $\mathcal{F}_I$. $M(\sigma_b)$ is the Jacobian matrix in the form of $M(\sigma_b) = \dfrac{(1-\|\sigma_b\|^2)I_3 + 2S(\sigma_b) + 2\sigma_b\sigma_b^T}{4}$ with $\|\cdot\|$ the vector 2-norm and $I_3$ the $3 \times 3$ identity matrix. Moreover, $M^T(\sigma_b)M(\sigma_b) = m(\sigma_b)I_3$ with $m(\sigma_b) = (1+\|\sigma_b\|^2)^2/16$. The transition matrix from $\mathcal{F}_I$ to $\mathcal{F}_B$ in terms of $\sigma_b$ is:

$$R(\sigma_b) = I_3 + \frac{8S(\sigma_b)S(\sigma_b) - 4(1-\|\sigma_b\|^2)S(\sigma_b)}{(1+\|\sigma_b\|^2)^2} \tag{5}$$

In this paper, the attitude reorientation control problem is considered. Our goal is reorienting the spacecraft from an arbitrary stationary attitude to a desired attitude with zero angular velocity. Denoting the attitude variables of $\mathcal{F}_D$, the desired frame, as $\sigma_d \in \mathbb{R}^3$ and $\omega_d \in \mathbb{R}^3$, the error attitude variables are defined as follows:

$$\sigma_e = \sigma_b \oplus \sigma_d^* \tag{6}$$

$$\omega_e = \omega_b - R(\sigma_e)\omega_d \tag{7}$$

where $\sigma_e \in \mathbb{R}^3$ is the error MRPs, $\oplus$ is the MRPs production operator, characterizing the successive rotations. For two MRPs expressed in their corresponding frames, e.g., $\sigma_1 \in \mathbb{R}^3$ and $\sigma_2 \in \mathbb{R}^3$, it is operated as follows:

$$\sigma_1 \oplus \sigma_2 = \frac{(1-\|\sigma_2\|^2)\sigma_1 + (1-\|\sigma_1\|^2)\sigma_2 - 2S(\sigma_1)\sigma_2}{1 + \|\sigma_2\|^2\|\sigma_1\|^2 - 2\sigma_2^T\sigma_1}$$

$\sigma_d^*$ is the inverse of $\sigma_d$, which is extracted from $R^{-1}(\sigma_d)$ and $\sigma_d^* = -\sigma_d$, $R(\sigma_e)$ and $R(\sigma_d)$ are the transition matrices from $\mathcal{F}_D$ to $\mathcal{F}_B$ and from $\mathcal{F}_I$ to $\mathcal{F}_D$, and their expressions in terms of $\sigma_e$ and $\sigma_d$ can be obtained by replacing $\sigma_b$ by $\sigma_e$ and $\sigma_d$ in (5). As $\omega_d = 0$, one has $\omega_e = \omega_b$. Therefore, the error attitude dynamics is expressed same as (3). As mentioned in [17], if the attitude variables pairs $(\sigma_b, \omega_b)$ and $(\sigma_d, \omega_d)$ satisfy the MRPs kinematics formulation described in (4), then the error attitude variables pair $(\sigma_e, \omega_e)$ also satisfies the MRPs kinematics formulation. Then, the attitude control system is governed by the following equations:

$$\begin{cases} \hat{J}\dot{\omega}_b + S(\omega_b)\hat{J}\omega_b = T_b + d \\ \dot{\sigma}_e = M(\sigma_e)\omega_b \end{cases} \tag{8}$$

### B. Problem Statement

The control objective can be summarized as follows: design a robust control algorithm to steer the attitude variables pair $(\sigma_b, \omega_b)$ from $(\sigma_b(0), 0)$ to $(\sigma_d, 0)$ (or equivalently render $\lim\limits_{t\to\infty} \sigma_e = \lim\limits_{t\to\infty} \omega_b = 0$) when the lumped uncertainty upper bound $d_{\max}$ is unknown in advance.

## III. Main Results

### A. Conventional ASMC Algorithm Design

In this section, the ASMC algorithm is applied to the attitude control problem under consideration and its major drawback will be revealed. First, define the following nonlinear sliding function $s \in \mathbb{R}^3$:

$$s = \omega_b + \lambda\frac{M^T(\sigma_e)}{m(\sigma_e)}\sigma_e \tag{9}$$

where $\lambda > 0$ is the sliding function gain and $m(\sigma_e)$ can be obtained by replacing $\sigma_b$ in (4) by $\sigma_e$.

According to the design principle presented in [8], following ASMC algorithm can be obtained:

$$T_b = S(\omega_b)\hat{J}\omega_b - \lambda\hat{J}\frac{4M(\sigma_e) - 2\sigma_e\sigma_e^T}{1+\|\sigma_e\|^2}\omega_b - \hat{d}\text{sgn}(s) \quad (10a)$$

with $\hat{d}$ the estimation of $d_{\max}$ which is given by

$$\hat{d} = c\int_0^t \|s\|_1 d\tau \quad (10b)$$

where $c > 0$ is the adaptive gain, $\text{sgn}(\cdot)$ is the sign function and $\|s\|_1 = s^T\text{sgn}(s)$ denotes the vector 1-norm of $s$.

By selecting the Lyapunov candidate function in the form of

$$V = \frac{1}{2}s^T\hat{J}s + \frac{1}{2c}\tilde{d}^2$$

where $\tilde{d} = \hat{d} - d_{\max}$ denotes the estimation error, it is easy to obtain that the time derivative of the above Lyapunov function is

$$\begin{aligned}\dot{V} &= s^T d - d_{\max}s^T\text{sgn}(s)\\ &\le -\eta\|s\|_1\end{aligned} \quad (11)$$

where $\eta = d_{\max} - \|d\|_\infty$. On the basis of the Barbalat lemma, one can conclude that $\lim\limits_{t\to\infty} s = 0$.

There are two major problems of the above ASMC algorithm. Theoretically speaking, $\lim\limits_{t\to\infty} s = 0$ cannot rigorously guarantee $\lim\limits_{t\to\infty}\omega_b = \lim\limits_{t\to\infty}\sigma_e = 0$, i.e., the asymptotic stability of the closed-loop system is not actually achieved. Practically speaking, the ASMC algorithm does not consider the dynamics of the reaching phase, which may arise an over-adaptation of the switching gain with respect to the lumped uncertainty bound and lead to an undesirable system performance. In the following, we try to address those problems by combining the ASMC technique with the backstepping design and present the B-ASMC algorithm.

### B. B-ASMC Algorithm Design

As demonstrated in [18], an important property of the system in (8) is that it describes the attitude control system in a *cascade interconnection*, which accords with the strict feedback form in backstepping design. With this in mind, it is possible to design the ASMC algorithm in the backstepping framework and use its key feature to guarantee the stability of the closed-loop system.

First, in the attitude kinematics subsystem, treat the angular velocity as an independent input, then there exists a state feedback stabilizing control law $\omega_b^*(\sigma_e)$ in the form of

$$\omega_b^*(\sigma_e) = -k_\sigma\frac{M^T(\sigma_e)}{m(\sigma_e)}\sigma_e = -k_\sigma\frac{4}{1+\|\sigma_e\|^2}\sigma_e \quad (12)$$

with $k_\sigma > 0$. Now, consider a Lyapunov candidate function for the attitude kinematics subsystem with the form of $V_\sigma = \|\sigma_e\|^2/2 = \sigma_e^T\sigma_e/2$. From (8) and (12), the derivative of the above Lyapunov candidate is

$$\dot{V}_\sigma = -k_\sigma\sigma_e^T\sigma_e \le 0 \quad (13)$$

If the angular velocity $\omega_b$ is identical to $\omega_b^*(\sigma_e)$, the attitude kinematics subsystem response is characterized by

$$\sigma_e = \exp(-k_\sigma(t-t_i))\sigma_e(t_i) \quad (14)$$

with $t_i$ the time when $\omega_b = \omega_b^*(\sigma_e)$, which implies a good error MRPs response would be achieved.

Then, in order to guarantee $\omega_b$ can track $\omega_b^*(\sigma_e)$, a coordinated transformation is utilized. Let $z = \omega_b - \omega_b^*(\sigma_e) \in \mathbb{R}^3$, the attitude control system described by $\sigma_e$ and $z$ is represented as:

$$\begin{cases}\hat{J}\dot{z} = T_b - S\left(\omega_b^*(\sigma_e)+z\right)\hat{J}\left(\omega_b^*(\sigma_e)+z\right) + d - \hat{J}\dot{\omega}_b^*(\sigma_e)\\ \dot{\sigma}_e = M(\sigma_e)z + M(\sigma_e)\omega_b^*(\sigma_e)\end{cases} \quad (15)$$

where $\dot{\omega}_b^*(\sigma_e)$ can be analytically expressed as

$$\dot{\omega}_b^*(\sigma_e) = -k_\sigma\frac{4M(\sigma_e) - 2\sigma_e\sigma_e^T}{1+\|\sigma_e\|^2}\left(\omega_b^*(\sigma_e)+z\right)$$

Here, using the ASMC methodology, the attitude control law for the attitude dynamics subsystem is designed as:

$$\begin{aligned}T_b =& S\left(\omega_b^*(\sigma_e)+z\right)\hat{J}\left(\omega_b^*(\sigma_e)+z\right) + \hat{J}\dot{\omega}_b^*(\sigma_e)\\ &- k_\omega\hat{J}z - \hat{d}\text{sgn}(z)\end{aligned} \quad (16a)$$

with

$$\hat{d} = c\int_0^t \|z\|_1 d\tau \quad (16b)$$

Consider a Lyapunov candidate function for the attitude dynamics subsystem in the form of

$$V_\omega = \frac{1}{2}z^T\hat{J}z + \frac{1}{2c}\tilde{d}^2 \quad (17)$$

According to (16a) and (16b), the derivative of the above Lyapunov function is

$$\begin{aligned}\dot{V}_\omega &= -k_\omega z^T\hat{J}z + z^T\left[d - \hat{d}\text{sgn}(z)\right] + (\hat{d} - d_{\max})z^T\text{sgn}(z)\\ &= -k_\omega z^T\hat{J}z + z^T d - d_{\max}z^T\text{sgn}(z)\\ &\le -k_\omega z^T\hat{J}z - \eta\|z\|_1 \le -k_\omega z^T\hat{J}z\end{aligned}$$

In order to obtain the control law for the entire system, we should explore the interconnection between the virtual control law in (12) and the attitude control law in (16a) and (16b). Thus, the B-ASMC algorithm is presented as:

$$\begin{aligned}T_b =& S\left(\omega_b^*(\sigma_e)+z\right)\hat{J}\left(\omega_b^*(\sigma_e)+z\right) + \hat{J}\dot{\omega}_b^*(\sigma_e)\\ &- M^T(\sigma_e)\sigma_e - k_\omega\hat{J}z - \hat{d}\text{sgn}(z)\end{aligned} \quad (18a)$$

with

$$\hat{d} = c\int_0^t \|z\|_1 d\tau \quad (18b)$$

Now, we are ready to state the following theorem:

*Theorem 1:* For the attitude control system described in (8), the B-ASMC algorithm in (18a) and (18b) can globally asymptotically stabilize the closed-loop system in the presence that the lumped uncertainty upper bound $d_{\max}$ is unknown in advance.

*Proof 1:* Chose the Lyapunov function for the entire system as

$$V = V_\sigma + V_\omega \tag{19}$$

By taking the time derivative along the system trajectory, one has:

$$\begin{aligned}\dot{V} =& \sigma_e^T M(\sigma_e) z + \sigma_e^T M(\sigma_e)\omega_b^*(\sigma_e) \\ & - k_\omega z^T \hat{J} z - z^T M^T(\sigma_e)\sigma_e + z^T\left[d - d_{\max}\mathrm{sgn}(z)\right] \\ =& - k_\sigma \sigma_e^T \sigma_e - k_\omega z^T \hat{J} z + z^T d - d_{\max} z^T \mathrm{sgn}(z) \\ \leq& - k_\sigma \sigma_e^T \sigma_e - k_\omega z^T \hat{J} z - \eta \|z\|_1 \\ \leq& - k_\sigma \sigma_e^T \sigma_e - k_\omega z^T \hat{J} z\end{aligned}$$

where we have used the fact that $\sigma_e^T M(\sigma_e) z = z^T M^T(\sigma_e)\sigma_e$ and $\|d\|_\infty < d_{\max}$.

Let $\chi = k_\sigma \sigma_e^T \sigma_e + k_\omega z^T \hat{J} z$. It is obvious that $\chi$ is uniformly continuous. By integrating the above equation from zero to $t$, one has:

$$\int_0^t \dot{V} d\tau \leq -\int_0^t \chi d\tau \Rightarrow V(0) \geq \int_0^t \chi d\tau \tag{20}$$

Taking the limits as $t \to \infty$ on both sides of (20) gives

$$\infty > V(0) \geq \lim_{t\to\infty} \int_0^t \chi d\tau \tag{21}$$

On the basis of Barbalat lemma, we can obtain $\lim_{t\to\infty} \chi = 0$, which implies that $\lim_{t\to\infty} \sigma_e = \lim_{t\to\infty} z = 0$. As $\lim_{t\to\infty} \sigma_e = 0$, one has $\lim_{t\to\infty} \omega_b^*(\sigma_e) = 0$. According to the definition of $z$, it is easy to obtain that $\lim_{t\to\infty} \omega_b = 0$. As $V$ is radially unbounded, then we can obtain the conclusion.

### C. Discussions

Here are some remarks:

*Remark 1:* By comparing the ASMC algorithm with the B-ASMC algorithm, one can find that the transformed variable $z$ is actually the sliding function $s$. Therefore, if we rewrite the B-ASMC algorithm in terms of $\sigma_e$ and $s$, there are two additional terms of the B-ASMC algorithm as compared to the ASMC algorithm, $-k_\omega \hat{J} s$ and $-M^T(\sigma_e)\sigma_e$.

The first term is used to improve the system performance by specifying dynamics in the reaching phase. Such a strategy belongs to the so called the *reaching law* method, which was presented in [19]. The *reaching law* is a differential equation which specifies the dynamics of the sliding function. When it is used in the ASMC design, additional benefit can be shown, which has not been fully explored in the literature. For this case, the sliding function dynamics is governed by $\hat{J}\dot{s} = -k_\omega \hat{J} s - \hat{d}\mathrm{sgn}(s)$. As the initial value of $\hat{d}$ is zero, before the adaptation scheme can produce a large enough $\hat{d}$ to satisfy the reaching condition, the term $-k_\omega \hat{J} s$ provides a necessary damping to speed up the reaching phase. On the other hand, it is well known that the basic idea of the ASMC method lies in that the switching gain can be adjusted by the departure from the sliding surface. However, from the adaptive law in (10b), one can see that the integral action starts from the very beginning and any departure from the sliding surface will results in an increase of the switching gain $\hat{d}$. Therefore, if the initial system error is large, or equivalently the initial system trajectory is located far from the sliding surface, the resulting $\hat{d}$ generated by the ASMC algorithm is much larger than the necessary value. Due to the fact that the chattering level is directly determined by the switching gain, the chattering phenomenon is serious in current ASMC design. However, by virtue of $-k_\omega \hat{J} s$, such an over adaptation problem would be weakened due to the fact that part of role of impelling the system trajectory to the sliding surface has been transferred from the $-\hat{d}\mathrm{sgn}(s)$ term to the $-k_\omega \hat{J} s$ term. The parameter $k_\omega$ serves as a tuning parameter dealing with the trade-off between the chattering level and control torque amplitude.

The second term, $-M^T(\sigma_e)\sigma_e$, is used to guarantee the asymptotical stability of the closed-loop system, which has already been verified in the above proof.

*Remark 2:* In [20], the sliding motion with an infinite frequency of the control switching is defined as the *ideal sliding*. In *ideal sliding*, the system trajectory is strictly constrained on the sliding surface. Whereas, due to the switching imperfection, i.e., the switching frequency is finite, sliding motion only takes place in a small neighborhood of the sliding surface, which is defined as the *real sliding*. Recalling the adaptive law in (18b), the switching gain will converge to a bounded value only in *ideal sliding*. However, in *real sliding*, as the sliding function is not identically equal to zero, $\hat{d}$ will become unbounded. For implementation in practice, the adaptive law has to be modified to get a bounded switching gain, such as the so-called $\sigma$-modification in [6]. In this paper, the approach proposed in [20] will be used, where the adaptive law in (18b) is modified as:

$$\hat{d} = \begin{cases} c\int_0^t \|s\|_1 \mathrm{sgn}(\|s\|_1 - \epsilon) d\tau & \text{if} \quad \hat{d} > \mu \\ \int_0^t \mu d\tau & \text{if} \quad \hat{d} \leq \mu \end{cases} \tag{22}$$

where $\mu > 0$ is a very small scalar to ensure $\hat{d}$ is positive and $\epsilon > 0$ is carefully chosen to deal with the trade-off in control accuracy and bounded switching gain. Further details on $\epsilon$ tuning can refer to [20].

## IV. Numerical Simulation

In this section, a large angle attitude maneuver is employed to verify the effectiveness of the proposed B-ASMC algorithm by comparing it with the ASMC algorithm.

The spacecraft inertia matrix for the controller design is $\hat{J} = \mathrm{diag}(48, 25, 61.8)$ (kg.m) and the uncertainty is 10% of the nominal value. $T_d = [\sin(0.2t), 2\cos(0.3t), 3\sin(0.4t)]^T \times 10^{-1}$ (N.m) is the external disturbance. The initial attitude variables of the spacecraft are $\sigma_b(0) = [-0.2, 0.3, 0.1]^T$ and $\omega_b(0) = [0, 0, 0]^T$ (rad/s). The desired attitude is $\sigma_d = [0.1, 0.2, -0.3]^T$ with the desired angular velocity $\omega_d = [0, 0, 0]^T$ (rad/s). The B-ASMC parameters are $k_\sigma = 0.2$, $k_\omega = 0.6$ and the adaptive gain is selected as $c = 1$. For comparison, the ASMC parameter is $\lambda = 0.2 = k_\sigma$ and the adaptive gain is also selected as $c = 1$. The simulation results are shown in Fig.1–Fig.5, where the superscripts $x, y, z$ denote the triaxial components of related vectors.

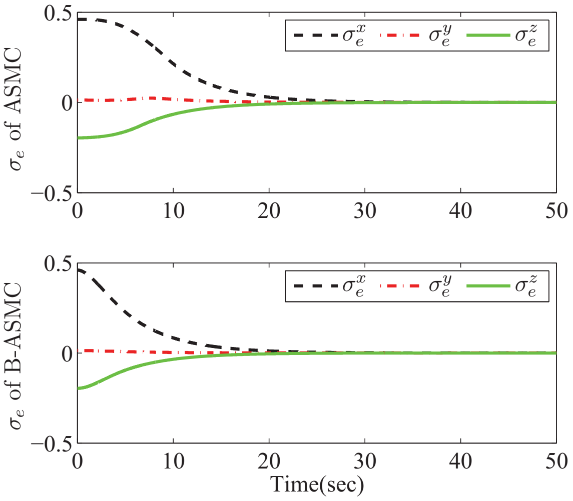


Fig. 1. Error MRPs response comparison

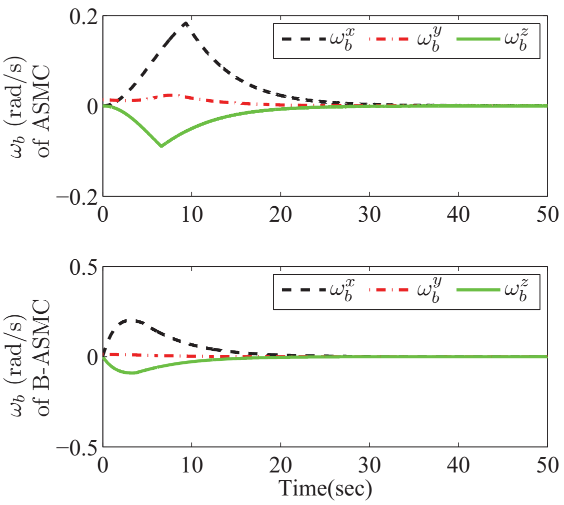


Fig. 2. Angular velocity response comparison

Fig.1 and Fig.2 illustrate the evolutions of the reorientation maneuver controlled by the ASMC algorithm and the B-ASMC algorithm in terms of error MRPs and angular velocity, with the corresponding control torque compared in Fig.3. From Fig.1, we can see that the convergence of the error MRPs controlled by the B-ASMC algorithm is faster than the ASMC algorithm. The fact is that the control torque computed by the ASMC algorithm is zero at the initial time according to (10a) and (10b). Therefore, the convergence is very slow at the beginning.

Fig.3 illustrates the control torque comparison, where the chattering problem in the ASMC algorithm is more serious

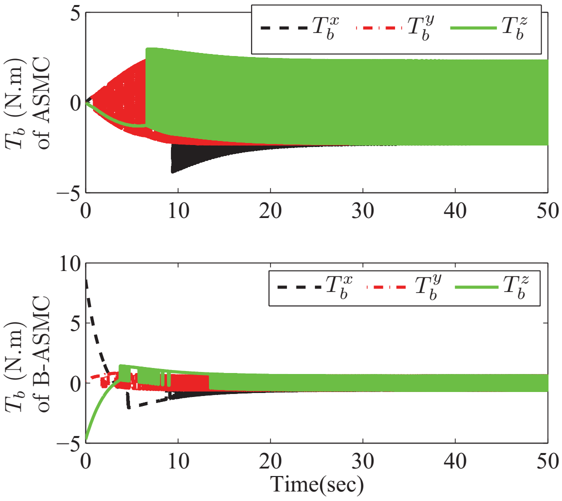


Fig. 3. Control torque response comparison

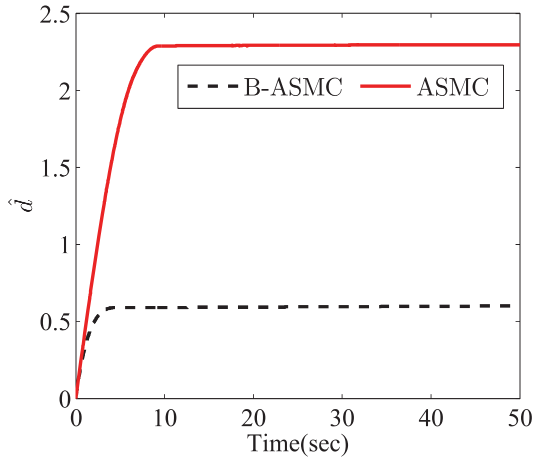


Fig. 4. $\hat{d}$ comparison between two algorithms

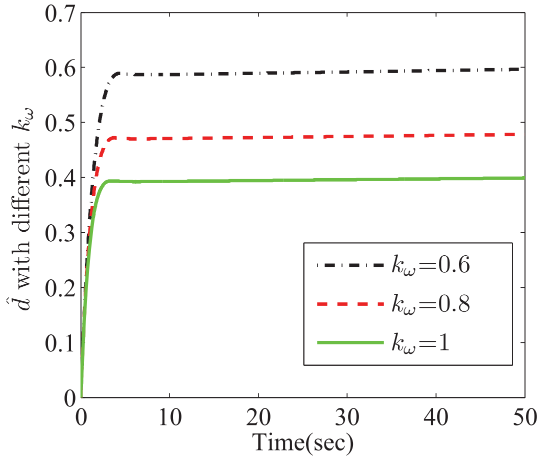


Fig. 5. $\hat{d}$ with different $k_\omega$ in the B-ASMC algorithm

than the B-ASMC algorithm. By examining the switching gains generated by the two control algorithms, as shown in Fig.4, it is clear that the resulting $\hat{d}$ updated by the B-ASMC algorithm is much smaller than the ASMC algorithm, which verifies that the B-ASMC algorithm can weaken the over adaptation problem in the ASMC algorithm and the chattering phenomena is correspondingly reduced. Moreover, $\hat{d}$ generated by the B-ASMC algorithm with different $k_\omega$ is illustrated in Fig.5, from which we can see that the larger the $k_\omega$ is, the smaller the $\hat{d}$ will be and consequently the lower chattering phenomena. However, it should be pointed out that large $k_\omega$ will result in a large initial control torque, as shown in Fig.3.

## V. Conclusion

This paper presents a B-ASMC design for the spacecraft attitude control problem. The proposed algorithm solves the theoretical inadequacy in current ASMC design, where the asymptotical stabilities of the sliding function and the entire closed-loop system are achieved. The system performance is improved by virtue of two additional terms in the control law as compared to current ASMC algorithms. Moreover, the over adaptation problem in ASMC design is also considered and a lower-chattering control signal is achieved. The issue of large initial control torque requirement in B-ASMC will be a topic of in the future work.

## References

[1] S.R. Vadali, “Variable structure control of spacecraft large-angle maneuvers”, *J. Guidance*, vol. 9, no. 2, pp. 235-239, 1986.

[2] S.C. Lo, Y.P. Chen, “Smooth sliding-mode control for spacecraft attitude tracking maneuvers”, *Journal of Guidance, Control, and Dynamics*, vol 18, no. 6, pp. 1345-1349, 1995.

[3] S.A. Kowalchuk, C.D. Hall, “Spacecraft attitude sliding mode controller using reaction wheels”, in *AIAA/AAS Astrodynamics Specialist Conference and Exhibit*, Honolulu, Hawaii, AIAA 2008-6260, 2008.

[4] J.L. Crassidis, “Sliding mode control using modified Rodrgiues parameters”, *Journal of Guidance, Control, and Dynamics*, vol. 19, no. 6, pp. 1381-1383, 1996.

[5] D.S. Yoo, M.J. Chung, “A variable structure control with simple adaptation laws for upper bounds on the norm of the uncertainties”, *IEEE Transactions on Automatic Control*, vol. 7, no. 6, pp. 860–865, 1992.

[6] G. Wheeler, C.Y. Su, Y. Stepanenko, “A sliding mode controller with improved adaption laws for the upper bounds on the norm of uncertainties”, in *Proc. 1996 IEEE Workshop on Variable Structure Systems*, Tokyo, Japan, pp. 154–159, 1996.

[7] Z. Zhu, Y.Q. Xia, and M.Y. Fu, “Adaptive sliding mode control for attitude stabilization with actuator saturation”, *IEEE Transaction on Industrial Electronics*, vol. 58, no. 10, pp. 4898–4907, 2011.

[8] F.J. Lin, S.L. Chiu, “Novel sliding mode controller for synchronous motor drive”, *IEEE Transactions on Aerospace and Electronic Systems*, vol. 34, no. 2, pp. 532–542, 1998.

[9] J.S. Souder, J.K. Hedrick, “Adaptive sliding mode control of air-fuel ratio in internal combustion engines”, *Int. J. Robust Nonlinear Control*, vol. 14, no. 6, pp. 525–541, 2004.

[10] R.J. Wai, “Adaptive sliding mode for induction servomotor drive”, *IEE Proc.-Electr. Power Appl.*, vol. 147, no. 6, pp. 553–562, 2000.

[11] Y.J. Huang, T.C. Kuo and S.H. Chang, “Adaptive sliding-mode conrol for nonlinear systems with uncertain parameters”, *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, vol. 38, no. 2, pp. 534–539, 2008.

[12] M. Krstić, Kanellakopoulos, and P. Kokotović, *Nonlinear and Adaptive Control Design*, New York: Wiley, 1995.

[13] G. Bartolini, A. Ferrara, L. Giacomini, E. Usai, “A combined backstepping/second order sliding mode approach to control a class of nonlinear systems”, in *Proc. 1996 IEEE Workshop on Variable Structure Systems*, Tokyo, Japan, pp. 205–210, 1996.

[14] G. Bartolini, A. Ferrara, L. Giacomini, E. Usai, “Properties of a Combined Adaptive/Second-Order Sliding Mode Control Algorithm for some Classes of Uncertain Nonlinear Systems”, *IEEE Transactions on Automatic Control*, vol.45, no. 7, pp. 1334–1341, 2000.

[15] C.C. Peng, W.T. Hsue Alber, C.L. Chen, “Variable structure based robust backstepping controller design for nonlinear systems”, *Nonlinear Dynamics*, vol. 63, pp. 253-262, 2011.

[16] H. Schaub, J.L. Junkins, *Analytical Mechanics of Space Systems*, AIAA, Virginia, 2009.

[17] J.T.-Y. Wen, K. Kreutz-Delgado, “The attitude control problem”, *IEEE Transactions on Automatic Control*, vol. 36, no. 10, pp. 1148–1162, 1991.

[18] P. Tsiotras, “Further passivity results for the attitude control problem”, *IEEE Transactions on Automatic Control*, vol. 43, no. 11, pp. 1597–1600, 1998.

[19] W.B. Gao, J.C. Hung, “Variable Structure Control of Nonlinear Systems: A new Approach,”, *IEEE Transactions on Industrial Electronics*, vol. 40, no. 1, pp. 45–55, 1993.

[20] F. Plestan, Y. Shtessel, et al, “New methodologies for adaptive sliding mode control”, *International Journal of Control*, vol. 83, no. 9, pp. 1907–1919, 2010.

[21] A. Lavant, “Sliding order and sliding accuracy in sliding mode control”, *International Journal of Control*, vol. 58, no.6, pp. 1247–1263, 1993.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Necessary and Sufficient Conditions for Consensus of Multi-Agent Systems with Nonlinear Dynamics and Variable Topology

Lijing Dong, Senchun Chai, Baihai Zhang
School of Automation
Beijing Institute of Technology
Beijing, China
yangyang304@bit.edu.cn

*Abstract*—**This paper studies some necessary and sufficient conditions for consensus of continuous multi-agent systems with nonlinear node dynamics and variable topology. The multi-agent systems are under variable topology. Basic theoretical analysis is carried out for the case where for each agent the nonlinear dynamics are governed by the position terms of the neighbor nodes. A necessary and sufficient condition associated with eigenvalues is given to ensure consensus of the nonlinear multi-agent system. Based on this result, a simulation example is given to verify the theoretical analysis.**

***Keywords- Nonlinear Multi-agent system; Consensus; Variable Topology***

## I. Introduction

Since consensus of multi-agent systems (MAS) is a fundamental problem in the MAS research area, it has attracted increasing attention of researchers from various disciplines of engineering, biology and science. In networks of agents, consensus means to reach an agreement regarding a certain quantity of interest that depends on the state of all agents. A consensus algorithm is an interaction rule that specifies the information exchange between an agent and all of its neighbors on the network. Such problems have been formulated as consensus of leaderless problems or leader-following problems [1-3]. For a cooperative multi-agent system, leaderless consensus means that each agent updates its state based on local information of its neighbors such that all agents eventually reach an agreement on a common value, while leader-following consensus means that there exists a virtual leader which specifies an objective for all agents to follow.

In the past few years, the multi-agent systems with integer dynamics [4-6] or invariant topology [1, 4, 5] have been widely studied by many researchers due to its simple construction and convenience to analyze. Certainly, there are some researchers spend efforts on multi-agent system with nonlinear dynamics [1, 7] or switching topologies [2, 6, 8], and there have been some outcomes. In [1] a pinning control algorithm was proposed to achieve leader-following consensus in a network of agents with nonlinear second-order dynamics. [7] proposed an adaptive distributed controller with a disturbance estimator to solve the consensus problem under fix topology. By using a common Lyapunov function, [2] extended leader-following consensus control for multi-agent systems, which ensured strong mean square consensus the result, to the switching topology case. In [6], the sampled control protocols are induced from continuous-time linear consensus protocol by using periodic sampling technology and zero-order hold circuit. Nevertheless, since consensus problem of multi-agent associated with both nonlinear dynamics and variable topology is extremely difficult and complicated, people rarely discuss it.

However, considering the fact that almost all the physical plants contain nonlinearity and the communication topology may change from time to time for that the velocity of each agent is time-varying and the communication radius is finite, nonlinear dynamics associated with variable topology consensus is of vital necessity. In order to overcome this problem, some knowledge of complex dynamical network is employed. Broadly speaking, the first-order nonlinear consensus problem of multi-agent systems can be treated as a special case of the synchronization problem of complex dynamical networks, which has been extensively studied in the past few decades [9-11]. [9] presents a sufficient and necessary condition of synchronization criteria for a time-varying complex dynamical network, which is cited as Lemma1 in this paper. Based on the author's work in [9], a necessary and sufficient condition associated with eigenvalues, which is briefer and easier to verify, is given to ensure consensus of the nonlinear multi-agent system with variable topology.

The paper is organized as follows. Some preliminaries of graph theory are briefly reviewed in Section 2. Main results are given in Section 3. To illustrate the proposed theoretical results, a numerical simulation is provided in Section 4. And finally, conclusions are drawn in Section 5.

## II. Problem Statement

### A. Notations

This work was supported by National Natural Science Foundation of Youth Science Fund(No. 61104086) the Key Technology Support Program of Scientific Research Base in Beijing Institute of Technology (No.2011CX02010), Excellent Young Scholars Research Fund of Beijing Institute of Technology(No. 2011CX04031) and Research Foundation of Beijing Institute of Technology(No. 20100642007)

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

Some mathematical notations are used throughout this paper. Denote $I_N \in \mathbb{R}^{N\times N}$ as an N-dimensional identity matrix, $1_N = [1,1,\cdots,1]^T \in \mathbb{R}^N$ as a vector of all ones. Let $A^T$ and $A^{-1}$ be the transpose and the inverse of matrix $A$ , respectively. $\lambda_{\max}(A)$ denotes the maximal eigenvalue of matrix $A$ . $\|\bullet\|$ denotes Euclidean norm.

### *B. Preliminaries in graph theory*

A directed graph, denoted by $G=\{v,\varepsilon,A\}$ be a weighted digraph with a node set $v=\{1,2,\ldots,N\}$ , an edge set $\varepsilon \subseteq v\times v$ and a weighted adjacency matrix $A=(a_{ij})_{N\times N}$ with nonnegative elements [12].We consider that $(i,j)\in\varepsilon$ if and only if vertex (node) $i$ can send its information to vertex $j$ . If $(i,i)$Î $e$ , we say that vertex $i$ has self-loop. In this paper, it is assumed that no self-loop exists. The set of neighbors of vertex $i$ is denoted by $N_i = \{j|\, j$ Î $v, (j,i)$ Î $e\}$, where $j \notin N_i$, which means there is no information flow from vertex $j$ to vertex $i$ , then $a_{ij}=0$ , otherwise $a_{ij}>0$ . The in-degree and out-degree of node $i$ are, respectively, defined as [13].

$$\deg_{in}(i) = \sum_{j=1,j\neq i}^{N} a_{ij} \quad , \quad \deg_{out}(i) = \sum_{j=1,j\neq i}^{N} a_{ji}$$

A digraph is called balanced if $\deg_{in}(i)=\deg_{out}(i)$, $\forall i \in v$ [12].

Denote $D \triangleq diag\{\deg_{in}(1),\deg_{in}(2),\ldots,\deg_{in}(N)\}$. Then the Laplacian matrix $L$ of the graph $G$ is defined as $L \triangleq D-A$ .

### *C. Equations*

The system to be considered in this paper is a multi-agent system composed of N nonlinear coupled agents, labeled from 1to N, which means $v=\{1,2,\ldots,N\}$ . The multi-agent system with variable topology is assumed to have the following dynamics:

$$\dot{x}_i(t) = f(x_i(t)) + u_i(t), i\in v \tag{1}$$

With control protocol

$$\begin{aligned} u_i(t) = \sum_{j\in \mathrm{N}_i(t)}^{N} & a_{ij}(t)(x_i(t)-x_j(t)) + \\ & b_i(t)(x_i(t)-s(t)), i\in v \end{aligned} \tag{2}$$

where $x_i \in \mathbb{R}$ and $u_i \in \mathbb{R}$ are the state and control input of agent $i$ , respectively, $f(x_i)\in\mathbb{R}$ is a nonlinear continuous function to describe the self-dynamics of agent $i$ . $a_{ij}(t)$ is the $(i,j)$ -th entry of the adjacency matrix $A(t)\in\mathbb{R}^{N\times N}$ at time $t$ , where $s\in\mathbb{R}$ is the state of the virtual leader for multi-agent system (1), which is an isolated agent described by $\dot{s}(t)=f(s(t))$ . $b_i(t)$ indicates the accessibility of $s(t)$ by agent $i$ at time $t$ . $b_i(t)>0$ indicates the case that $s(t)$ is accessible by agent $i$ , and $b_i(t)=0$ indicates the case that $s(t)$ is not accessible by agent $i$ .

Let

$$B(t) = diag\{b_1(t), b_2(t), \cdots, b_N(t)\}$$
$$b(t) = col\{b_1(t), b_2(t), \cdots, b_N(t)\}.$$

***Definition 1:*** The multi-agent system (1) is said to achieve synchronization if $\lim_{t\to\infty}\|x_i(t)-s(t)\|=0, i\in \upsilon$ for any initial condition.

Denote $\bar{G}=\{\bar{v},\bar{\varepsilon},\bar{A}\}$ as the graph considering $s(t)$ into $G$ . Denote $\bar{L}(t)=L(t)+B(t)$ as the Laplacian matrices of $\bar{G}$ at time $t$ .

## III. MAIN RESULTS

This section presents necessary and sufficient conditions for nonlinear multi-agent system with variable topology.

Define

$$x(t) = col\{x_1(t), x_2(t), \cdots, x_N(t)\}$$

$$S(t) = col\{s(t), s(t), \cdots, s(t)\}$$

With the fact that $L1_N=0$ and $B1_N=b$, we can rewrite the multi-agent system (1) as

$$\begin{aligned} \dot{x}(t) &= f(x(t)) + (L(t)+B(t))x(t) - B(t)1_N s(t) \\ &= f(x(t)) + (L(t)+B(t))(x(t)-S(t)) \end{aligned} \tag{3}$$

Define

$\xi(t) = col\{\xi_1(t), \xi_2(t), \cdots, \xi_N(t)\}$ with $\xi_i(t) = x_i(t) - s(t)$ .

Then the error closed-loop system can be deduced,

$$\dot{\xi}(t) = f(\xi(t)+s(t)) - f(s(t)) + (L(t)+B(t))\xi(t) \tag{4}$$

***Lemma1****[9]:* Suppose that $Df(s(t))$ , which is the Jacobian of $f$ evaluated at $s(t)$ , is bounded. There exists a real matrix $\phi(t)$ , nonsingular for all $t$ , such that $\phi^{-1}(t)(L(t)+B(t))^T\phi(t) = diag\{\lambda_1(t),\lambda_2(t),\cdots,\lambda_N(t)\}$ and $\dot{\phi}^{-1}(t)\phi(t) = diag\{\beta_1(t),\beta_2(t),\cdots,\beta_N(t)\}$ . The control protocol (2) solves the consensus problem of nonlinear multi-agent system with variable topology is stable if and only if the linear time-varying systems

$$\begin{aligned}&\dot{w}_k(t) = \left[Df(s(t)) + \lambda_k(t) - \beta_k(t)\right] w_k(t)\\ &k = 1,2,\cdots,N\end{aligned} \tag{5}$$

are stable.

For linear continuous time-varying system given by

$$\dot{y}(t) = A(t)y(t), y(t_0) = y_0 \tag{6}$$

with $y(t) \in \mathbb{R}^n$, it is assumed that $A(t)$ is continuous in $t$. The following classical notation of stability, known as stability in sense of Lyapunov is recalled.

***Definition 2:*** Let $t_0$ be any real number. If for all $\delta > 0$, there exists $r(t_0,\delta) > 0$ such that $\left\|y(t_0)\right\| < r(t_0,\delta)$ implies $\left\|y(t)\right\| < \delta$ for all $t \ge t_0$, then the system (6) is stable in the sense of Lyapunov.

Define

$$w(t) = col\{w_1(t), w_2(t), \cdots, w_N(t)\}$$

System (5) can be rewritten as

$$\dot{w}(t) = \Lambda(t)w(t), w(t_0) = w_0 \tag{7}$$

with

$$\begin{aligned}\Lambda(t) = diag\{&Df(s(t)) + \lambda_1(t) - \beta_1(t), Df(s(t)) + \lambda_2(t) - \beta_2(t)\\ &,\cdots, Df(s(t)) + \lambda_N(t) - \beta_N(t)\}\end{aligned}$$

Consider the following sets

$$\Omega_0 = \left\{w_0 \in \mathbb{R}^n : \left\|w_0\right\|^2 \le \rho_0^2, \rho_0 > 0\right\} \tag{8}$$

and

$$E_t = \left\{w \in \mathbb{R}^n : \left\|w(t)\right\|^2 \le \rho(t)^2, \rho(t) > 0, \forall t \ge t_0\right\} \tag{9}$$

with $\Omega_0 \subseteq E_{t0}$, then the following results can be deduced.

***Theorem 1:*** Consider the linear time-varying system (7) and suppose that $w(t_0) = w_0 \in \Omega_0$ . Then $w(t) \in E_t$ for all $t \ge t_0$ if and only if

$$\rho(t) \ge \hat{\rho}(t) = \rho_0 \lambda_{\max}\left(\Phi(t,t_0)\right) \tag{10}$$

where $\Phi(t,t_0)$ is the state transition matrix of linear time-varying system (7).

***Proof***

*Sufficiency:* Consider the following optimization problem for a given instant of time $t$

$$\begin{cases}\max\limits_{w_0} \rho(t)^2 = w_0^T \Phi(t,t_0)^T \Phi(t,t_0) w_0\\ w_0^T w_0 \le \rho_0^2\end{cases} \tag{11}$$

The Lagrangian of this optimization problem is

$$L(w_0,\gamma) = w_0^T \Phi(t,t_0)^T \Phi(t,t_0) w_0 + \gamma\left(w_0^T w_0 - \rho_0^2\right) \tag{12}$$

with $\gamma \le 0$. The optimal condition $\frac{\partial L}{\partial w_0} = 0$ yields

$$2\Phi(t,t_0)^T \Phi(t,t_0) w_0 + 2\gamma w_0 = 0 \tag{13}$$

Considering the reversibility of state transition matrix $\Phi(t,t_0)$ ,(13) can be rewritten as

$$2\gamma\Phi(t,t_0)^T \left(\Phi(t_0,t)^T \Phi(t_0,t) + \gamma^{-1} I\right)\Phi(t,t_0) w_0 = 0 \tag{14}$$

The optimal condition $\frac{\partial L}{\partial \gamma} = 0$ yields

$$w_0^T w_0 - \rho_0^2 = 0 \tag{15}$$

Multiplying (13) on the left by $w_0^T$ , (16) is obtained,

$$w_0^T \Phi(t,t_0)^T \Phi(t,t_0) w_0 + \gamma w_0^T w_0 = 0 \tag{16}$$

and considering (11) and (15), it yields that

$$\rho(t)^2 = -\rho_0^2 \gamma, \forall t \ge t_0 \tag{17}$$

To obtain a solution $w_0 \ne 0$ satisfying condition (14)and (15), it must have

$$\det\left(\Phi(t_0,t)^T \Phi(t_0,t) + \gamma^{-1} I\right) = 0 \tag{18}$$

which implies that $-\gamma^{-1}$ is an eigenvalue of matrix $\Phi(t_0,t)^T \Phi(t_0,t)$ . Then, from (17) and the reversibility of state transition matrix $\Phi(t,t_0)$ , it follows that

$$\begin{aligned}\hat{\rho}(t)^2 &= \max \rho(t)^2\\ &= \rho_0^2 \lambda_{\max}\left(\Phi(t,t_0)^T \Phi(t,t_0)\right)\end{aligned} \tag{19}$$

In this paper, the state matrix of system (7) , $\Lambda(t)$, is diagonal, then the state transition matrix $\Phi(t,t_0)$ is also diagonal, which results that

$$\lambda_{\max}\left(\Phi(t,t_0)^T \Phi(t,t_0)\right) = \lambda_{\max}^2\left(\Phi(t,t_0)\right)$$

Namely that

$$\hat{\rho}(t) = \lambda_{\max}\left(\Phi(t,t_0)\right) \tag{20}$$

If $\rho(t) \ge \hat{\rho}(t)$ for all $t \ge t_0$ , the state of the system is confined in the family of sets $E_t$ .

*Necessity:* Suppose $w(t) \in E_t$ and that for $t > t_0$ , $\rho(t) < \hat{\rho}(t)$ . This implies that there exist both $w_0 \in \Omega_0$ and

$t > t_0$ such that $w(t)$ does not belong to the ellipsoid $E_t$, which leads to a contradiction. The proof is completed.

For the multi-agent system (1) with nonlinear dynamics and variable topology $\bar{G} = \{\bar{v}, \bar{\varepsilon}, \bar{A}\}$, the necessary and sufficient conditions of the protocol (2) solving the consensus problem will be presented as follows.

***Theorem 2:*** Consider the nonlinear multi-agent system (1) with variable topology $\bar{G} = \{\bar{v}, \bar{\varepsilon}, \bar{A}\}$. Then the control protocol (2) solves the consensus problem if and only if

$$\rho_M = \max_{t \geq t_0} \left\{ \rho_0 \lambda_{\max} \left( \Phi(t, t_0) \right) \right\} < \infty$$

where $\Phi(t, t_0)$ is the state transition matrix of linear time-varying system (7).

***Proof.*** According to Lemma1, it is apparent that we just need to proof the stability of system (7) if we want to proof the stability of system (1).

*Sufficiency:* According to Theorem1, we know that for initial condition $\|w_0\|^2 \leq \rho_0^2$, it yields $\|w(t)\|^2 \leq \rho(t)^2$.

Associated with equation (17) $\rho(t)^2 = -\rho_0^2 \gamma, \forall t \geq t_0$, choose $-\gamma = \left( \dfrac{\delta}{\lambda_{\max}\left( \Phi(t, t_0) \right)} \right)^2$, where $\delta > 0$ is an arbitrary scalar.

Consider the set

$$\left\{ w_0 \in \mathbb{R}^n : \|w_0\|^2 \leq r_0^2 \right\} \qquad (21)$$

with $r_0 = \min\limits_{t \geq t_0} \delta \lambda_{\max}^{-1} \left( \Phi(t, t_0) \right) \leq \dfrac{\delta \rho_0}{\rho_M}$

For all time $t \geq t_0$, (21) presents the set of initial conditions such that $\|w(t)\|^2 \leq \delta^2$.

Then, for all $\delta > 0$, there exists $r = \dfrac{\delta \rho_0}{\rho_M}$ such that $\|w_0\| < r$ which results in $\|w(t)\| < \delta, \quad \forall t \geq t_0$ provided that $\rho_M < \infty$. Then, according to Definition2, the system (7) is stable in sense of Lyapunov. This completes the proof of sufficiency.

*Necessity:* Theorem1 defines in the state space a tube containing all the trajectories of the system (7) under $\forall t \geq t_0$. Then, there exists $t = \hat{t}$ corresponding to the $\hat{\rho}(t)$ such that $\|w(t)\|^2 = \rho_M^2$. If the system (7) is stable, $\rho_M$ must be finite, which yields the fact that *r* is independent of the initial time $\hat{t}_0 \geq t_0$. This completes the proof of necessity.

According to the above, we can obtain that the condition $\rho_M = \max\limits_{t \geq t_0} \left\{ \rho_0 \lambda_{\max} \left( \Phi(t, t_0) \right) \right\} < \infty$ ensures the stability of system (7). From Lemma1, the protocol (2) solves the consensus problem of nonlinear multi-agent system (1) with variable topology if and only if $\rho_M < \infty$. This completes the prove.

## IV. Numerical Example

In this section, a numerical example is given to illustrate the Theorem2 with network (1) under control protocol (2). Consider the nonlinear multi-agent system with variable topology:

$$\dot{x}(t) = -\arctan\left(x(t)\right) + \left(L(t) + B(t)\right)\left(x(t) - s(t) 1_N\right) \quad (22)$$

with $\dot{s}(t) = -\arctan\left(s(t)\right)$, $x(t) = \left[x_1(t), x_2(t), x_3(t)\right]$ and $N = 3$ .Then the Jacobian matrix is $Df\left(s(t)\right) = -\dfrac{1}{1 + s(t)^2}$. Assume that the coupling matrix of system (22) is

$$L(t) + B(t) = \frac{1}{2e^2 - e - 1} \begin{pmatrix} \left(e^2 - 1\right) th(t) + e\arctan(t) & (1 - e) th(t) - 2e\arctan(t) & \left(e - e^2\right) th(t) + e\arctan(t) \\ 2\left(e^2 - 1\right) th(t) + e^2 \arctan(t) & 2(1 - e) th(t) - 2e^2 \arctan(t) & 2\left(e - e^2\right) th(t) + e^2 \arctan(t) \\ 3\left(e^2 - 1\right) th(t) + \arctan(t) & 3(1 - e) th(t) - 2\arctan(t) & 3\left(e - e^2\right) th(t) + \arctan(t) \end{pmatrix}$$

It is easy to verify that there exists a nonlinear real matrix

$$\phi(t) = \frac{1}{2e^2 - e - 1} \begin{pmatrix} 3e^2 - 2 & \left(1 - e^2\right) e^t & -e^{1 + \sin(t)} \\ 1 - 3e & (e - 1) e^t & 2e^{1 + \sin(t)} \\ 2e - e^2 & \left(e^2 - e\right) e^t & -e^{1 + \sin(t)} \end{pmatrix}$$

Such that

$$f^{-1}(t)\left(L(t) + B(t)\right)^T f(t) = diag\,\{0, -\,th(t), -\arctan(t)\}$$

and $\dot{f}^{-1}(t) f(t) = diag\,\{0, -1, -\cos(t)\}$

Thus, system (7) is developed as (23) for instantiation.

$$\dot{w}(t) = \breve{L}(t)\, w(t) \qquad (23)$$

with

$$\breve{L}(t)=diag\left\{-\frac{1}{1+s(t)^{2}},-\frac{1}{1+s(t)^{2}}-th(t)+1,\right.$$
$$\left.-\frac{1}{1+s(t)^{2}}+\cos(t)-\arctan(t)\right\}$$

The state transition matrix is

$$\breve{F}(t)=diag\left\{e^{\int_{t_0}^{t}-\frac{1}{1+s(t)^{2}}dt},e^{\int_{t_0}^{t}\left(\frac{s(t)^{2}}{1+s(t)^{2}}-th(t)\right)dt},e^{\int_{t_0}^{t}\left(-\frac{1}{1+s(t)^{2}}+\cos(t)-\arctan(t)\right)dt}\right\}$$

It is obvious that $\max_{t\geq t_0}\left\{\rho_0\lambda_{\max}\left(\Phi(t,t_0)\right)\right\}<\infty$. Therefore, from Theorem2, we know that the consensus of nonlinear multi-agent system (22) with variable topology and control protocol (2) is achieved.

Fig. 1 describes the errors of three followers with the leader in the multi-agent system under the initial condition $s(0)=1$, $x(0)=[-0.2\ \ 0.5\ \ 2]$.

From Fig. 1, we can see the errors of the three following agents with the leader agent tend to zero as time goes on, which means the multi-agent system (22) reaches an absolute consensus on the state $x$. This verified that the necessary and sufficient conditions can solve the consensus problem of multi-agent system with both nonlinear dynamics and variable topology effectively.

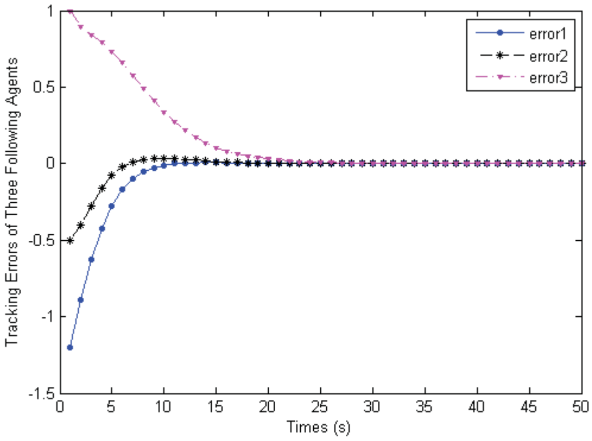


Figure 1. Errors of agents with the leader

## V. Conclusions

A necessary and sufficient condition associated with eigenvalues has been given to ensure consensus of the nonlinear multi-agent system with variable topology. The condition merges the work of [9], which studied controlled synchronization criteria of complex dynamical networks, and some stability theory on linear time-varying system. The necessary and sufficient condition is expressed in a brief way and it is easy to verify.

In the future, second-order consensus problem will be studied based on the work in this paper. Second-order multi-agent system better confirms to the actual situation than the first-order. Therefore, the extension of consensus algorithms from first-order to second-order is non-trivial. The second-order consensus problem is more complicated and challenging.

## References

[1] Q. Song, J. D. Cao and W. W. Yu. Second-order leader-following consensus of nonlinear multi-agent systems via pinning control. Systems & Control Letters 59 (2010) 553-562.

[2] W. Ni and D. Cheng. Leader-following consensus of multi-agent systems under fixed and switching topologies. Systems & Control Letters, 59(3C4)(2010) 209C217.

[3] W. Ren. Consensus tracking under directed interaction topologies: algorithms and experiments. IEEE Trans. Control Syst. Technology, 18(1) (2010)230 -237.

[4] S. H. Li, H. B. Dua, X. Z. Lin. Finite-time consensus algorithm for multiagent systems with double-integrator Dynamics. Automatica 47 (2011) 1706C1712.

[5] J. P. Hua, G. Feng. Distributed tracking control of leader-follower multiagent systems under noisy measurement. Automatica 46 (2010) 1382-1387.

[6] G. M. Xie, H. Y. Liu, L. Wang and Y. M. Jia. Consensus in Networked Multi-Agent Systems via Sampled Control: Switching Topology Case. 2009 American Control Conference, June 10-12, (2009)4525-4530.

[7] K. Sumizaki, L. Liu and S. Hara. Adaptive Consensus on a Class of Nonlinear Multi-Agent Dynamical Systems. SICE Annual Conference 2010 August 18-21(2010) 1141-1145.

[8] U. M¨unz, A. Papachristodoulou, and F. Allg¨ower. Consensus in Multi-Agent Systems With Coupling Delays and Switching Topology. IEEE Trans. Automat. Control, VOL. 56, NO. 12, Decemeber(2011) 2976-2982.

[9] J. H. L¨u, G. R. Chen. A Time-Varying Complex Dynamical Network Model and Its Controlled Synchronization Criteria. IEEE Trans. Automat. Control, VOL. 50, (6) (2005)841-846.

[10] L. M. Pecora, T. L. Carroll. Master stability functions for synchronized coupled systems. Phys. Rev. Lett. 80 (10) (1998) 2109-2112.

[11] W. Lu, T. Chen. New approach to synchronization analysis of linearly coupled ordinary differential systems. Physica D 213 (2006) 214-230.

[12] R. Olfati-Saber, R.M. Murray. Consensus problems in networks of agents with switching topology and time-delays. IEEE Trans. Automat. Control 49 (9) (2004) 1520-1533.

[13] Q. Song, J. Cao. On pinning synchronization of directed and undirected complex dynamical networks. IEEE Trans. Circuits Syst. I 57 (3) (2010) 672-680.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Distributed Practical Consensus in Multi-agent Networks with Communication Constrains

Lulu Li
Department of Mathematics
City University of Hong Kong
Hong Kong, China
Email: lilulu01@gmail.com

Daniel W.C. Ho
Department of Mathematics
City University of Hong Kong
Hong Kong, China
Email: madaniel@cityu.edu.hk

Jianquan Lu
Department of Mathematics
Southeast University
Nanjing 210096, China
Email: jqluma@seu.edu.cn

***Abstract*—This paper deals with the multi-agent consensus problem subject to communication constrains. Two types of communication constrains are discussed in this paper: i) each agent can only exchange quantized data with its neighbors and ii) each agent can only obtain the delayed information from its neighbors. Solutions of the resulting system are defined in the Filippov sense. For the consensus protocol which only considers quantization effect, we prove that Filippov solutions converge to a practical consensus set in a finite time. For the consensus protocol considering quantization and time delay simultaneously, it is shown that Filippov solutions converge to a practical consensus set asymptotically. Moreover, we also present how initial state of the agents, time delay and quantization parameter affect the final practical consensus set. Numerical examples are provided to demonstrate the effectiveness of the obtained theoretical results.**

***Index Terms*—Multi-agent networks, consensus, time delay, quantization.**

## I. Introduction

The concept of consensus originates from the cooperative control, which means all agents reach an agreement on certain quantities of interest. Recently, consensus control has become one of the most focused problems in distributed coordination control of multi-agent networks due to its broad applications in the fields of unmanned aerial vehicles(UAVs), clusters of satellites, automated highway systems and congestion control in communication networks (see, e.g., [1]–[5]).

Due to the physical location, communication bandwidth and unavoidable information losses, only limited information can be sent, transmitted and received by agents in real systems. Hence, consideration of the communication constrains is necessary and important for the design of control strategy or algorithm (see, e.g., [6]–[11]). Two important phenomena in communication are signal quantization and time delay. Unlike the error-free information exchange, the signals in real-world systems are required to be quantized before transmission when high data rate are not available. Some related results about network control system with quantized data have been reported in [8], [11]. In [8], robust $H_\infty$ estimation problems for uncertain systems subject to quantization are investigated. In [11], $H_\infty$ filter is designed for a class of nonlinear discrete time-varying stochastic systems with quantization effects. On the other hand, consensus problems with quantized data are also challenging and desired to be investigated. Discrete-time consensus protocols with quantization have been extensively studied (see [7], [9], [10], [12]–[14]).

Besides quantization, another significant communication constraint in multi-agent networks is the time delay, which is usually caused by an agent waiting to send out message via a busy channel, or by a signal processing and propagation [15]–[17]. It has been shown that conventional consensus protocol with time delay may lead to unexpected results [16].

From the viewpoint of both mathematics and engineering, it is worth noting that the quantization will lead to a system with no solutions in classical sense. Hence, considering solutions in a more general sense is necessary. A ground work have been laid in [18], which extend the quantized consensus model to the continuous-time case and quantized consensus results have been obtained for the network model. In this paper, we shall further extend the previous results of [18] by using different methods. Moreover, we shall address the consensus problems with consideration of the time delay and quantization simultaneously.

The main contributions of this paper are presented as follows:

- Two consensus protocols considering different communication constraints are proposed in this paper. The convergence analysis of the proposed protocols are discussed in detail.
- The effect of the time delay and quantization on the practical consensus set is obtained in this work.

The organization of the remaining part is given as follows. In Section II, some preliminaries about algebra graph theory and discontinuous differential equations are summarized. In Section III and IV, consensus analysis of the proposed protocol are presented in detail. In Section V, two numerical simulation examples are given to show the effectiveness of the theoretical results. In Section VI, concluding remarks are drawn.

$Notations$ : The standard notations will be used in this paper. Throughout this paper, $\mathbb{R}^n$ and $\mathbb{R}^{m\times n}$ represent $n$-dimensional Euclidean space and the set of $m\times n$ real matrices, respectively. $\mathbb{Z}$ denotes a set containing all integer numbers. The superscript "$T$" represents the transpose. $\lfloor\cdot\rfloor$ denotes the lower integer function. $[a, b]$ means the closed interval with endpoints $a$ and $b$. Let $\mathcal{C}([-\tau, 0]; \mathbb{R})$ denote the family of all

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

continuous $\mathbb{R}$-valued functions $g(s)$ on $[-\tau, 0]$ with the norm $\|g(.)\| = sup_{-\tau \leq s \leq 0}|g(s)|$.

## II. Preliminaries and problem formulation

### A. Basic graph theory

Let $\mathcal{G}(\mathcal{V},\, \mathcal{E},\, \mathcal{A})$ be a *weighted directed graph* with the set of nodes $\mathcal{V} = \{v_1,\, v_2,\, \dots, v_N\}$, the set of edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ and a weighted adjacency matrix $\mathcal{A} = [\bar{a}_{ij}]$ with nonnegative adjacency elements $\bar{a}_{ij}$. An edge of $\mathcal{G}$ is denoted by $e_{ij} = (v_i,\, v_j)$, where $v_i$ and $v_j$ are called the parent and child vertices, respectively. For adjacency matrix $\mathcal{A}$, $(v_i,\, v_j) \in \mathcal{E} \Longleftrightarrow \bar{a}_{ji} > 0$. For the undirected graph, one has $\bar{a}_{ij} = \bar{a}_{ji}$. The set of *neighbors* of $v_i$ in $\mathcal{G}$ is denoted by $\mathcal{N}_i(v_i) = \{v_j : (v_j,\, v_i) \in \mathcal{E}\}$. Let $A = \mathcal{A} - \Delta = [a_{ij}]$, where $\Delta = [\Delta_{ij}]_{N\times N}$ is a diagonal matrix with $\Delta_{ii} = \sum_{j=1}^{N} \bar{a}_{ij}$. Then, we have $a_{ij} = \bar{a}_{ij} \geq 0$ for $i \neq j$ and $a_{ii} = -\sum_{j=1,\, j\neq i}^{N} a_{ij}$. The matrix $-A$ is called the *Laplacian* of the directed graph. A *path* in a digraph is an ordered sequence of vertices such that any two consecutive vertices are an directed edge of the digraph. A directed graph $\mathcal{G}$ is *strongly connected* if there is a path for any pair of distinct vertices in $\mathcal{G}$.

### B. Discontinuous differential equations

For differential equations with discontinuous right-hand sides we understand the solutions in terms of differential inclusions following Filippov (1988).

Now we introduce the concept of Filippov solution. Consider the following system

$$\frac{d\mathbf{x}(t)}{dt} = f(\mathbf{x}(t)), \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}^n$ is Lebesgue measurable and locally essentially bounded.

*Definition 1:* A set-valued map is defined as

$$\mathcal{K}(f(\mathbf{x})) = \bigcap_{\delta>0} \bigcap_{\mu(N)=0} \bar{co}[f(B(\mathbf{x}, \delta) \setminus N)], \tag{2}$$

where $\bar{co}(\Omega)$ is the closure of the convex hull of set $\Omega$, $B(\mathbf{x}, \delta) = \{y : \|y - \mathbf{x}\| \leq \delta\}$, and $\mu(N)$ is Lebesgue measure of set $N$. A solution in the sense of Filippov of the Cauchy problem for equation (1) with initial condition $\mathbf{x}(0) = \mathbf{x}_0$ is an absolutely continuous function $\mathbf{x}(t), t \in [0, T]$, which satisfies $\mathbf{x}(0) = \mathbf{x}_0$ and differential inclusion:

$$\frac{d\mathbf{x}}{dt} \in \mathcal{K}(f(\mathbf{x})), \quad a.e.\ t \in [0, T], \tag{3}$$

where $\mathcal{K}(f(\mathbf{x})) = (\mathcal{K}[f_1(\mathbf{x})], \cdots, \mathcal{K}[f_n(\mathbf{x})])$.

Let $h : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz function and $S_h$ be the set of points where $h$ fails to be differentiable. Clarke generalized gradient of $h$ at $\mathbf{x}$ is the set $\partial_c h(\mathbf{x}) = co\{\lim_{i\to+\infty} \nabla h(\mathbf{x}^{(i)}) : \mathbf{x}^{(i)} \to \mathbf{x},\, \mathbf{x}^{(i)} \notin S \cup S_h\}$, where $S$ can be any set of zero measure [19]. A Filippov solution to (3) is a maximal solution if it cannot be extended further in time.

## III. Finite-time Practical Consensus under the effects of quantization

Consider the following multi-agent system with dynamics

$$\dot{x}_i(t) = u_i(t), \quad i = 1, \dots, N,$$

where $x_i(t) \in R$ is the state of the agent $i$, and $u_i(t)$ is called the consensus protocol. The following consensus protocol

$$u_i(t) = \sum_{j\in\mathcal{N}_i} a_{ij}(x_j(t) - x_i(t)), \quad i = 1, \dots, N, \tag{4}$$

has been proposed in [1], which requires that each agent receives information from its neighbors timely and accurately.

Due to the communication bandwidth constraints in many real multi-agent networks, the agents can only use the quantized information of the neighboring agents. The following consensus protocol will be studied in this part.

$$\frac{dx_i(t)}{dt} = \sum_{j\in\mathcal{N}_i} a_{ij}[q_\mu(x_j(t)) - q_\mu(x_i(t))], \qquad i = 1, \dots, N, \tag{5}$$

where $q_\mu(z)$ denotes one-parameter family of uniform quantizers defined by $q_\mu(z) = \lfloor \frac{z}{\Delta\mu} + \frac{1}{2} \rfloor \mu$. Here, $\mu$ and $\Delta$ are called the *quantization parameters* and *error bound* of the quantizer, respectively. Moreover, if $\mathbf{x} = (x_1,\, x_2,\, ...,\, x_N)^T \in \mathbb{R}^N$, we denote $q_\mu(\mathbf{x}) = (q_\mu(x_1),\, q_\mu(x_2),\, ...,\, q_\mu(x_N))^T$.

We know that system (5) may not have the global solution in the sense of *Carathéodory* due to the discontinuous of function $q_\mu(.)$ [18]. Hence, we shall consider solutions in a more general sense, i.e, the Filippov solution of system (5). The concept of the Filippov solution to the differential equation (5) is as follows:

*Definition 2:* A absolutely continuous function $\mathbf{x} : [0,\, T) \to R^N$ is a solution in the sense of Filippov for the discontinuous system (5) if $\mathbf{x}$(t) satisfy that

$$\frac{dx_i(t)}{dt} \in \mathcal{K}[\sum_{j\in\mathcal{N}_i} a_{ij}(q_\mu(x_j(t)) - q_\mu(x_i(t)))], \quad i = 1, \dots, N.$$

Based on the measurable selection theorem of set-valued function (see [20], p. 308, Th. 8.1.3), if $\mathbf{x}(t)$ is a Filippov solution of system (5), then there exists a measurable function $\gamma(t) \in \mathcal{K}[q_\mu(\mathbf{x}(t)))]$ such that for almost all $t \in [0,\, T)$, the following equation is true:

$$\frac{dx_i(t)}{dt} = \sum_{j=1,\, j\neq i}^{N} a_{ij}(\gamma_j(t) - \gamma_i(t)), \qquad i = 1, \dots, N. \tag{6}$$

Any function $\gamma$ as in (6) is called an *output* associated to the solution $\mathbf{x}$.

*Remark 1:* Due to the introduction of the quantization effect, complete consensus cannot be ensured by the proposed protocol, but only *practical consensus* can be achieved. That is, the final consensus values lie within an interval, so called *practical consensus set*, as discussed in [18] and [21].

*Lemma 1:* Suppose $x(t)$ be a Filippov solution to (5). Let $\mathcal{N}=\{1,\ldots,N\}$, $M(t)=\max\limits_{i\in\mathcal{N}} x_i(t)$ and $m(t)=\min\limits_{i\in\mathcal{N}} x_i(t)$. Then, $M(t)$ is a non-increasing function for $t$ and $m(t)$ is a non-decreasing function for $t$.

*Proof:* The proof is similar to the one of Lemma 3, [22]. We omit here due to the length limit. ■

Let $\xi=\{\xi_1,\,\xi_2,\,\ldots,\,\xi_N\}$ be the normalized left eigenvector of matrix $A$ with respect to the zero eigenvalue satisfying $\sum\limits_{i=1}^{N}\xi_i = 1$. It can be obtained that $\xi_i > 0$ from Perron-Frobenius theorem (see [23]). Denote $\mathcal{N}=\{1,\ldots,N\}$ in the following part.

*Theorem 1:* Consider the multi-agent network (5) with a strongly connected graph $G$. The initial conditions associated with (5) are given as $x_i(0)$, $(i=1,2,\ldots,N)$. Let $k=\lfloor\frac{\sum_{i=1}^{N}\xi_i x_i(0)}{\mu\Delta}+\frac{1}{2}\rfloor$. Then $x_i(t)$ will converge to the set $\mathcal{D}=[(k-\frac{1}{2})\Delta\mu,\,(k+\frac{1}{2})\mu\Delta]$ for any $i\in\mathcal{N}$ in a finite time, where $\mu$ and $\Delta$ are quantization parameters and error bound of the quantizer.

*Proof:* The proof of Theorem 1 is divided into two parts.

Part (I) we shall take three steps to prove that each agent in the network will converge to a set $[(k-\frac{1}{2})\mu\Delta,\,(k+\frac{1}{2})\mu\Delta]$ in finite time.

Step 1. Consider the Lyapunov functional as

$$V(t)=\sum_{i=1}^{N}\xi_i\int_0^{x_i(t)} q_\mu(s)ds. \tag{7}$$

Note that $cq_\mu(c)\geq 0$ for any $c\in\mathbb{R}$, we have $V(t)\geq 0$.

Notice that for $p_i(s)=\int_0^s q_\mu(u)du$, we have $\partial p_i(s)=\{v\in\mathbb{R}: q_\mu^-(s)\leq v\leq q_\mu^+(s)\}$. Based on the chain rule (for details, see Proposition 6, [24]), $V(t)$ is differentiable for a.e. $t\geq 0$. Differentiating $V(t)$ along the solution of (6) gives that

$$\begin{aligned}\frac{dV(t)}{dt} &= \sum_{i=1}^{N}\xi_i\gamma_i(t)\sum_{j=1,\,j\neq i}^{N}a_{ij}[\gamma_j(t)-\gamma_i(t)]\\ &= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1,\,j\neq i}^{N}\xi_i a_{ij}[2\gamma_i(t)\gamma_j(t)-2\gamma_i^2(t)].\end{aligned} \tag{8}$$

Notice that

$$\begin{aligned}&\sum_{i=1}^{N}\sum_{j=1,\,j\neq i}^{N}\xi_i a_{ij}\gamma_i^2(t)=\sum_{i=1}^{N}(-a_{ii})\xi_i\gamma_i^2(t)\\ &=\sum_{j=1}^{N}(-a_{jj})\xi_j\gamma_j^2(t)=\sum_{j=1}^{N}\sum_{i=1,\,i\neq j}^{N}\xi_i a_{ij}\gamma_j^2(t)\\ &=\sum_{i=1}^{N}\sum_{j=1,\,j\neq i}^{N}\xi_i a_{ij}\gamma_j^2(t),\end{aligned} \tag{9}$$

we have

$$\frac{dV(t)}{dt}=-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1,\,j\neq i}^{N}\xi_i a_{ij}(\gamma_i(t)-\gamma_j(t))^2. \tag{10}$$

Step 2. Let $\Phi=\{\mathbf{x}(t)\in\mathbb{R}^N :\mid\gamma_i(t)-\gamma_j(t)\mid<\frac{\mu}{N+1},\ \forall i,\,j\in\mathcal{N},\,i\neq j,\,a_{ij}\neq 0\}$. We claim that the agents in the network converge to the set $\Phi$ in finite time.

Let $J=\{t\geq 0:\mathbf{x}(t)\notin\Phi\}$. For $\mathbf{x}(t)\in\mathbb{R}^N$ and $t\in J$, there exist $i,\,j\in\{1,\,2,\,\ldots,\,N\}$, $i\neq j$ and $a_{ij}\neq 0$ such that $\mid\gamma_i(t)-\gamma_j(t)\mid\geq\frac{\mu}{N+1}$. Hence, for $a.e.\,t\in J$

$$\begin{aligned}\dot{V}(t) &\leq -\frac{1}{2}\xi_i a_{ij}(\frac{\mu}{N+1})^2\\ &\leq -\frac{1}{2}\zeta\mu^2,\end{aligned} \tag{11}$$

where $\zeta=\min\limits_{1\leq i,\,j\leq N,\,a_{ij}>0}[(\frac{1}{N+1})^2 a_{ij}\xi_i]$. Further, we can obtain

$$\begin{aligned}0 &\leq V(t)\\ &\leq V(0)-\frac{1}{2}\zeta\mu^2 t,\ t\geq 0.\end{aligned} \tag{12}$$

It follows from $V(0)\geq 0$ that (12) holds if and only if $t\leq\frac{2V(0)}{\zeta\mu^2}$. Therefore, $\mathbf{x}(t)$ will arrive to the set $\Phi$ in finite time.

Step 3. we shall prove that there exists $k\in\mathbb{R}$ such that $x_i(t)$ will converge to the set $[(k-\frac{1}{2})\mu\Delta,\,(k+\frac{1}{2})\mu\Delta]$ in finite time for every $i\in\mathcal{N}$.

Note that $\gamma_i(t)\in\mathcal{K}[q_\mu(x_i(t))]$ and $\gamma_j(t)\in\mathcal{K}[q_\mu(x_j(t)]$, we can get that there exist $k_{ij}\in\mathbb{R}$ such that $x_i(t)$ and $x_j(t)$ belong to the set $[(k_{ij}-\frac{1}{2})\mu\Delta,\,(k_{ij}+\frac{1}{2})\mu\Delta]$ if $\mid\gamma_i(t)-\gamma_j(t)\mid<\frac{\mu}{N+1}$. Hence, based on the proof of step 2, there exists a $T_0\geq 0$ such that $\forall i,\,j\in\mathcal{N},\,i\neq j,\,a_{ij}\neq 0,\,x_i(T_0)$ and $x_j(T_0)$ belong to the set $[(k_{ij}-\frac{1}{2})\mu\Delta,\,(k_{ij}+\frac{1}{2})\mu\Delta]$. Due to the network is strongly connected, there exists a $k\in\mathbb{R}$ such that $k_{ij}=k$.

It follows from Lemma 1 that $x_i(t)\in[(k-\frac{1}{2})\mu\Delta,\,(k+\frac{1}{2})\mu\Delta]$ for $\forall i\in\mathcal{N}$, $t\geq T_0$.

Part (II) Estimate the value of $k$.

Up till now, we have proved the first part of the Theorem 1. Next, we shall give the value of $k$ which is shown to be depending on the initial values of the multi-agent network. Let $\eta(t)=\sum_{i=1}^{N}\xi_i x_i(t)$. We can calculate the derivative of $\eta(t)$ as follows:

$$\begin{aligned}\dot{\eta}(t) &= \sum_{i=1}^{N}\xi_i\sum_{j=1,\,j\neq i}^{N}a_{ij}[\gamma_j(t)-\gamma_i(t)]\\ &= \sum_{i=1,\,i\neq j}^{N}\xi_i a_{ij}\sum_{j=1}^{N}\gamma_j(t)-\sum_{i=1}^{N}\xi_i\gamma_i(t)\sum_{j=1,\,j\neq i}^{N}a_{ij}\\ &= -\sum_{j=1}^{N}\xi_j a_{jj}\gamma_j(t)+\sum_{i=1}^{N}\xi_i a_{ii}\gamma_i(t)\\ &= 0.\end{aligned}$$

Due to $\dot{\eta}(t)=0$ for a.e. $t\in[0,\,\infty)$ and the continuity of $\eta(t)$, it can be easily obtained that $\eta(t)$ is a constant. That is, $\eta(t)=\eta(0)=\sum_{i=1}^{N}\xi_i x_i(0)$.

Next, we need to estimate the value of $k$. Let $\mathcal{D}=[(k-\frac{1}{2})\mu\Delta,\,(k+\frac{1}{2})\mu\Delta]$, we have proved that there exists a $T_0$ such that $x_i(t)\in\mathcal{D},\quad\forall t\geq T_0,\quad\forall i\in\mathcal{N}$.

It follows from $\sum_{i=1}^{N}\xi_i = 1$ that $\sum_{i=1}^{N}\xi_i x_i(t) \in \mathcal{D}, \quad \forall t \geq T_0$. Thus, $\sum_{i=1}^{N}\xi_i x_i(0) \in [(k-\frac{1}{2})\mu\Delta, (k+\frac{1}{2})\mu\Delta]$.

We consider the following two cases:

Case 1: If there exists a $k_0 \in \mathbb{Z}$ such that $\sum_{i=1}^{N}\xi_i x_i(0) = (k_0 - \frac{1}{2})\mu\Delta$, then $k = k_0$ or $k_0 - 1$. Since $x_i(t) \in \mathcal{D}, \quad \forall t \geq T_0, \quad \forall i \in \mathcal{N}$, we have $x_i(t) = (k_0 - \frac{1}{2})\mu\Delta, \quad \forall t \geq T_0$. In this case, we can select $k = k_0$. That is, $k = \frac{1}{\mu\Delta}\sum_{i=1}^{N}\xi_i x_i(0) + \frac{1}{2} = \lfloor \frac{1}{\mu\Delta}\sum_{i=1}^{N}\xi_i x_i(0) + \frac{1}{2}\rfloor$.

Case 2: If $\sum_{i=1}^{N}\xi_i x_i(0) \neq (k_0 - \frac{1}{2})\mu\Delta$ for any $k \in \mathbb{Z}$, then, $\sum_{i=1}^{N}\xi_i x_i(0) \in ((k-\frac{1}{2})\mu\Delta, (k+\frac{1}{2})\mu\Delta)$. Hence, $k = \lfloor \frac{1}{\mu\Delta}\sum_{i=1}^{N}\xi_i x_i(0) + \frac{1}{2}\rfloor$.

Therefore, $k = \lfloor \frac{1}{\mu\Delta}\sum_{i=1}^{N}\xi_i x_i(0) + \frac{1}{2}\rfloor$. This completes the proof of this theorem. ■

*Remark 2:* As discussed in the Theorem 2 and Proposition 4 of [18], the practical consensus results for the model (5) with $\Delta = 1$ are investigated via LaSalle invariance principle of differential inclusions. While in this paper, using different methods, we extend the previous results from the following three aspects:

- We do not assume the network is undirected or balanced.
- We show that the Filippov solutions of (5) reach set $\mathcal{D} = [(k-\frac{1}{2})\mu, (k+\frac{1}{2})\mu]$ in a finite time even if $x_{ave}(0) = \frac{1}{N}\sum_{i=1}^{N} x_i(0) = (k_0 + \frac{1}{2})\mu$ for some $k_0 \in \mathbb{Z}$.
- We present an explicit relationship between the practical consensus set and initial conditions.

*Corollary 1:* Consider the multi-agent network (5) with a strongly connected graph $G$. The initial conditions associated with (5) are given as $x_i(0), (i = 1, 2, \ldots, N)$. Let $k = \lfloor \frac{1}{\mu\Delta}\sum_{i=1}^{N}\xi_i x_i(0) + \frac{1}{2}\rfloor$. Then $x_i(t)$ will converge to the set $\Omega = [\sum_{i=1}^{N}\xi_i x_i(0) - \mu\Delta, \sum_{i=1}^{N}\xi_i x_i(0) + \mu\Delta]$ in a finite time, where $\mu$ and $\Delta$ are quantization parameters and error bound of the quantizer.

*Proof:* According to Theorem 1, $x_i(t)$ converges to the set $[(k-\frac{1}{2})\mu\Delta, (k+\frac{1}{2})\mu\Delta]$ in a finite time, where $k = \lfloor \frac{1}{\mu\Delta}\sum_{i=1}^{N}\xi_i x_i(0) + \frac{1}{2}\rfloor$. It follows from $\sum_{i=1}^{N}\xi_i x_i(0) - \frac{1}{2}\mu\Delta \leq k\mu\Delta \leq \sum_{i=1}^{N}\xi_i x_i(0) + \frac{1}{2}\mu\Delta$ that $x(t)$ will converge to the set $\Omega = [\sum_{i=1}^{N}\xi_i x_i(0) - \mu\Delta, \sum_{i=1}^{N}\xi_i x_i(0) + \mu\Delta]$ in a finite time. ■

*Remark 3:* From Corollary 1, how the initial condition of the agents and quantization parameter $\mu$ affect the practical consensus set $\Omega$ can be observed explicitly. It is interesting to observe that the size of the practical consensus set can be made arbitrarily small by decreasing the quantization parameter $\mu$.

## IV. Practical consensus under quantization and time delay

Considering time delay as another very important communication constraint in the process of information exchange, we propose the following practical consensus protocol:

$$\frac{dx_i(t)}{dt} = \sum_{j\in\mathcal{N}_i} a_{ij}[q_\mu(x_j(t-\tau)) - q_\mu(x_i(t))], \qquad i = 1, \ldots, N, \qquad (13)$$

where $\tau$ is the communication delay from agent $j$ to agent $i$ and $q(.)$ is the same as in model (5). The initial conditions associated with (13) are given as $x_i(s) = \phi_i(s) \in \mathcal{C}([-\tau, 0], R) \quad (i = 1, 2, \ldots, N)$.

Next, we will present the the Filippov solution of system (13).

*Definition 3:* A function $\mathbf{x} : [-\tau, T) \to R^N$ ($T$ might be $\infty$) is a solution in the sense of Filippov for the discontinuous system (13) on $[-\tau, T)$, if

1) $\mathbf{x}$ is continuous on $[-\tau, T)$ and absolutely continuous on $[0, T)$;
2) $\mathbf{x}$(t) satisfy that

$$\frac{dx_i(t)}{dt} \in \mathcal{K}[\sum_{j\in\mathcal{N}_i} a_{ij}(q_\mu(x_j(t-\tau)) - q_\mu(x_i(t)))], \quad i = 1, \ldots, N. \qquad (14)$$

It follows from Theorem 1 in [25] that

$$\mathcal{K}[\sum_{j\in\mathcal{N}_i} a_{ij}(q_\mu(x_j(t-\tau)) - q_\mu(x_i(t)))] \subseteq \sum_{j\in\mathcal{N}_i} a_{ij}(\mathcal{K}[q_\mu(x_j(t-\tau))] - \mathcal{K}[q_\mu(x_i(t))]). \qquad (15)$$

Similar to (6), if $\mathbf{x}(t)$ is the solution of system (13), then there exists a measurable function $\gamma(t) \in \mathcal{K}[q_\mu(\mathbf{x}(t)))]$ such that for a.e. $t \in [0, T)$, the following equation is true:

$$\frac{dx_i(t)}{dt} = \sum_{j=1}^{N} a_{ij}(\gamma_j(t-\tau) - \gamma_i(t)), \quad i = 1, \ldots, N. \qquad (16)$$

Now, we shall present the definition of an initial value problem associated to (13).

*Definition 4:* For any continuous function $\phi : [-\tau, 0] \to \mathbb{R}^N$ and any measurable selection $\psi : [-\tau, 0] \to \mathbb{R}^N$, such that $\psi(s) \in \mathcal{K}[q_\mu(\phi(s))]$ for a.e. $s \in [-\tau, 0]$, an absolute continuous function $\mathbf{x}(t) = \mathbf{x}(t, \phi, \psi)$ is said to be a solution of Cauchy problem for system (13) on $[0, T)$ with initial value $(\phi, \psi)$, if

$$\begin{cases} \dot{x}(t) = \sum_{j=1}^{N} a_{ij}(\gamma_j(t-\tau) - \gamma_i(t)) \\ for\ \ a.e.\ t \in [0, T), \quad i = 1, \ldots, N, \\ \mathbf{x}(s) = \phi(s), \quad \forall s \in [-\tau, 0], \\ \gamma(s) = \psi(s) \quad a.e. \quad s \in [-\tau, 0]. \end{cases} \qquad (17)$$

Note that the solution of the system (17) depends on the initial function $\phi$ and also on the selection of the output $\psi(s) \in \mathcal{K}[q_\mu(\phi(s))]$. In the following part, we shall show that the global solution for system (17) exists. The proof is omitted here due to the length limit.

*Theorem 2:* For any initial function $\phi$ and the selection of the output $\psi(s) \in \mathcal{K}[q_\mu(\phi(s))]$, there exists the global solution for system (17).

Next, we shall present the consensus result under the effects of quantization and time delay simultaneously. The proof will be presented in the full-length version of the paper. The initial conditions associated with (5) are given as $x_i(s) = \phi_i(s) \in \mathcal{C}([-\tau, 0], \mathbb{R}), \quad (i = 1, 2, \ldots, N)$. The Filippov solution of

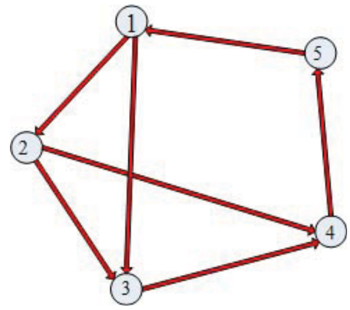


Fig. 1. Network topology in Example 1.

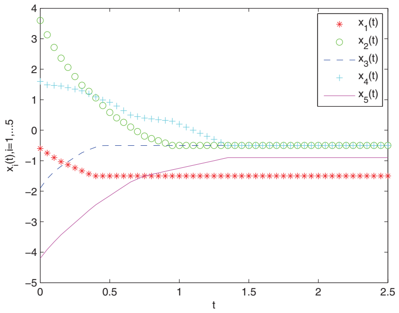


Fig. 2. The states of the multi-agent network in Example 1.

system (13) is defined in (17) and $\psi_j(s)$ $(s \in [-\tau, 0])$ is the initial condition of measurable selection of $\gamma_j(s)$. Let $\eta(0) = \frac{1}{N}\sum_{i=1}^{N} x_i(0) + \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1,\, j\neq i}^{N} a_{ij}\int_{-\tau}^{0}\psi_j(s)ds$ and $A = \mu(\Delta - \frac{\tau}{N}\sum_{i=1}^{N} a_{ii})$, where $\mu$ and $\Delta$ are quantization parameters and error bound of the quantizer.

*Theorem 3:* Consider the undirected multi-agent network (13) with connected graph $G$. Then, for any finite communication delay $\tau$, each agent in the network will converge to the set of $\Omega_1 = [(k-\frac{1}{2})\mu\Delta,\ (k+\frac{1}{2})\mu\Delta]$ asymptotically, where $k = \lfloor\frac{\eta(0)}{A}\rfloor$ or $\lfloor\frac{\eta(0)}{A}\rfloor + 1$.

## V. Numerical examples

In this section, two examples are given to illustrate the correctness of the theoretical results.

*Example 1:* Consider the multi-agent system (5) with five agents, where $\mu = 1$ and $\Delta = 1$. The directed network topology is displayed in Figure 1, and the weight of each edge is set as 1.

Figure 2 shows the state trajectories of (5) with the initial condition randomly chosen from (-5,5). It can be observed from Figure 1 that the state of each agent converges to a practical consensus set in a finite time, which illustrates Theorem 1 very well.

In many real multi-agent consensus problems, the size of this consensus set is required to be very small. In the following example, we will show that the size of the practical consensus set can be reduced by decreasing the quantization parameter $\mu$, which also illustrates Corollary 1 well.

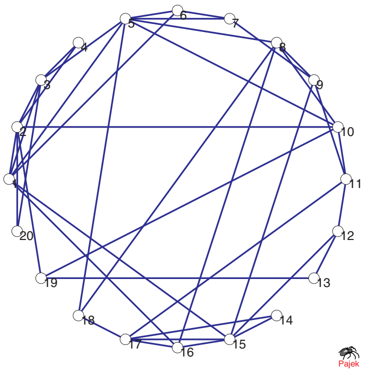


Fig. 3. Network topology in Example 2.

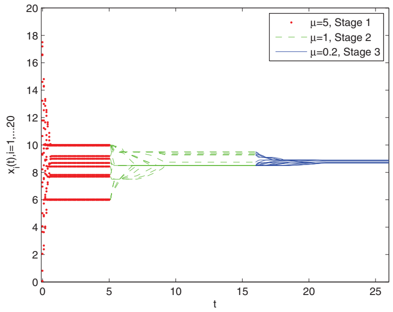


Fig. 4. The states trajectories of the multi-agent networks at different stages with respect to different $\mu$.

*Example 2:* In order to illustrate how quantization parameter $\mu$ affects the practical consensus set and how to reduce the size of the practical consensus set, a multi-agent network (5) with 20 agents is considered. The graph (Figure 3) is assumed to be a small-world network (undirected), in which each node has 2 nearest neighbors and the rewiring probability of the edges is 0.5 (see [26]). Let $\Delta = 1$ and initial conditions are randomly chosen from (0,20). Let the quantization parameter $\mu$ be updated at each stage of practical consensus:

$$\mu = \begin{cases} 5, & 5 > t \geq 0\ (Stage\,1), \\ 1, & 16 > t \geq 5,\ (Stage\,2) \\ 0.2, & 42 > t \geq 16\ (Stage\,3). \end{cases} \tag{18}$$

Note that the value of parameter $\mu$ is reduced by $80\%$ after each stage of practical consensus in order to reduce the size of the consensus set.

Figure 4 shows the states of the 20 agents at different stages with respect to $\mu = 5$, $1$ and $0.2$, respectively. At each stage, all the agents converge to a practical consensus set in a finite time. After a smaller value of $\mu$ is chosen, the agents converge to a new set with a smaller range. The final consensus states of the 20 agents at different stages are respectively shown in Figure 5. From Figures 4 and 5, we can observe that the size of the consensus set is greatly reduced as $\mu$ decreases at different stages. These two figures also verify Corollary 1, that is, the size of consensus set is controlled by $\mu$.

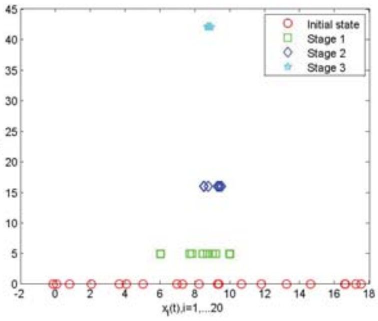


Fig. 5. The final states of the 20 agents at different stages with respect to different $\mu$.

## VI. Conclusion

This paper investigates the continuous-time consensus problem of multi-agent network where each agent can only obtain the quantized and delayed measurements of the states of its neighbors. Filippov solutions of the resulting system exist for any initial condition. We have proved that under certain network topology, the states of the multi-agent network which only considers quantization effect will converge to a practical consensus set in a finite time. For the multi-agent network model considering quantization and time delay simultaneously, it is shown that Filippov solutions converge to a practical consensus set asymptotically. Moreover, we also present how initial state of the agents, time delay and quantization parameter affect the final practical consensus set. The theoretical results have been well illustrated by two numerical examples.

## Acknowledgment

The work was jointly supported by CityU grant 7008188, UK Royal Society, the National Natural Science Foundation of China under Grant 61175119, the Natural Science Foundation of Jiangsu Province of China under Grant BK2010408, Program for New Century Excellent Talents in University (NCET-10-0329), and the Alexander von Humboldt Foundation of Germany.

## References

[1] R. Olfati-Saber and R. Murray, "Consensus problems in networks of agents with switching topology and time-delays," *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1520–1533, 2004.

[2] W. Ren and R. Beard, "Consensus seeking in multiagent systems under dynamically changing interaction topologies," *IEEE Transactions on Automatic Control*, vol. 50, no. 5, pp. 655–661, 2005.

[3] W. Ren, R. Beard, and E. Atkins, "Information consensus in multivehicle cooperative control," *IEEE Control Systems Magazine*, vol. 27, no. 2, pp. 71–82, 2007.

[4] M. Cao, A. Morse, and B. Anderson, "Reaching a consensus in a dynamically changing environment: A graphical approach," *SIAM Journal on Control and Optimization*, vol. 47, no. 2, pp. 575–600, 2008.

[5] B. Shen, Z. Wang, and Y. Hung, "Distributed consensus H-infinity filtering in sensor networks with multiple missing measurements: The finite-horizon case," *Automatica*, vol. 46, no. 10, pp. 1682–1688, 2010.

[6] H. Gao and T. Chen, "H-infinity estimation for uncertain systems with limited communication capacity," *IEEE Transactions on Automatic Control*, vol. 52, no. 11, pp. 2070–2084, 2007.

[7] A. Kashyap, T. Basar, and R. Srikant, "Quantized consensus," *Automatica*, vol. 43, no. 7, pp. 1192–1203, 2007.

[8] H. Gao and T. Chen, "A new approach to quantized feedback control systems," *Automatica*, vol. 44, no. 2, pp. 534–542, 2008.

[9] T. Aysal, M. Coates, and M. Rabbat, "Distributed average consensus with dithered quantization," *IEEE Transactions on Signal Processing*, vol. 56, no. 10, pp. 4905–4918, 2008.

[10] S. Kar and J. Moura, "Distributed consensus algorithms in sensor networks: quantized data and random link failures," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1383–1400, 2010.

[11] B. Shen, Z. Wang, H. Shu, and G. Wei, "Robust h-infinity finite-horizon filtering with randomly occurred nonlinearities and quantization effects," *Automatica*, vol. 46, no. 11, pp. 1743–1751, 2010.

[12] A. Nedic, A. Olshevsky, A. Ozdaglar, and J. Tsitsiklis, "On distributed averaging algorithms and quantization effects," *IEEE Transactions on Automatic Control*, vol. 54, no. 11, pp. 2506–2517, 2009.

[13] P. Frasca, R. Carli, F. Fagnani, and S. Zampieri, "Average consensus on networks with quantized communication," *International Journal of Robust and Nonlinear Control*, vol. 19, no. 16, pp. 1787–1816, 2009.

[14] T. Li, M. Fu, L. Xie, and J. Zhang, "Distributed consensus with limited communication data rate," *IEEE Transactions on Automatic Control*, vol. 56, no. 2, pp. 279–292, 2011.

[15] R. Olfati-Saber, "Flocking for multi-agent dynamic systems: algorithms and theory," *IEEE Transactions on Automatic Control*, vol. 51, no. 3, pp. 401–420, 2006.

[16] F. Xiao and L. Wang, "Consensus protocols for discrete-time multi-agent systems with time-varying delays," *Automatica*, vol. 44, no. 10, pp. 2577–2582, 2008.

[17] J. Lu, D. Ho, and J. Kurths, "Consensus over directed static networks with arbitrary communication delays," *Physical Review E*, vol. 80, p. 066121, 2009.

[18] F. Ceragioli, C. De Persis, and P. Frasca, "Discontinuities and hysteresis in quantized average consensus," *Automatica*, vol. 47, no. 9, pp. 1916–1928, 2011.

[19] F. Clarke, "Optimization and nonsmooth analysis," 1983.

[20] J. Aubin and H. Frankowska, *Set-valued analysis*. Birkhčauser (Boston), 1990.

[21] Q. Hui, "Quantized near-consensus via quantized communication links," *International Journal of Control*, vol. 84, no. 5, pp. 931–946, 2011.

[22] M. Forti, P. Nistri, and D. Papini, "Global exponential stability and global convergence in finite time of delayed neural networks with infinite gain," *IEEE Transactions on Neural Networks*, vol. 16, no. 6, pp. 1449–1463, 2005.

[23] R. Horn and C. Johnson, *Matrix Analysis*. Cambridge University Press, 1990.

[24] M. Forti and P. Nistri, "Global convergence of neural networks with discontinuous neuron activations," *IEEE Transactions on Circuits and Systems*, vol. 50, no. 11, pp. 1421–1435, 2003.

[25] B. Paden and S. Sastry, "A calculus for computing filippov's differential inclusion with application to the variable structure control of robot manipulators," *IEEE Transactions on Circuits and Systems*, vol. 34, no. 1, pp. 73–82, 1987.

[26] D. Watts and S. Strogatz, "Collective dynamics of 'small-world' networks." *Nature*, vol. 393, pp. 440–442, 1998.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Smooth Second-Order Sliding Mode Control Design of PEM Fuel Cell System

**S. Saqib H. Rizvi, A. I Bhatti, Saira Rehman, Qadeer Ahmed, Sohail Iqbal, Zahid Butt**
**CASPR, Dept. of Electronics Engineering**
Mohammad Ali Jinnah University
Islamabad, Pakistan

***Abstract*—A novel smooth second-order sliding mode (SSOSM) control is applied to tackle the problem of PEM fuel cell system output voltage stabilization along with controlling oxygen excess ratio. The oxygen excess ratio is controlled in the inner control loop using PID while the regulation of output voltage is being treated in the outer loop through smooth second-order sliding mode control. The control is applied on the validated nonlinear model of the PEM fuel cell system. The simulation results show that the scheme with SSOSM control provides smoothness (chattering free) with robust performance and the fuel consumption minimization.**

***Keywords-component; Higher (Second) order Smooth Sliding Mode Control; PEM Fuel Cell System***

## I. INTRODUCTION

Petroleum is a fossil fuel and is obtained from the earth's resource of fossil which are finite. The current trend of petroleum consumption is highly increasing with respect to new explorations. If the oil discovery and consumption follow the current trends, estimation shows that the world oil reserves will hardly be used by 2038. Transportation sector is the main user of petroleum and is growing very fast specially in developing countries. This trend needs to optimal use of petroleum with improved efficiency and also require to switch to alternate energy resources.

Fuel cell is a promising alternate energy resource with zero emission environment friendly technology and recent researches in this area show that fuel cell is capable in wide range of application as power source. Different types of fuel cells are suitable for different applications. Polymer Electrolyte Membrane Fuel Cell (PEMFC) shows its potential in, stationary, portable and specially propulsion applications and now in the phase of commercialization.

Fuel Cell System is an electrochemical device that converts chemical energy into electricity directly. Hydrogen and oxygen are its inputs as fuel and oxidant respectively. The voltage is its output which doesn't inherit stabilized form. The voltage varies due to chemical reactions and fluid dynamics. It also degrades due to increase in the load. The variable and degraded voltage is not useful for all electric appliances because they need regulated voltage for their proper working and their long life. That is why; the stabilization of output voltage is a challenging problem for researchers.

The desired output voltage can be maintained through an effective control system. The controller can stabilize the voltage efficiently manipulating inputs. This manipulation can made the system economical avoiding conservative use of inputs. A typical propulsion system based on PEM fuel cell is shown in Fig. 1.

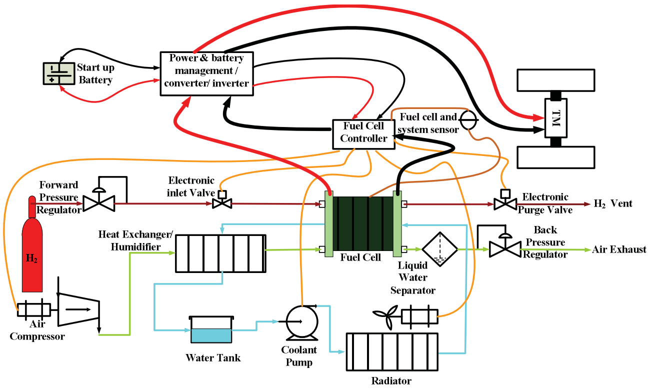


Figure 1. PEM Fuel Cell System for Propulsion

Most of the vast literature available on fuel cells deals with the steady-state conditions. Many researchers attempted to develop electrochemistry based models of PEM fuel cell [8, 9]. The work presented by Lu Ying Chiu and Tanrioven is aimed at developing dynamic models of PEM fuel cell [10, 11]. The dynamic model and design of combined fuel cell and ultra-capacitor system for stand-alone residential applications is developed in [12].

Some models focused on the description of the fuel cell stack and formed the basis for many 2D and 3D finite element models that allow predicting local reactant concentration, temperature and current density [13-15]. Kim et al, Lee et al, and Mann et al proposed generic models that can be adjusted to any fuel cell by adjusting a certain number of fitting parameters [16-18]. Rodatz developed a one dimensional, steady state fuel cell model. Based on stack temperature, current, anode and cathode humidity and pressures, it describes the cell voltage [19].

Although these models provide certain understanding of PEM fuel cell, they cannot be used to perform control algorithm development for PEM fuel cell system. Such models were intended for simulation or parameter estimation rather

This work is supported by Higher Education Commission (HEC) of Pakistan.

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

than control analysis. Other PEM fuel cell models available in the open literature cannot be used for simulations and control purposes because the information about the parameters was not listed completely. Some of them were incomplete and too complex as well as computationally intensive to be used for real-time applications. To develop refined control strategies for PEM fuel cell system, an accurate mathematical model of PEM fuel cell system is necessary.

In this wok, a dynamic non-linear mathematical model for 500Watt PEM fuel cell [6, 7], is used. This model describes the parameters that have a physical significance so that it can be adapted to a given system. It describes correctly the interaction between the different subsystems from a controls point of view. The results of the non-linear PEM fuel cell model simulation in the Simulink/Matlab environment ensure the perfectness of the Model.

The Avista SR-12 system is designed to produce power up to 500Watt (25V at 20A). The specifications of the SR-12 system, as provided by Avista [20], are shown in Appendix C [37].

In this work, PEM fuel cell model is studied in section II and a proposed scheme is developed in section III, controller synthesis and their stability proof along with simulation results for verification is provided in section IV, and then the final conclusion.

## II. PEM Fuel Cell Model

The block diagram in Fig. 2 describes the relationship developed in mathematical model. The model has three inputs hydrogen pressure, oxygen pressure, disturbance load current and an exogenous input i.e. room temperature whereas it has only two outputs fuel cell voltage and stack temperature. The partial pressures of hydrogen, oxygen and water vapors are calculated by mass balance and material conservation equations. Then, the open circuit potential ($V_{OFC}$) of the fuel cell is determine by the Nernst equation. The voltage losses are calculated by ohmic voltage drop equation, activation voltage drop equations and concentration voltage drop equations. The terminal (output) voltage of the fuel cell is determined by voltage losses together with the double-layer charging effect equation. The thermodynamic equations are used to calculate stack temperature.

The states description of non-linear fuel cell system model [37] is as follows:

$x_1 = \left(m_{O_2\_net}\right)_{net}$ Net flow rate of Oxygen (mol/s)

$x_2 = \left(m_{H_2\_net}\right)_{net}$ Net flow rate of Hydrogen (mol/s)

$x_3 = \left(m_{H_2O\_net}\right)_{net}$ Net flow rate of Water (mol/s)

$x_4 = T$ Stack temperature (K)

$x_5 = P_{H_2}$ Partial pressure of hydrogen (atm)

$x_6 = P_{O_2}$ Partial pressure of oxygen (atm)

$x_7 = P_{H_2O}$ Partial pressure of water (atm)

$x_8 = Q_C$ Heat generated due to electrochemical reaction (J)

$x_9 = Q_E$ Heat generated due to electricity (J)

$x_{10} = Q_L$ Heat loss by air Convection (J)

$x_{11} = V_c$ Voltage across the Capacitor (V)

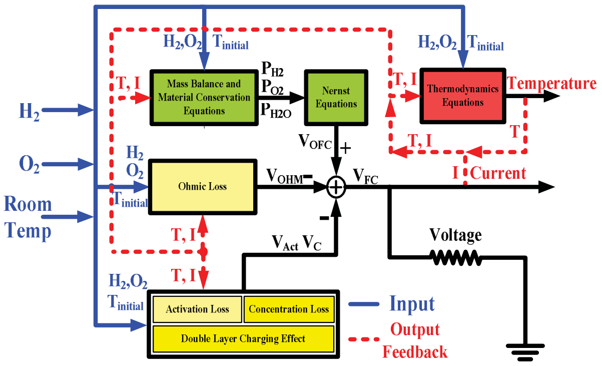


Figure 2. Block Diagram of PEM Fuel Cell Model

The V-I and P-I characteristic curves obtained in [37] of PEM fuel cell model is presented in Fig. 3.

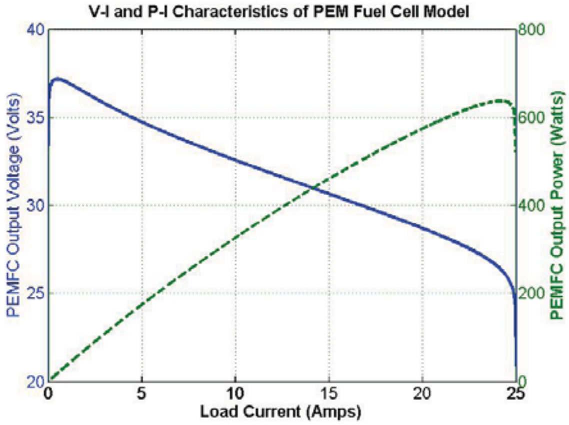


Figure 3. V-I & P-I Characteristics of PEM Fuel Cell Model

## III. Problem Formulation

The strategy contains two control loops. The first loop tries to maintain the stoichiometric oxygen excess ratio whereas the second loop stabilizes the output voltage of the system as shown in Fig. 4. The basic problem with the operation of fuel cell system is that it does not produce steady output voltage whereas most of the electrical appliances require stable one. There is a variation in the output voltage as the load varies and it drops with the passage of time even if the fuel supply is kept steady. It may conclude that to get rid of variable and degraded output voltage, there is an inevitable need of voltage stabilization.

The rapid and efficient control of air flow is required for avoiding oxygen starvation and the life of the stack [24]. The purpose of sustaining the excess ratio is to optimize the conversion of energy in the fuel cell and maximize the net power by the system operating under different load conditions, while minimizing the parasitic losses of feeding more oxygen into the cathode. If during the operation, the partial pressure of oxygen in the air stream of the cathode falls down to a certain critical level, a complicated phenomenon called oxygen starvation occurs [25].This phenomenon causes a sudden decrease in fuel cell output voltage, which can causes a hot spot or even burn the surface of a membrane in some severe situations [26].The sufficient mass flow of oxygen through the stack will satisfy the load demand. In this way not only the fuel consumption is minimized but also oxygen starvation will be avoided. The optimal mass flow of oxygen is achieved by maintaining the oxygen excess ratio to its optimal value at cathode. The stoichiometric ratio or the oxygen excess ratio is defined as [27].

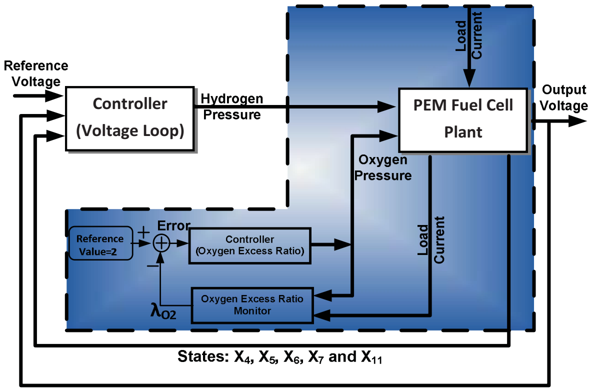


Figure 4. Control Scheme for PEM Fuel Cell System

$$\lambda_{O_2} = \frac{m_{O_2 in}}{m_{O_2 used}} \tag{1}$$

Where $m_{O_2 in}$ is the number of moles of oxygen going into the cathode, which depends on the pressure generated by the air blower and the vapors injected by the humidifier, while $m_{O_2 used}$ is the number of moles of oxygen consumed in the reaction. The total stack current depends upon this rate of oxygen consumption.

The optimum value of oxygen excess ratio is estimated by varying load to the stack on different values of stoichiometry operating as open-loop system [28]. It depends on the load condition and is normally taken as two which can be clearly seen from Fig. 5 that at higher load current, the peak power can be obtained at $\lambda_{O_2} = 2$ before and after that point, it decreases. Many fuel cell systems exhibits the same behavior with some ignorable deviations in the system operating range so $\lambda_{O_2}$ can be kept constant. Otherwise, variable $\lambda_{O_2}$ can easily be derived as a function of load current.

The advantage of the oxygen excess ratio control is to avoid oxygen starvation on the cathode side that can causes serious problems to the polymer membrane's life; the control will allow the oxygen to go below when it is needed [28, 29]. Another significance of maintaining oxygen excess ratio is that it maximizes or optimizes performance, efficiency and lifetime of the fuel cell.

On the other hand if the oxygen excess ratio crosses the optimal value then extra amount of energy is required to pump the oxygen into the cathode while the fuel cell system output power also degrades. Once the optimized stoichiometric ratio is obtained then keeping the oxygen excess ratio within optimal values by controlling the oxygen molar flow rate can be written as.

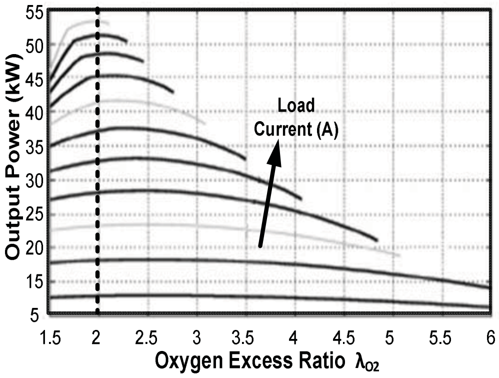


Figure 5. Analysis of oxygen excess ratio on net output power at different load conditions [28]

$$\lambda_{O_2} = \frac{4F8.614x10^{-5}U_{pc}}{n_s\, I} \cong 2 \tag{2}$$

The output voltage stabilization is carried out via five different approaches i.e. Proportional, Integral and Derivative (PID) control, first order SMC, Higher Order Sliding Mode (HOSM) control using super twisting algorithm [37], Smooth Sliding Mode Control (SSMC) [38] and Smooth Second Order Sliding Mode Control. The simulation experiments are performed on a nonlinear state space mathematical model that is validated with experimental results available in public literature.

## IV. Controller Design

### A. PID Control

To regulate the oxygen excess ratio a Proportional Differential and Integral (PID) controller is used. The reasons for selecting PID controller is that it is a simple controller and it can serve the particular task well without moving forward to advanced complicated robust controllers.

The control law for PID controller is given as:

$$U = K_p\, e + K_i \int e\, dt + K_d \frac{d}{dt} e \tag{3}$$

The PID controller is stabilizing the oxygen excess ratio at optimum value that is two in the presence of external varying

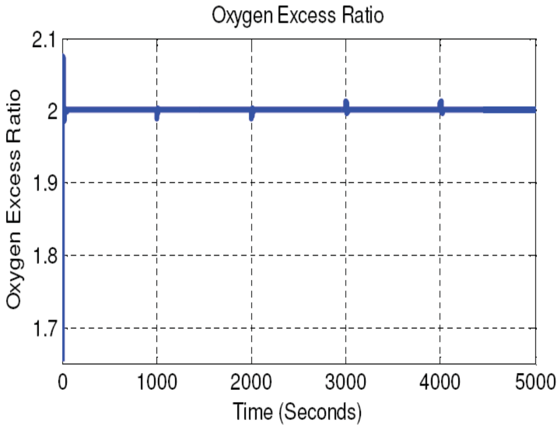


Figure 6. Controlled oxygen excess ratio

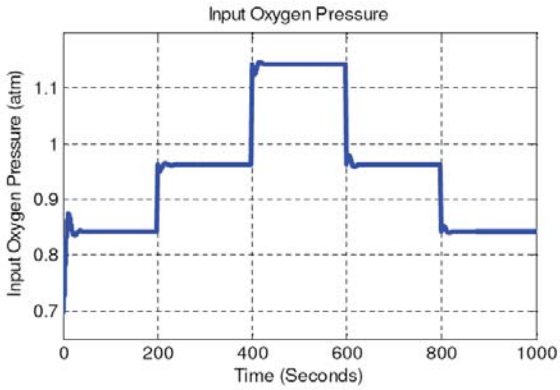


Figure 7. Input Oxygen Pressure (PID)

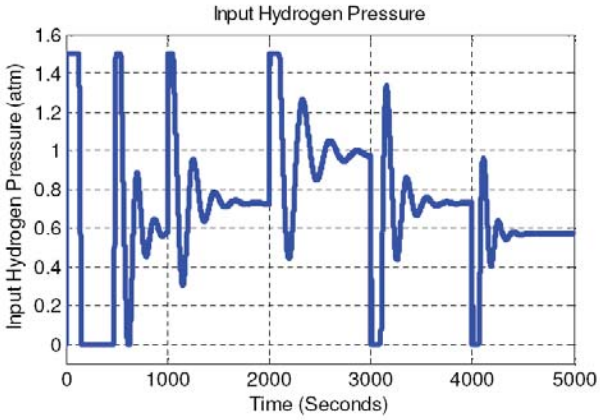


Figure 8. Input Hydrogen Pressure (PID)

disturbance. There are little over shots and under shots after each 1000 seconds interval. This is due to the step variation in the load current as in Fig. 6. This band of variation is bearable for the plant.

Controlled input oxygen pressure formed by the PID controller on oxygen excess ratio loop in order to maintain the oxygen excess ratio at 2. Oxygen excess ratio is maintained at an optimal value by an oxygen excess ratio controller as shown in Fig. 7. Oxygen excess ratio is regulating pretty well at optimal value via PID controller. A PID controller is applied on the voltage loop of the fuel cell to track the reference voltage in the presence of the oxygen excess ratio as shown in Fig. 9 and Fig. 8 represents the control effort generated by the PID controller to stabilize the fuel cell output voltages by manipulating input hydrogen pressure. The oscillations is because of the variation in load current. The settling time for the PID is high. The Fig. 10 shows error in the output voltage of the fuel cell; in ideal case error should be zero. The variations in the error are due to the load current variations on the output of fuel cell.

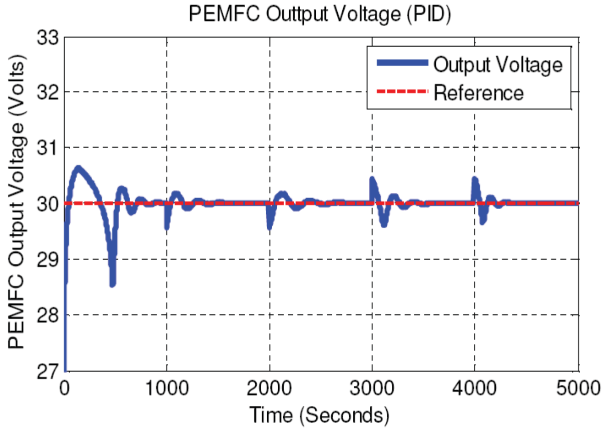


Figure 9. PEM Fuel Cell Output Voltage (PID)

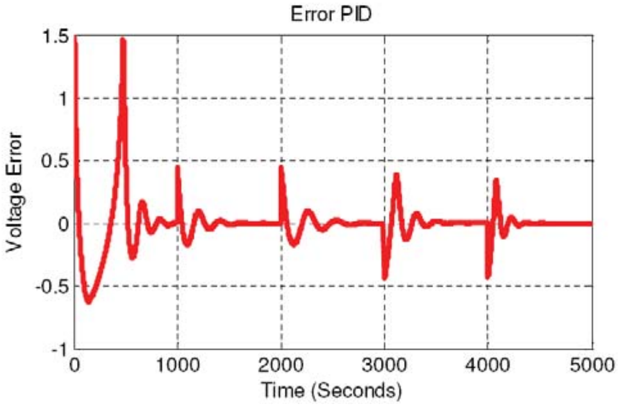


Figure 10. PID Output Voltage Error

## B. Smooth Second Oder Sliding Mode Control

A Higher-order Smooth Sliding Mode Control (HSSMC)/ Smooth Second-Order Sliding Mode Control (SSOSMC) strategy is investigated for Proton Exchange Membrane Fuel Cell (PEMFC) system to achieve the smoothness in control law. In this strategy first loop tries to maintain the stoichiometric ratio whereas the second loop stabilizes the output voltage of the system.

Smooth Second Order Sliding Mode Control is synthesized as follows. Consider SISO $\sigma$ –dynamics

$$\dot{\sigma} = g(t) + u \tag{4}$$

Where $\sigma$ is the sliding surface, the uncertainty *g(t)* is a smooth function and $u \in \Re$ is a control input. The sliding variable $\sigma$ is the system motion along the system trajectories and $\sigma = 0$ is the system motion on the sliding surface. The smoothness in the control input $u$ is the main objective of this synthesis. This synthesized control law $u$ pushes the sliding variable and its derivative to zero, i.e. $\sigma, \dot{\sigma} \rightarrow 0$ in finite time. In this technique, by using Levant exact differentiator based observer, the drift term *g(t)* is estimated and cancelled. Here it is assumed that *g(t)* is available in real-time. The sliding surface $\sigma$ in (4) can be selected as

$$\dot{x}_1 = \alpha_1 |x_1|^{(P-1)/_P} sign(x_1) + x_2$$
$$\dot{x}_2 = \alpha_2 |x_1|^{(P-2)/_P} sign(x_1);\ \sigma = x_1 \qquad (5)$$

**Lemma** Let p ≥ 2, $\alpha_1, \alpha_2 > 0$ then the system (5) is finite time stable. The settling time is a continuous function and depends on the initial condition. Also it vanishes at the origin.

The Lemma can be proved by selecting the Lyapunov candidate function as

$$V = \frac{{x_2}^2}{2} + \int_0^{x_1} \alpha_2 |z|^{(P-2)/_P} sign(z)\, dz \qquad (6)$$

and then applying LaSalle theorem [39]. It is also homogeneous with the dilation $dk{:}(x_1,\ x_2) \mapsto (k^p x_1,\ k^{p-1} x_2)$ and the degree of homogeneity is $-1$.

The term *g(t)* is unknown bounded disturbance and the sliding variable $\sigma$ is very sensitive to this disturbance. It is assumed that the terms *g(t)* and *u(t)* are available, $g^{(m-1)}(t)$ *has* a known Lipshitz constant $L$>0 *by* $m$−1 times differentiability of *g(t)*. The *u(t)* is Lebesgue-measurable input control function and $\sigma(t)$ is continuous function defined ∀$t$ >0 i.e. it is understood in the Filippov sense. Then

$$\dot{\sigma} = \dot{V}_{fc} - \dot{V}_r \qquad (7)$$

Putting values $\dot{V}_{fc}$ (see appendix B) [37].

$$\dot{\sigma} = f_1\, \dot{x}_4 + f_2 \dot{x}_5 + f_3 \dot{x}_6 - f_4 \dot{x}_7 - n_s \dot{x}_{11} + f_5 - \dot{V}_r \qquad (8)$$

Now applying $\dot{x}_5$ values and then by simplifying we get

$$\dot{\sigma} = g(t) + b\, U_{P_A} \qquad (9)$$

Where

$$g(t) = f_1\, \dot{x}_4 + f_2\, [-2\xi_1(x_4)x_5 - \xi_2(x_4)I] + f_3 \dot{x}_6 - f_4 \dot{x}_7 - n_s \dot{x}_{11} + f_5 - \dot{V}_r \qquad (10)$$

$$b = 2 f_2\, \xi_1(x_4) \qquad (11)$$

The control law $U_{P_A}$ that would maintain $\dot{\sigma} = 0$, i.e.

$$U_{P_A} = \frac{1}{b}[-g(t) + U_d] \qquad (12)$$

$$U_d = \alpha_1 |\sigma|^{2/_3} sign(\sigma_1) + \alpha_2 \int |\sigma|^{1/_3} sign(\sigma)\, d\tau \qquad (13)$$

Since the function $g(x)$ is differentiable, so it has a Lipshitz constant. As it is very difficult to estimate the exact value of the Lipshitz constant L for $\dot{g}(x) \leq L$ , it is the only parameter for proposed observer, so can be estimated through some hit and trial method. $g(x)$ is estimated using observer with the paramters m=2, $\lambda_0 = 2$, $\lambda_1 = 1.5$, $\lambda_2 = 1.1$, [40].

$$\dot{z}_0 = \nu_0 + b\, U_{P_A}$$
$$\nu_0 = -\lambda_0 L^{1/3} |z_0 - \sigma|^{2/3} sign(z_0 - \sigma) + z_1\ ; \qquad \lambda_0 = 2$$
$$\dot{z}_1 = \nu_1$$
$$\nu_1 = -\lambda_1 L^{1/2} |z_1 - \nu_0|^{1/2} sign(z_1 - \nu_0) + z_2; \quad \lambda_1 = 1.5$$
$$\dot{z}_2 = -\lambda_2 L\, sign(z_2 - \nu_1); \qquad \lambda_2 = 1.1$$
$$\hat{g}(x) = z_1 \qquad (14)$$

If there is no noise at the input, after finite time as the transients die out, we obtain $\hat{g}(x) = g(x)$ in (14). The SSOSM Control Law in terms of hydrogen input at anode $U_{P_A}$ with parameters $p = 3, m = 2$ will become

$$U_{P_A} = \frac{1}{b}\left(\alpha_1 |\sigma|^{2/3} sign(\sigma) + \alpha_2 \int |\sigma|^{1/3} sign(\sigma)\, d\tau + \hat{g}(x)\right) \qquad (15)$$

$$U_{P_A} = \frac{1}{b}\left(U_d + \hat{g}(x)\right) \qquad (16)$$

Which is the SSOSM Control with robust finite time convergent term, $U_d$.

## C. Simulation Results:

The smooth control signal generated by HSSM controller to control input hydrogen pressure is represented in Fig. 11; there is no chattering at all. This result is the main objective of HSSM controller and is better than presented in [37,38].

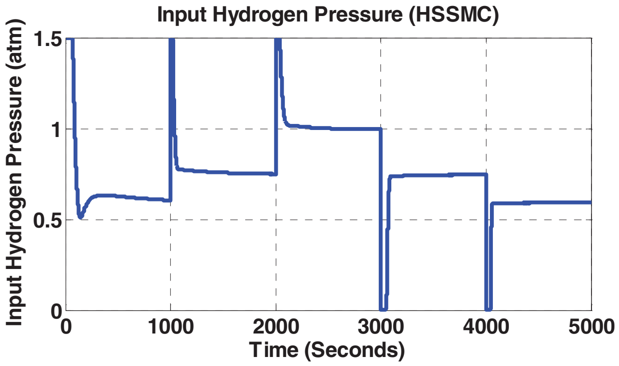


Figure 11. PEMFC Input Hydrogen Pressure HSSMC

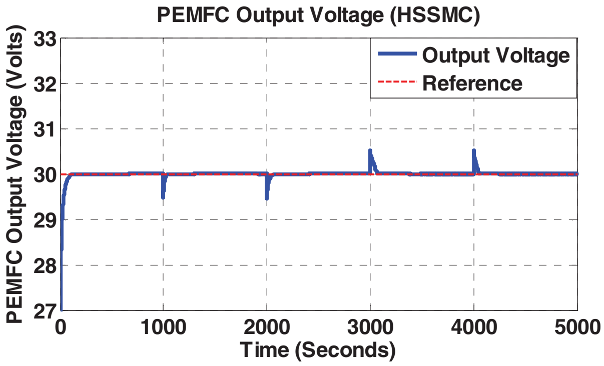


Figure 12. PEMFC Output Voltage HSSMC

The PEM fuel cell output voltage stabilized at 30 volts by HSSM controller is shown in Fig. 12. It can be seen that for two undershoots at 1000 and 2000 seconds the output voltage comes back to the reference value quickly where as in case of two overshoots at 3000 and 4000 seconds the output takes little more time to returns the reference value. The load current increases at 1000 and 2000 whereas it decreases at 3000 and 4000 seconds. When load current increases the fuel cell needs more fuel to maintain the voltage level that surplus amount can be rapidly provided by the controller where as in the case when the load current decreases the high pressure fuel stored in the gas chambers need some time to consume it up.

The sliding surface in case of HSSM controller ideally it remains at zero. The peaks are due to the load current profile given to the system. The Fig. 13 shows the robustness of the controller. The sliding surface leaves the manifold to cater the load disturbance for a few seconds and remains there forever.

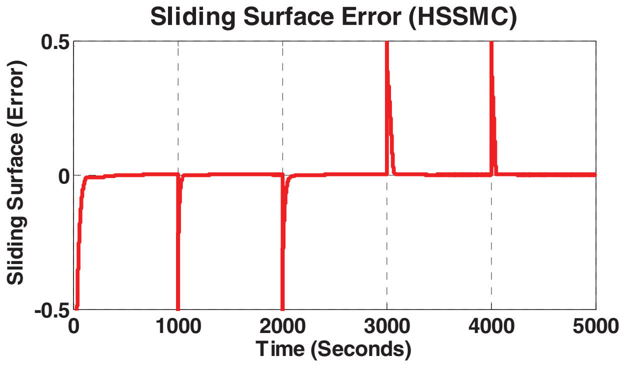


Figure 13. Sliding Surface HSSMC

## Conclusion

A novel chattering free smooth second order sliding mode control (SSOSMC)/HSSMC is studied to tackle the problem of PEM fuel cell system output voltage stabilization problems along with controlling oxygen excess ratio. The oxygen excess ratio is controlled in the inner control loop using PID while the regulation of output voltage is being treated in the outer loop through HSSM control. The control law is designed for the formulated problem and stability analysis is provided. The simulation is carried out on the nonlinear, validated model. Finally, it is shown from the simulation that HSSM controller provides the smooth/ chattering free control which is robust against uncertain disturbances and minimize fuel consumption by showing minimal control efforts. The proposed controller is far better in avoiding chattering and easy for implementation than the previously applied different techniques for the formulated scheme.

## REFERENCES

[1] Larminie James, Andrew Dicks, Fuel Cell Systems Explained. John Wiley and Sons, UK, 2000.

[2] G. Pede, A. Iacobazzi, S. Passerini, A. Bobbio, and G. Botto, “FC vehicle hybridisation: An affordable solution for an energy-efficient FC powered drive train,” Journal of Power Sources, 2004.

[3] G. Hoogers, ed., Fuel Cell Technology Handbook. CRC Press, 2003.

[4] M. Ehsani, Y. Gao, S.E Gays, and A. Emadi, Modern Electric, Hybrid Electric and Fuel Cell Vehicle—Fundamentals, Theory and Design. Boca Raton: CRC press, 2005.

[5] A. Emadi, M. Ehsani, and J. Miller, Vehicular Electric Power Systems: Land, Sea, Air and Space Vehicles. New York: Marcel Dekker, 2004.

[6] Ali Keyhani and Sachin Puranik, “Dynamic Modeling Of Proton Exchange Membrane Fuel Cell”, Green Energy Systems Laboratory, Ohio State University 2009.

[7] F.Khorrami, S.Puranik, A. Keyhani, P. Krishnamurthy and Y.She , “PEM Fuel Cell Distributed Generation System: Modeling and Robust Nonlinear Control” , IEEE Conference on Decision and Control, December 2009.

[8] T. E. Springer, T. A. Zawodzinski, and S. Gottesfeld, "Polymer Electrolyte Fuel Cell Model", Journal of The Electrochemical Society, August 1991.

[9] J. C. Amphlett, R. M. Baumert, R.F. Mann, B.A. Peppley, P. R. Roberge, and Harris T.J, "Performance Modeling of the Ballard Mark IV Solid Polymer Electrolyte Fuel Cell II. Empirical Model Development", Journal of The Electrochemical Society, January 1995.

[10] Tanrioven, M.; Alam, M. S., “Modeling, Control, and Power Quality Evaluation of a PEM Fuel Cell-Based Power Supply System for Residential Use”, IEEE Trans.Industry Applications, Nov-Dec 2006.

[11] Lu-Ying Chiu, Randall S. Gemmen and Bill Diong, “An improved small-signal model of the dynamic behavior of PEM fuel cells, “IEEE 38th IAS Annual Meeting, Oct. 2003.

[12] M. Uzunoglu, M.S. Alam, “Dynamic modeling, design, and simulation of a combined PEM fuel cell and ultracapacitor system for stand-alone residential applications”, IEEE Trans. Energy Conversion, Sept 2006.

[13] N. P. Siegel, M. W. Ellis, D. J. Nelson, and M. R. von Spakovsky, “A two-dimensional computational model of a PEMFC with liquid water transport,” Journal of Power Sources,2004.

[14] S. Um and C. Y. Wang, “Three-dimensional analysis of transport and electrochemical reactions in polymer electrolyte fuel cells,” Journal of Power Sources., 2004.

[15] B. R. Sivertsen and N. Djilali, “CFD-based modelling of proton exchange membrane fuel cells,” Journal of Power Sources, 2005.

[16] J. Kim, S. Lee, and S.Srinivasan, “Modeling of proton exchange membrane fuel cell performance with an empirical equation,” J. Electrochem. Soc., 1995.

[17] J. H. Lee, T. R. Lalk, and A. J. Appleby, “Modeling electrochemical performance in large scale proton exchange membrane fuel cell stacks,” Journal of Power Sources, 1998.

[18] R. Mann, J. Amphlett, M. Hooper, H. Jensen, B. Peppley, and P. Roberge, “Development and application of a generalised steadystate electrochemical model for a PEM fuel cell,” J. of Power Sources, 2000.

[19] P. H. Rodatz, Dynamics of the Polymer Electrolyte Fuel Cell: Experiments and Model-Based Analysis. PhD thesis, Eidgenössische Technische Hochschule Zürich, 2004.

[20] Avista Laboratories, Inc. SR-12 Modular PEM Generator, Operator’s Manual, Spokane, WA, Jul. 2000.

[21] Caisheng Wang, M. Hashem Nehrir, and Steven R. Shaw, “Dynamic Models and Model Validation for PEM Fuel Cells Using Electrical Circuits”, IEEE Trans.Energy Conversion, June 2005.

[22] National Energy Technology laboratory, Fuel Cell Hand Book 6th edition, November 2002.

[23] Sachin V. Puranik, Ali Keyhani and Farshad Khorrami, “Neural Network Modeling of Proton Exchnage Membrane Fuel Cell”, IEEE Transactions on Energy Conversion, 2010.

[24] W.-C. Yang, B. Bates, N. Fletcher, and R. Pow, “Control challenges and methodologies in fuel cell vehicle development,” SAE Paper 98C054.

[25] T.E. Springer, R. Rockward, T.A. Zawodzinski, and S. Gottesfeld, “Model for polymer electrolyte fuel cell operation on reformate feed,”J. Electrochem. Soc., vol. 148, no. 1, pp. A11–A23, 2001.

[26] Jay T. Pukrushpan, Anna G. Stefanopoulou and Huei Peng “Control of Fuel Cell Breathing” IEEE Control Systems Magazine, April 2004.

[27] Yiyao A. Chang and Scott J. Moura "Air Flow Control in Fuel Cell Systems: An Extremum Seeking Approach", American Control Conference, 2009.

[28] Cristian Kunusch, Paul F. Puleston and Jordi Riera “Sliding Mode Strategy for PEM fuel Cells Stacks Breathing Control Using a Super-Twisting Algorithm”, IEEE Transactions on Control System Technology, January 2009.

[29] Carlos Andres Ramos-Paja,Carlos Bordons, Alfonso Romero and Luis Martínez-Salamero "Minimum Fuel Consumption Strategy for PEM Fuel Cells", IEEE Transactions on Industrial Electronics, Vol. 56, No. 3, March 2009 James Larminie and Anderw Dicks, Fuel Cell Systems Explained, 2nd edition. New York: Wiley, 2003.

[30] Rolands S.Burns "Advanced Control Engineering", First published 2001.

[31] Karl J. Astrom and Tore Hagglund "PID Controllers: Theory, Design and Tunning", 2nd Edition, 1995.

[32] Slotine Jean-Jacques E., and Li Weiping, Applied Nonlinear Control. Prentice Hall,1991.

[33] Levant A. Higher-order sliding modes, differentiation and output feedback control International Journal of Control 2003.

[34] Levant A. Universal SISO sliding mode controllers with finite-time convergence. IEEE Transactions on Automatic Control 1998.

[35] S.V. Emelyanove, S.K. Korovin and A. Levant, ‘Higher-order sliding modes in control systems, Differential Equations, 29, pp. 1627–1647, 1993.

[36] Wilfrid Perruquetti, Sliding Mode Control in Engineering, CRC Press, 1st edition 2002.

[37] S. Saqib H. Rizvi, A. I Bhatti, Qudrat Khan, Qadeer Ahmad and Asad Hameed “HOSM Control Design of PEM Fuel Cell using Super Twisting Algorithm”, 9th IBCAST IEEE , January 2012.

[38] S. Saqib H. Rizvi, A. I Bhatti, Qudrat Khan, Qadeer Ahmad, Asad Hameed and Zahid Butt, “Smooth Sliding Mode Control for PEM Fuel Cell System”, 24th Chinese Control and Decision Conference, IEEE, May 2012.

[39] Y.B. Shtessel, I. A. Shkolnikov and A. Levant, “Smooth second-order sliding modes: Missile guidance application”, 1470-1476, 43 Automatica, 2007.

[40] J. Davila, L. Fridman & A. Levant, “Second-order sliding-mode observer for mechanical systems”, 1785–1789, *50*(11), *IEEE Transactions of Automatic Control*, 2005.

.

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# Rapid Alignment Method of INS with Large Initial Azimuth Uncertainty under Complex Dynamic Disturbances

Xin Liu
School of Automation
Beijing Institute of Technology
Beijing 100081,China
Email: xiaoyuanxiao163@163.com

Bo Wang
School of Automation
Beijing Institute of Technology
Email: wb1020@bit.edu.cn

Zhihong Deng
and Shunting Wang
Beijing Institute of Technology
Beijing 100081, China
Telephone: (010) 68914350
Fax: (010) 68914382

*Abstract*—**For the rapid alignment of the ship-borne weapon INS with large initial azimuth attitude error under complex environment disturbances, a nonlinear error propagation model augmented by sensor errors and disturbance sources was proposed. Velocity plus angular rate matching method was applied in the implementation of the alignment. Simulation results show that comparing with the conventional solutions, this method can accomplish the transfer alignment of a mooring ship's weapon INS with large heading error rapidly and accurately. Meanwhile, it has strong adaptability to the sea condition and can improve the precision of the alignment under complex environment disturbances.**

## I. Introduction

Aligning SINS (slave inertial navigation system) of lower accuracy by MINS (master inertial navigation system) is called transfer alignment (TA). TA method can align the SINS with short time and high accuracy. Conventional TA error propagation model based on the hypothesis of small misalignment angle between SINS and MINS is linear; however, there might be large heading error between the frames of the two INSs in ship-borne TA process[1],[2]. Thus, it is necessary to study the methodology of TA with large azimuth misalignment.

In the works [3]and [4], velocity matching was used to execute the TA with large heading misalignment angle of SINS, but the long alignment time cannot meet the operational requirements. In the research [5] , velocity plus angular rate matching TA method estimated attitude errors accurately and rapidly only by virtue of swaying movement caused by sea waves, when the misalignment angles are small. Meanwhile, it can give good estimations of gyro-drifts. However, there are few researches on this matching method under the condition of large azimuth attitude error between SINS and MINS.

There are some special problems for the transfer alignment between the ship-borne weapon INS and the Master INS. On the one hand, it is inappropriate for large warship to execute maneuver in order to accomplish the alignment. On the other hand, ship-borne TA is badly affected by environment disturbance. Based on the above considerations, a new approach of the TA for ship-borne weapon with large azimuth error was studied in this paper. This alignment method used the differences of velocity and angular rates between SINS and MINS as the measurements. Meanwhile, dynamic and static lever arm errors and flexure deformation are augmented to system model of the TA. The alignment method was executed by UKF. Performance of this approach was tested under two typical sea conditions by the simulation.

## II. Error propagation Model of the Transfer Alignment

### A. State-Equations of Transfer Alignment

According to [6], the error propagation equation with the large azimuth misalignment is

$$\dot{\phi} = (I - C_n^{n'})\omega_{in}^n - C_b^{n'}\varepsilon + \eta_\phi \tag{1}$$

In the research [8], lever arm effects were compensated in the velocity error propagation equation, it needs angular acceleration information to calculate compensation item, and however, the angular acceleration cannot be obtained directly. Difference calculation is needed in order to get angular acceleration, but it will bring extra alignment error. Therefore, in this paper, instead of compensating lever-arm effects in velocity error propagation equation, those errors were compensated in the velocity measurement. Thus, the velocity error propagation equation is

$$\delta\dot{v}_s^n = [I-(C_n^{n'})^T]C_b^{n'}f^b - (2\omega_{ie}^n+\omega_{en}^n)\times\delta v_s^n + C_b^{n'}\bigtriangledown^b + \eta_v \tag{2}$$

In (1)and (2), $\phi$ is the attitude misalign vector between SINS′s and MINS′s navigation frames, $\delta v_s^n$ is the velocity error vector of SINS, $\eta_\phi$and $\eta_v$ are the noise vectors of attitude and velocity, $C_n^{n'}$ is the transform matrix from the true navigation frame($n$) to the SINS′s navigation frame($n'$), $C_b^{n'}$is the body-axes-to-navigation direction cosine matrix of SINS. When the azimuth attitude error is large, $C_{n'}^n$ is given by the following form

$$C_{n'}^n = \begin{bmatrix} cos\phi_z & -sin\phi_z & \phi_y cos\phi_z + \phi_x sin\phi_z \\ sin\phi_z & cos\phi_z & \phi_y sin\phi_z - \phi_x cos\phi_z \\ -\phi_y & \phi_x & 1 \end{bmatrix} \tag{3}$$

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

The flexure deformation, which is generated by sea waves, winds and engine vibration, can be approximated by a second-order Markov process [6] [7]

$$\ddot{\theta}_{fi} + 2\beta_i \dot{\theta}_{fi} + \beta_i^2 \theta = \eta_i \quad (i = x, y, z) \tag{4}$$

Where, $\theta = [\theta_x, \theta_y, \theta_z]^T$ is the flexure deformation vector, and its variance is $\sigma = [\sigma_x, \sigma_y, \sigma_z]^T$, $\eta = [\eta_x, \eta_y, \eta_z]^T$ is a white noise process with strength given by the diagonal matrix $Q_n$. $\beta = [\beta_x, \beta_y, \beta_z]^T$ is a constant vector. The relationship between $Q_{ni}$, $\sigma_i$and $\beta_i$ is $Q_{ni} = 4\beta_i^3 \sigma_i^2 \quad (i = x, y, z)$ ,and the correlation time $\tau_i$ of the stochastic process is related with $\beta_i$ by $\beta_i = 2.146/\tau_i \quad (i = x, y, z)$ .

For the time of the TA is short, the random errors can be treated as white noises besides the zero-biases of the gyros and accelerometers. Thus, the models of gyro and accelerometer can be approximated as

$$\varepsilon = \varepsilon_b + \varepsilon_w \tag{5}$$

$$\bigtriangledown = \bigtriangledown_b + \bigtriangledown_w \tag{6}$$

where, $\varepsilon_b$ is constant zero-drift, $\varepsilon_w$ is Gaussian white noise. $\bigtriangledown_b$ is constant zero-bias, $\bigtriangledown_w$ is Gaussian white noise.

The measurement error vector of static lever arm is constant, thus $\delta\dot{r}_{meas}^T = 0$ . According to the investigation [8], since the deformation angle is small, the dynamic lever arm length $(\delta r_f)$, which is caused by the deformation, can be described as $\delta r_f = R\theta_f$, thus

$$\delta\dot{r}_f = R\dot{\theta}_f = R\omega_f \tag{7}$$

where,

$$H = \begin{bmatrix} 0 & z_0 & 0 \\ 0 & 0 & x_0 \\ y_0 & 0 & 0 \end{bmatrix} \tag{8}$$

Comparing with the conventional TA model of large azimuth misalign, in order to estimate those environment disturbances in the alignment process, this augmented model is established.

### *B. Measurement Equations*

The measurements of this TA method are velocity and angular rate differences of of SINS and MINS.

As discussed above, the lever-arm effects should be compensated in the velocity matching measurement, thus

$$\Delta v = v_s^n - v_m^n = \delta v_s^n + v_r^n \tag{9}$$

The compensation for velocity measurement is governed by (10), if the measurement of the lever-arm is accurate

$$v_r^n = C_m^n(\omega_{im}^m \times r_0^m) + C_m^n \dot{r}_0^m \tag{10}$$

where, $r_0^m$ is the deterministic lever-arm in the body-frame of MINS, which can be measured beforehand, and $\dot{r}_0^m = 0$ in this situation,$\times$ stands for the skew matrix of the vector.

However, there are some uncertain errors, which cannot be measured accurately. These errors are the static lever-arm measurement error $(\delta r_{meas})$ and the dynamic lever-arm error $(\delta r_f)$ caused by deformation. Thus, the real lever-arm is as following

$$r = r_0 + \delta r_{meas} + \delta r_f \tag{11}$$

Accordingly, the compensation item of the velocity measurement is

$$v_r^n = C_m^n[\omega_{im}^m \times (r_0^m + \delta r_{meas}^m + \delta r_f^m)] + C_m^n \delta\dot{r}_f \tag{12}$$

where, $\omega_{im}^m \times$ is the skew matrix of vector $\omega_{im}^m$.

After compensating the deterministic lever-arm effect by (10), the velocity matching measurement is still affected by $\delta r_{meas}$ and $\delta r_f$, and it is governed by (13)

$$\begin{aligned} \Delta v &= \delta v_s^n + C_m^n[\omega_{im}^m \times (\delta r_{meas}^m + \delta r_f^m)] + C_m^n \delta\dot{r}_f \\ &= \delta v_s^n + C_m^n[\omega_{im}^m \times (\delta r_{meas}^m + \delta r_f^m)] + C_m^n R\omega_f \end{aligned} \tag{13}$$

The angular rate measurement is the difference of angular rates of SINS and MINS in each navigation frame, therefore

$$\Delta\omega = \omega_{ibs}^{n'} - \omega_{ibm}^n = C_{bs}^{n'} \omega_{ibs}^{bs} - C_{bm}^n \omega_{ibm}^{bm} \tag{14}$$

where, $bm$ and $bs$ are the body frames of MINS and SINS respectively. $\omega_{ibs}^{bs}$ and $\omega_{ibm}^{bm}$are the outputs of angular rates of SINS and MINS respectively. The relationship of $\omega_{ibs}^{bs}$ and $\omega_{ibm}^{bm}$ is

$$\omega_{ibs}^{bs} = C_{bm}^{bs}(\omega_{ibm}^{bm} + \omega_f^{bm}) + \varepsilon \tag{15}$$

Substituting (15) into (14), then the

$$\begin{aligned} \Delta\omega &= C_{bm}^{n'} \omega_{ibm}^{bm} + C_{bs}^{n'} \omega_f^{bs} + C_{bs}^{n'} \varepsilon - C_{bm}^n \omega_{ibm}^{bm} \\ &= C_n^{n'} \omega_{ibm}^n - \omega_{ibm}^n + C_{bs}^{n'} \omega_f^{bs} + C_{bs}^{n'} \varepsilon \\ &= (C_n^{n'} - I)\omega_{ibm}^n + C_{bs}^{n'} \omega_f^{bs} + C_{bs}^{n'} \varepsilon \end{aligned} \tag{16}$$

Comparing with the traditional TA method, the circumstance disturbances are augmented to this TA model, and these errors can also be reflected in the measurement equations.

## III. Simulation verification of the TA method

Augmenting the deformation and lever-arm errors to the TA model, the states of the TA filter are as follows

$$X = \begin{bmatrix} \phi^T & \delta v^T & \varepsilon^T & \bigtriangledown^T & \theta_f^T & \omega_f^T & \delta r_{meas}^T & \delta r_f^T \end{bmatrix}^T \tag{17}$$

Summarizing the TA filter model, the state vector is defined as (17), and the corresponding state equation consists of the attitude error (1), velocity error (2), the disturbances (4) and (7) ,and constant zero-bias of IMU (5)(6).

Adding velocity measurement noise and angular rate measurement noise to (13) and (16), the measurement equations of this TA filter are

$$\begin{cases} Z_{\Delta v} = \Delta v + \lambda_{\Delta v} \\ Z_{\Delta\omega} = \Delta\omega + \lambda_{\Delta\omega} \end{cases} \tag{18}$$

where, $\lambda_{\Delta v}$ and $\lambda_{\Delta\omega}$ are the measurement noises of velocity matching and angular rate matching respectively.

The implementation of this TA method is by UKF [4], [10].

The ship is mooring in the TA process. Under the waves excitation, the ship sways around the pitch axis, the roll axis

TABLE I
PARAMETER SETTINGS OF FLEXURE DEFORMATION

| Sea condition | Parameter | | | |
|---|---|---|---|---|
| | Variable | Roll | Pitch | Yaw |
| Medium sea | Amplitude | $1'$ | $15'$ | $3'$ |
| | Correlation time | 60s | 60s | 60s |
| Peaceful sea | Amplitude | $1'$ | $6'$ | $2'$ |
| | Correlation time | 60s | 60s | 60s |

TABLE II
ERRORS OF THE SINSS IMUS

| Sensor error | Accelerometer | | Gyro | |
|---|---|---|---|---|
| | Constant zero-bias | Random walk coefficient | Constant zero-drift | Random walk coefficient |
| Degree1 | 100 $\mu$g | 10$\mu$g/$\sqrt{h}$ | 0.01°/h | 0.001°/$\sqrt{h}$ |
| Degree2 | 400$\mu$g | 40 $\mu$g/$\sqrt{h}$ | 1°/h | 0.05°/$\sqrt{h}$ |

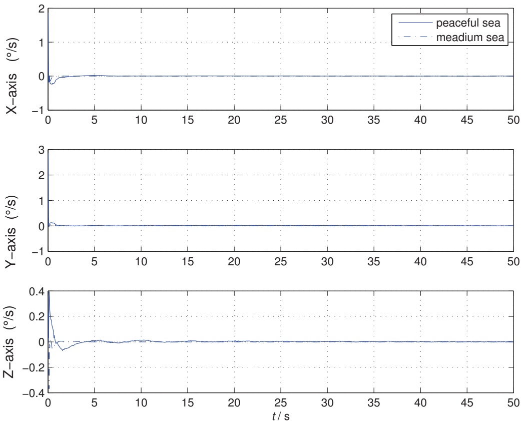


Fig. 1. Estimation bias of flexure deformation angular rate.

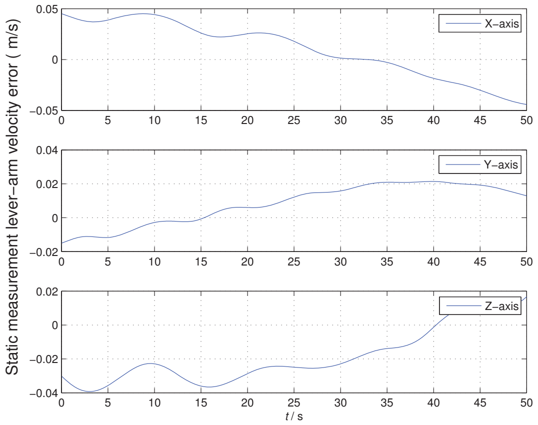


Fig. 2. Velocity error caused by the static lever-arm measurement error.

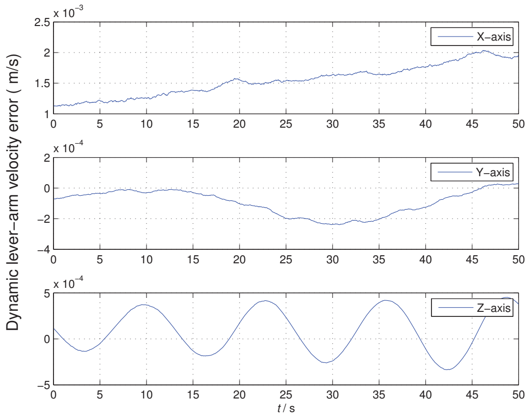


Fig. 3. Velocity error caused by the dynamic lever-arm error.

and the heading axis separately for sinusoidal motion. The model is given as follows

$$\begin{cases} \psi = \psi_m \sin(\omega_p t + \varphi_p) \\ \gamma = \gamma_m \sin(\omega_r t + \varphi_r) \\ H = H_m \sin(\omega_h t + \varphi_h) \end{cases} \tag{19}$$

where, $\psi_m$ ,$\gamma_m$ and $H_m$ are the amplitudes of the sway. $\omega_i = 2\pi/T_i(i = p, r, h)$ ,and $T_i$ is the period of sways. $\varphi_i(i = p, r, h)$ is the initial phase, and $\varphi_i(t)$ is a random value in $(-2\pi, 2\pi)$. Parameters of the above model are determined by the sea condition, two sea conditions are studied in this paper, one is medium sea condition, and another is peaceful sea condition, the parameters of these two conditions are

For medium sea condition,
$\psi_m = 2.5°$ ,$\gamma_m = 12°$ ,$H_m = 4°$ ;
$T_p = 8s$,$T_r = 13s$ ,$T_h = 40s$ .
For peaceful sea condition,
$\psi_m = 0.5°$ ,$\gamma_m = 2°$ ,$H_m = 2°$ ;
$T_p = 5s$,$T_r = 10s$ ,$T_h = 60s$

The parameter settings of the flexure deformation are shown in table I. The TA method is also tested by two different accuracies (see table II) of the SINSs IMUs

This TA filter can estimate flexure deformation angular rate accurately, the results are shown in Figure 1.

The disturbance of the lever-arm effect becomes bigger as the strength of the ships sway increases. Thus, in the following simulation, the influence of the lever-arm error is analyzed in the medium sea condition. where, let $\delta r^m_{meas} = [0.5\text{m}, 0.5\text{m}, 0.5\text{m}]^T$ . The velocity errors caused by $\delta r_{meas}$ and $\delta r_f$ are shown in Figure2 and Figure3.

From Figure2 to Figure5, we can see that the velocity error caused by $\delta r_{meas}$ and $\delta r_f$ influenced the velocity measurement in certain degree, and this disturbance is even more harmful to SINS of high sensor accuracy (such as a shipboard aircraft). Thus, in order to get useful velocity measurement for the TA filter, these errors must be compensated. This TA method can estimate $\delta r_{meas}$ and $\delta r_f$ accurately, the estimation errors of them are shown in Figure6 and Figure7.

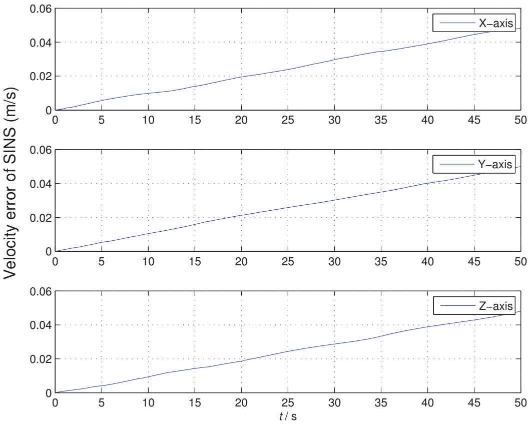


Fig. 4. Velocity error of SINS with IMU error degree 1.

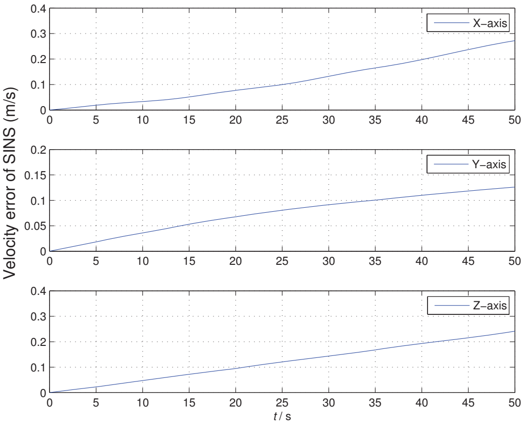


Fig. 5. Velocity error of SINS with IMU error degree 2.

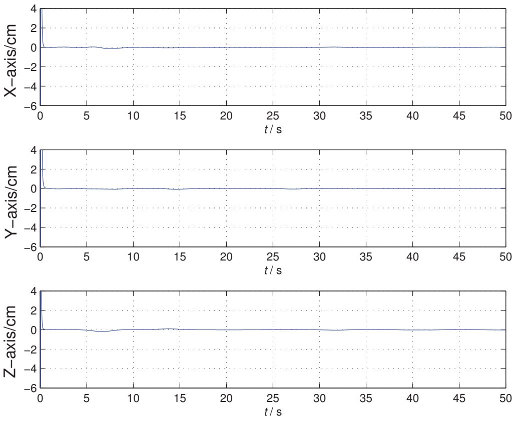


Fig. 6. Estimation bias of the static lever-arm measurement error.

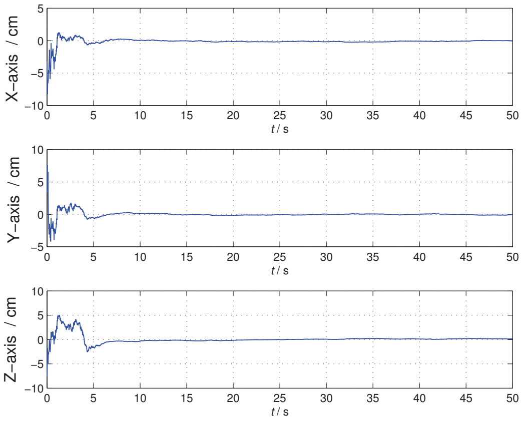


Fig. 7. Estimation bias of the dynamic lever-arm error.

TABLE III
TA RESULTS OF DIFFERENT COMPENSATION SITUATIONS

| Estimation error | Different compensation situations | | | |
|---|---|---|---|---|
| | Situation1 | Situation2 | Situation3 | Situation4 |
| $\phi_x$ | $9.9701'$ | $9.4511'$ | $5.1821'$ | $4.0645'$ |
| $\phi_y$ | $16.5467'$ | $15.9928'$ | $4.3780'$ | $3.9019'$ |
| $\phi_z$ | $13.7641'$ | $12.7643'$ | $6.4335'$ | $5.0075'$ |

The static lever-arm measurement error of a shipboard aircraft can be more indeterminate, thus, let $\delta r^m_{meas} = [1.5\text{m}, 1.5\text{m}, 1.5\text{m}]^T$ in the following simulation. Table III shows the TA results of partial or all disturbance compensations. In situation 1, there are no compensations of both lever-arm error and deformation. In situation 2, there is only lever-arm error compensation. In situation 3, there is only deformation compensation. In situation 4, all the disturbances are compensated by the method of this paper.

The simulation results shows that both flexure deformation and lever-arm error affect the accuracy of the TA in some degree, see Table III, whereas the TA method in this paper can compensate all these disturbances very well, and the alignment accuracy can be improved a lot. From the alignment results in the four situations of different sea condition and sensor error , as shown in Figure8 through Figure11, we can get that this TA method has strong adaptability to the sea condition, and the degree of sensor error does not affect the TA results too much (see table IV).

The above results are acquired in the situation that the measurement noises of actual value and the TA filter are same.

TABLE IV
ESTIMATION BIASES OF MISALIGNMENT ANGLES OF THIS TA METHOD

| Sea condition | Sensor error degree | $\phi_x$ | $\phi_y$ | $\phi_z$ |
|---|---|---|---|---|
| Medium sea | Degree 1 | $4.0645'$ | $3.9019'$ | $5.0075'$ |
| | Degree 2 | $4.5735'$ | $4.3017'$ | $5.5971'$ |
| Peaceful sea | Degree 1 | $4.8917'$ | $4.0133'$ | $5.4335'$ |
| | Degree 2 | $5.0103'$ | $4.3855'$ | $5.8011'$ |

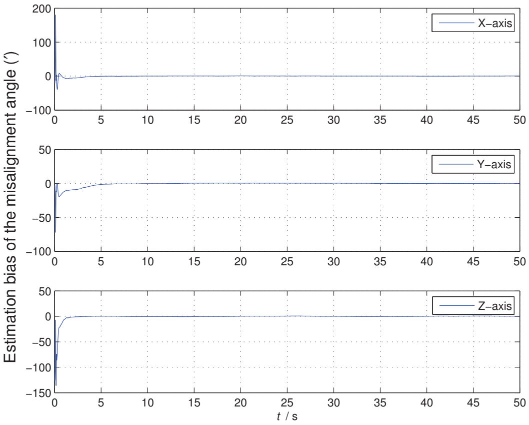


Fig. 8. Estimation bias of the misalignment angle, in the condition of medium sea and sensor error degree 1.

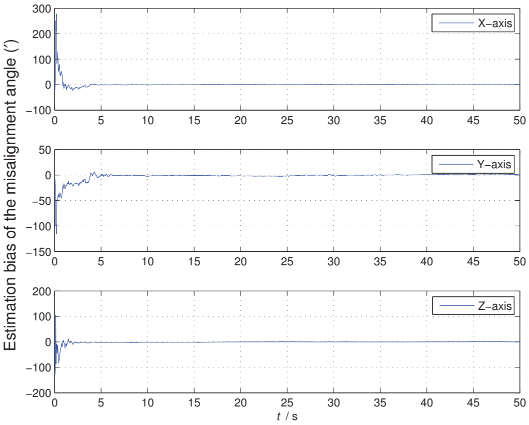


Fig. 9. Estimation bias of the misalignment angle, in the condition of medium sea and sensor error degree2 .

From(16), we can see that deformation angular rate is a main disturbance of angular rate matching measurement. In the TA method of this paper, the flexure deformation disturbances are augmented to the states of the TA filter, the measurement disturbance could be compensated to some extent in the process of alignment filtering. Thus, the actual measurement noise will just change in a small range. Assuming that the actual measurement noise is two times larger than the TA filter's. The comparison simulations are carried out in the condition of peaceful sea and sensor error degree 1, and the results are shown in Figure 12. The estimation bias of the misalignment angle is $\delta\phi = [5.5011', 4.7566', 5.6179']^T$.

The simulation results show that the bigger actual measurement noise can cause lower accuracy of the misalignment angle estimation. Meanwhile, the research [11] has shown that the adaptive Kalman filter could be one solution of this problem.

Besides, time-delay is another influencing factor of TA.

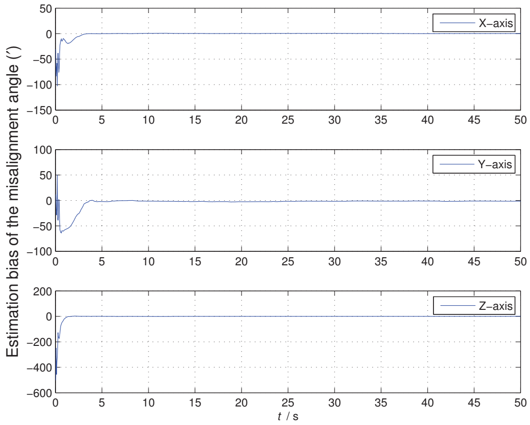


Fig. 10. Estimation bias of the misalignment angle, in the condition of peaceful sea and sensor error degree 1.

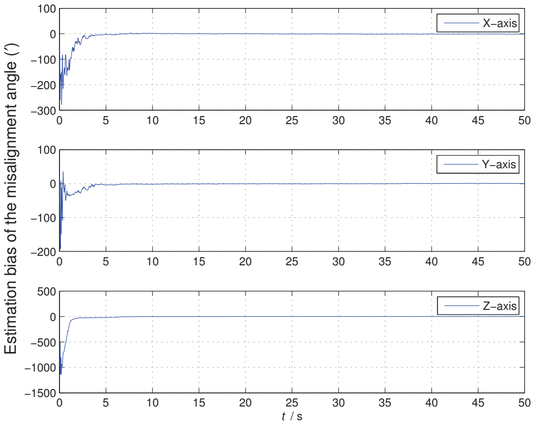


Fig. 11. Estimation bias of the misalignment angle, in the condition of peaceful sea and sensor error degree 2.

According to the research [12] , time delay may cause some misalignment between the MINS and SINS, which can reduce the accuracy of the TA. Recent research [12] has shown that the fire control computer of the MINS can give the time-delay information to the SINS, then the data can be processed by Kalman filtering and the time origin is unified to the moment corresponding to the MINS data. Therefore, the time-delay barely affects the TA process in this situation.

## IV. Conclusion

This transfer alignment method is for the purpose of reducing alignment errors induced by the disturbances and aligning the weapon INS with large azimuth attitude error on a mooring ship. The method has been established through the disturbances state augmentation and velocity plus angular rate matching, and via qualitative explanations and interpreting computer simulation results, it turns out to be that comparing with the conventional TA method the present one is more

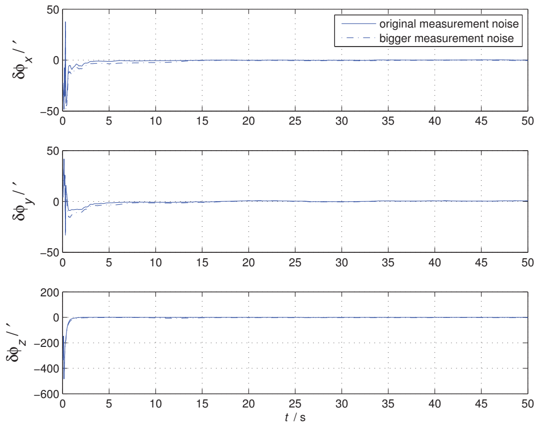


Fig. 12. Comparison of estimation bias of the misalignment angle under two different measurement noise situations

suitable for weapon INS with large azimuth attitude error of a mooring ship under complex environment disturbances.

A related further research of TA with large azimuth attitude error should develop an adaptive state estimation method in face of uncertain measurement noise and an error compensation scheme based on uncertain time-delay.

## Acknowledgment

This work is supported by the grant from National Program on Key Basic Research Project (973 Program) (No.613121) and National Natural Science Foundation of China (No. 61104189).

## References

[1] H.S. Hong, J. G. Lee, "In-flight alignment of SDINS under large initial heading error," *AIAA Guidance, Navigation, and Control Conference and Exhibit*, pp.1-6, Aug. 2001.

[2] S.P. Dmitriyev, O.A. Stepanov, S.V. Shepel, "Nonlinear filtering methods application in INS alignment," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 33, no. 1 pp. 260-271, January 1997.

[3] G.R. Zhao, Q.W. Gao, X.B Wang, "Transfer Alignment of Carrier-Based Aircraft Inertial Navigation System," *Journal of Naval Aeronautical and Astronautical University*, vol. 24, no. 1, pp. 39-42, Jan. 2009.

[4] G.M. Yan, W.S. Yan, D.M. Xu, "Application of simplified UKF in SINS initial alignment for large misalignment angles," *Journal of Chinese Inertial Technology*, vol.16, no.3, pp. 253-264, Jun. 2008

[5] R.M. Rogers, "Velocity-plus-rate matching for improved tactical weapon rapid transfer alignment," *AIAA Guidance, Navigation and Control Conference*, pp.1580-1588, Aug. 1991.

[6] Y.X. Xiao,H.Y. Zhang, "Study on Transfer Alignment with the Wing Flexure of Aircraft," *Journal of Aerospace Control*, no.2, pp. 27-35, 2001.

[7] J. Hu, B.L. Zhou, X.H. Cheng, "Comparison on Two Methods of Transfer Alignment with Carrier Flexure," *Journal of Chinese Inertial Technology*, vol.13, no.4, pp. 15-18, Aug. 2005.

[8] Q.W. Gao ,G.R. Zhao, X.B. Wang, "Incorporate Modeling and Simulation of Transfer Alignment with Flexure of Carrier and Lever-arm Effect," *Journal of Acta Aeronautica Et Astronautica Sinica*, vol.30, no.11, pp.2172-2177, Nov. 2009.

[9] D.H. Titterton, J.L. Weston, "The alignment of ship launched missile IN systems," *IEE Colloquium on Inertial Navigation Sensor Development*, pp.1-16, Jan. 1990.

[10] S.J. Julier, J.K. Uhlmann, H.F. Durrant-Whyte, "A new approach for filtering nonlinear systems, "*Proceedings of the American Control Conference*, Seattle, WA,USA, 1995, pp. 1628-1632.

[11] C. M. Xie, Y. Zhao, J. Y. Deng. "Improved adaptive square root filtering algorithm for transfer alignment," *Journal of Systems Engineering and Electronics*, Vol. 33 ,No. 3, pp. 622-626,Mar. 2011.

[12] J. H. Xia, Y. Y. Qin, C. S. Zhao. "Study on the Master Inertial Navigation System Reference Data Delay Processing Method in Transfer Alignment," *Journal of Acta Armamentarii*, Vol. 30 No. 3, pp.342-345, Mar. 2009.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Study on The Method of Multi-agent Generation Algorithm within Special Artificial Society Scene

Zongchen Fan, Wei Duan, Bin Chen, Yuanzheng Ge, Xiaogang Qiu

School of Mechatronics Engineering and Automation
National University of Defense Technology
Changsha, Hunan
andy_van1@sohu.com

*Abstract*—**Multi-agent generation within a special scene is a basic problem for the research on artificial society. In the context of public health emergency, the multi-agent model of an artificial campus is built; based on the deep analysis of the problem, we consider that a special scene in artificial society is approximate to a polygon determined by some key points (border points). Therefore the problem of multi-agent generation is converted to inner-point generation in the polygon. The grid-based method is proposed to divide the scene polygon reasonably consistent to the density. The filtering mechanism of grids generated is proposed. Finally, the experiment shows that the method could be used to generate some agents with some properties in a special scene randomly, and has certain reference value.**

***Keywords-artificial society; scene; grid method; infinite element***

## I. Introduction

The problem of modeling highly integrated and complex artificial society involves many aspects of natural and artificial systems, social organizations, individual psychology and behavior, with a variety of uncertainties and opposability [1]. The wide application of the computing technology and the maturity of numerical method have make it true to research the dynamics of social system in use of simulation. In early 1990s，the Land Corporation proposed the concept of artificial society, and from that on, artificial society has begun to spring up. The core method of artificial society is based on the modeling and analysis of the Agents. Its basic idea is described as following: human society is a complex system which contains a lot of individuals, and the individuals could be modeled in the computer (that is the software agent); the agents should follow certain rules of interaction; the social or group phenomena emerged by observing the whole behavior of the agents, and the model of the complex social system could be built to study information technology impact on the society, politics and culture [1, 2, 3]. The simulation of artificial society based on multi-agents modeling has become an effective solution to solve complex social systems.

In recent years, various unconventional emergencies have occurred one after another in China, and emergency management has become the major challenge to the social management. The method and technique of computer simulation used to build the scenario of the incident's happening, development, transformation and evolution has been considered the most important way to solve "scene-response" problem of unconventional emergency management. The research has been carried out abroad for some years. Arizona State University in the United States built a 8000 square feet named Decision Theater for decision support in the emergency scene in 2005. In 2007 and 2008, the U.S. Department of Homeland Security implemented the Golden Guardian for the terrorist attack and earthquake in California. The key component is the Exercise Control System to support scenario generation, learning and training, tissue remodeling, and so on [4, 5].

The scene could be considered as the environment and condition where the agents are located in artificial society, while usually the agents are constrained to act in a special area. The paper mainly studies the random generation of agents with some particular characteristics within the special scene of artificial society, and the groundwork is built for the next research.

## II. Multi-Agent Model

Based on the complex network modeling technology and the real system information acquisition technology, multi-agent system could be introduced to describe the groups and individuals in human society, which could provide the good technology ground for recurring the complexity of the social system, and approaching the social context of unconventional emergencies [6]. While social system is the complex system which contains a large number of intelligent agents, the method of agent-based modeling and simulation is feasible to describe the individual or group behavior, and analyze the social structure evolution and group behavior law with the interaction between agents. Multi-agent system is composed of many agents interacting with each other. There are also some interactions between agents and environment. Generally, each individual agent could make decision independently, and has the ability of adaptive, learning and cognition [7, 10].

The paper mainly focuses on the spread of infectious diseases within a campus, and builds the artificial campus in the context of unconventional public health emergency. The population agent in artificial society should be set the basic demographic attribute to establish the population statistics model within a specified scene in use of the population census statistics. The agents with different demographic attributes

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

and different spatial distribution could be generated automatically in geography environment. Agent modules in a campus include population statistics model, agent behavior model, disease-related model, the social relationship model, and agent interaction model [8-10].

$$Agent_{individual}:\left\{\begin{array}{l} demography\ model(t) \\ activity\ model(t) \\ disease\ referred\ model(t) \\ social\ relationship\ model\ (t) \\ communication\ model(t) \end{array}\right\}$$

The agent's population census statistics can be described from two aspects. One is public population census attributes, owned by any agent in any application area, such as ID, sex, age, role, and so on ; the other is domain-related population census attribute, for example, in public health event, including the immunity ability, symptoms, duration of symptoms, spread probability, and so on.

The agents in artificial society can be categorized according to the role of population census. A class of agent represents a kind of entity in real world. Based on the agent model with public population census attributes, the special attributes and behavior rules can be added newly. For example, except for the basic attributes (ID, sex, age) for the student agent, the special attributes (grade, professional, class) and related behaviors (entertainment, community activities, sports) should be added. A common agent could be described as shown in fig.1.

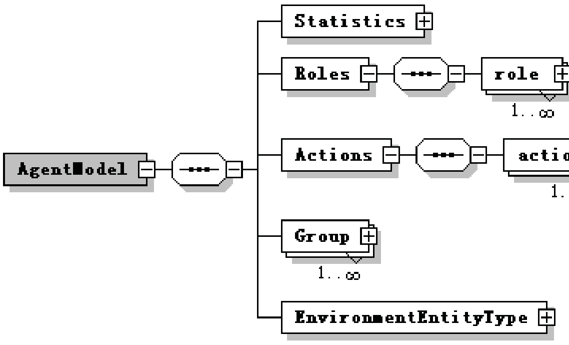


Figure 1. Agent model

## III. Boundary of special scene

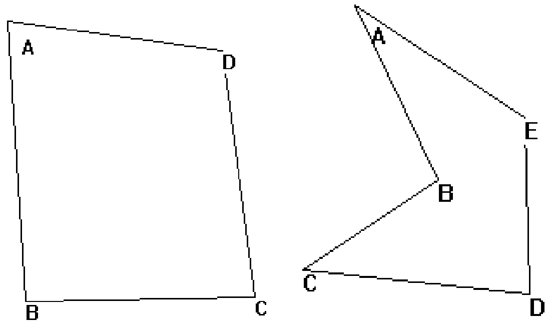
### A. Setting the Boundary

The scene is the place where the agents act, and has its own boundary and constraint. The agents always act within a certain area, and execute some related activities.

There are some criteria to justify the algorithm. The agents must be generated inside the scene, and the distribution of different agents should be consistent with the density of different agent in the scene. In order to determine the scene boundary, the coordinate of some key points which reflect the boundary shape should be given, and the scene can be generated based on the points. Especially for the complex shape, it's feasible to depict the scene in use of polygon approximation, and in error permission the complex shape between two points can be approximated by a straight line segment [13]. For the sake of simplicity, the scenes generated in the paper are closed polygon shape.

Given a point array $\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$ , the algorithm generating n edges shape is described as following: firstly, the index counter should be defined, and record the number of connected nodes. The initial state of the index counter is Zero. Secondly, the point $P(x_p,y_p)$ selected from n key points ($1\le p\le n$) is regarded as the start point, and the index counter should be added by 1. Then the adjacent points should be connected in sequence. Whenever a point is connected, the index counter should be added by 1. In the step the closure of polygon should be judged. If the value of the index counter is greater than n, and the point returns to the start point, the shape could be considered to be closed. So the lines between adjacent points as the sides of polygon are regarded to be the boundary of the scene.

Usually, the expressed way of the polygon in the computer is to sort the order of adjacent vertices arranged in an array. When shown the vertices can be connected by a straight line according to the store order. Geometry shape is ultimately described by the polygon.



(a) (b)

Figure 2. Polygon generated

### B. Judging the Relationship between Points and Scene

After the boundary of the scene generated, the activity range is determined in artificial society. It's important to make sure that agent should act in a special area, and can't exceed any boundary. The section mainly proposes the ray method to judge the relationship between the points and scene through refining an agent to be a single point in geography space.

At present, the ray method is the most widely used to judge the space relationship between points and polygon in engineering area [15, 16]. It's basic principle is described as following: Given a judging point P, the ray L is drawn along the direction of the polygon from the point P(Generally the ray parallels x axis or y axis); then the number of the intersection points is calculated. The relationship could be judged according to the parity of the number. That is the "odd number internal, even number external" rule. However, the rule isn't always suitable for any condition. There are still some special conditions to deal with.

(1)If the ray and a polygon vertex intersects, and two sides of the vertex lie in both sides of the ray L, the number of the intersection points is just one;

(2)If the ray and a polygon vertex intersects, and two sides of the vertex lie in same sides of the ray L, the number of the intersection points is Zero;

(3)If the ray L comes across two adjacent vertexes in succession (It means that the ray L superposes a side of the polygon), the two vertexes could be considered an intersection point. According to the computing method described in condition (1)(2), if the two fore-and-aft vertexes of the intersection point lie in different sides of the ray L, the intersection point should be regarded one point added to the number of the intersection point; otherwise, it can't be computed.

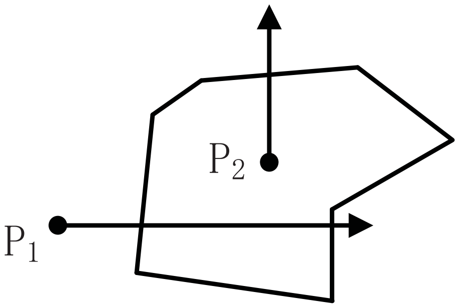


Figure 3. Ray method

The ray could always come through the polygon, because the polygon is bounded. Based on the ray method, the algorithm process could be described as following:

(1)Given the start point $P(x_p, y_p)$, and the polygon formed by a point array $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$;

(2)The ray $y = y_p, x \in (x_p, \infty)$ paralleling x axis is drawn from the start point $P(x_p, y_p)$;

(3)Each side $S_i((x_i, y_i), (x_{i+1}, y_{i+1})), i \in [0, n)$ of the polygon is obtained circularly, and there should be some needs to judge whether each side parallels x axis. If paralleling x axis, enter into the next step; otherwise, continue to cycle;

(4)Priority is to judge whether the point $P(x_p, y_p)$ lies on a side of the polygon. If it is, it's thought that the point lies in the polygon and return true, otherwise enter into the next step;

(5)Judge whether the ray $y = y_p$ and the vertex of the polygon intersect. If it is, the position relationship among the ray and the two sides of the vertex can be judged by comparing the abscissa of the vertex with $P(x_p, y_p)$, and if the position lies in different sides, the intersection point counter should be added by 1, otherwise the counter can't be changed; if not, enter into the next step;

(6)Judge whether the ray $y = y_p$ and a side of the polygon overlap. If it is, it could be considered an intersection point. If the adjacent vertexes of the ersatz point lie in the different sides of the ray, the ersatz point can be regarded as a intersection point added to the counter, otherwise the counter unchanged. If not, enter into the next step;

(7)Judge the ray and a side of the polygon intersect. If it is, the intersection point counter should be added by 1; otherwise return to the third step;

(8)Judge the overall number of the intersection points. If the number is odd number, it means the point P lies in the polygon and return true; if even number, it means the point P lies outside the polygon and return false.

## IV. GRID METHOD

Obviously, it's very important to generate some inner points within the scene automatically and randomly. As described above, the auto generation of the inner points within a scene is transformed into the auto generation of the inner points within a polygon. Recently the generation technology of the inner points within a polygon has been developed, such as the approaches based on triangle and MBR [11, 12]. These approaches have made good attempts to solve the problem of inner point generation and received ideal effects. However, the algorithm process may be so trivial and complex that too many computing resources are consumed. We proposed and introduced grid method to make sure that individual agents be ergodic and diversiform, but also not increase the scale of the inner points blindly.

### *A. Grid Density*

The basic idea of grid method is described as follows: according to the boundary of the scene polygon (for example, the value range in x axis or y axis), the scene polygon is divided into many interzone boundaries. All the interzone boundaries are intersected to form many small quadrilateral grids. The intersection points in the quadrilateral grids are the point in search space, as shown in fig. 4. When the initial points are generated, individuals can be full of the whole search space.

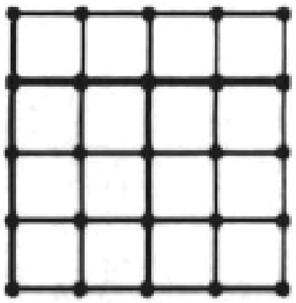


Figure 4. Grid method

In artificial society scene different area might have different population distribution density. There are some concentrated and collective phenomena in real world. It must require that the density of the grid should be considered when the scene is divided.

For the algorithms of quadrangle grids, the most representative is geometric decomposition method. Within it the two best methods are Looping and Paving, while it's hard for the two methods to control the grid density. In order to reflect the density differences among the grids, it's required that the density control should integrate with geometric decomposition.

The density of the grid can be defined as the reciprocal of the grid element length from the view point of mathematics. We adapted Laplace equation to express grid density distribution function within artificial society scene. Grid density distribution should be continuous in the area divided. It could ensure that the grid elements generated based on the grid density have the property of catholicity and universality [13, 14]. Given the grid density $U(x,y)$, the Laplace equation of the grid density distribution function can be described as following:

$$\frac{\partial^2 U}{\partial x^2}+\frac{\partial^2 U}{\partial y^2}=0 \qquad (\text{in } \Omega) \qquad (1)$$

Where $\Omega$ means the area of the scene.

The boundary condition of Eq. (1) is expressed as:

$$\frac{\partial U}{\partial n}=0 \qquad (\text{on } \Gamma) \qquad (2)$$

Where Γ is the boundary of the scene; n is normal direction of the boundary Γ.

Numerical solution of Laplace method can be attained in use of the finite element method [14]. In order to attain the grid density of certain area, the area should be divided into discrete computing elements. As shown in fig.5, the scene is a hexagon area. Based on the minimum rectangle carving, the elements outside the scene are removed, and the scene area is divided into the computing elements.

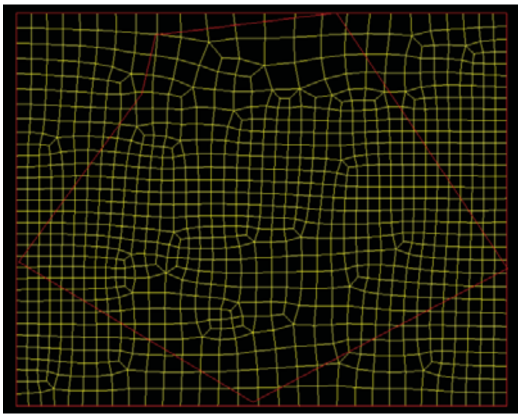


Figure 5. Scene division

The density of each point within grid elements is solved by the interpolation of the density of grid's four vertexes. The shape functions of the four vertexes are described as [14]:

$$\begin{cases} q_1=(1-\xi)(1-\eta)/4 \\ q_2=(1+\xi)(1-\eta)/4 \\ q_3=(1+\xi)(1+\eta)/4 \\ q_4=(1-\xi)(1+\eta)/4 \end{cases} \qquad (3)$$

The density of any point can be expressed as:

$$u(\xi,\eta)=\sum_{i=1}^{4} q_i(\xi,\eta)u_i \qquad (4)$$

Where $\xi,\eta$ are normalized coordinates of elements; $u_i$ is the density value of ith node.

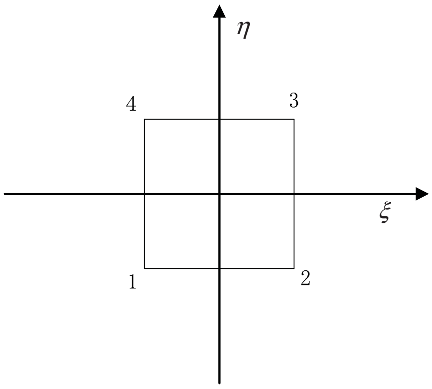


Figure 6. Grid element

Under the known boundary conditions, Eq. (1, 2, 3, 4) are united to solve the density of elements. In fact, the boundary condition of density is equal to determining the density value of the boundary points. The density value of the boundary points can be attained according the curvature. Because the area of the scene is expressed by diagnostic polygon, the length of each side in diagnostic polygon can roughly reflect the curvature of the boundary [13,14].

The division of the grid would need to consider the human activity scale, and the number of the grids divided could be given ahead of time. According to the grid density distribution, the nodes are generated in the boundary. The nodes lie in the sides of diagnostic polygon, generated between two key points. Then these nodes are numbered sequently according to the order. There need to be two steps to complete the process: the first is computing the number of the nodes; the second is determining the location of the nodes.

### B. *Computing the Number of Node*

In order to compute the number of the nodes generated, the integral for the density should be made along two key points, shown as:

$$\int_L u(s)ds = R \cong N-1 \qquad (5)$$

Where N is the overall number of the nodes including two vertexes; L is the connected line between two vertexes; u(s) is the boundary density.

In actual numerical computing, Eq. (5) could be computed according to the following method: the line L can be divided into some same line segment. The density in the middle of the small line segment is regarded as the mean density. Product by multiplying the mean density and the length of the line segment sums into the approximate integral result R. As a real number, the integral result R should be rounded up to the nearest integer N-1. In addition, the overall number of the boundary nodes must be an even number. So to meet the demand, there need to be adjust N. As the result of round, there must be some errors between the grid unit number and setting value.

### C. Determining the Location of Nodes

The finite element iteration technique is used to determining the location of the nodes. Except for the beginning and end nodes, the others are unknown. The location of the nodes(relative to the start point of the length) can be regarded as a dimension variable and the variable should meet certain equation:

$$u(s)ds = Ad\varepsilon = C \qquad (6)$$

Where $u(s)$ is the boundary density function; $d\varepsilon$ is the derivative in computing space (the nodes meet the even distribution in computing space); A is the proportion constant; C is the constant.

The Eq. (6) could be transformed into two-order derivative equation, shown as:

$$\frac{d}{dx}\left[u(s)\frac{ds}{d\varepsilon}\right]=0 \qquad (7)$$

In Eq. (7), the numerical value could be solved in use of emergence element iteration technique. The computing space is divided into N-1 units. Through the unit analysis, the unit stiff equation could be described as:

$$u[\tfrac{s_i+s_{i+1}}{2}]\begin{bmatrix}1 & -1\\ -1 & 1\end{bmatrix}\begin{bmatrix}s_i\\ s_{i+1}\end{bmatrix}=\begin{bmatrix}0\\ 0\end{bmatrix} \qquad (8)$$

Where $s_i$ means the location of the ith node.

The whole stiff equation can be attained by fitting all the unit stiff equations together. The location of the first node is set $s_1 = 0$, and the location of the Nth node is set $s_N = L$ (the length between two key points). The others are assumed to meet the even distribution. The location of each node could be attained by solving the whole stiff equation iteratively.

### D. Generating Randomly

In the paper we use middle-square method to generate the stochastic number. The square method firstly starts from some initial seeds. Based on the algorithm of the square method, false stochastic array could be acquired. After generating false stochastic number, $R_{new}$ is defined according to the number of elements selected in the scene, described as follows:

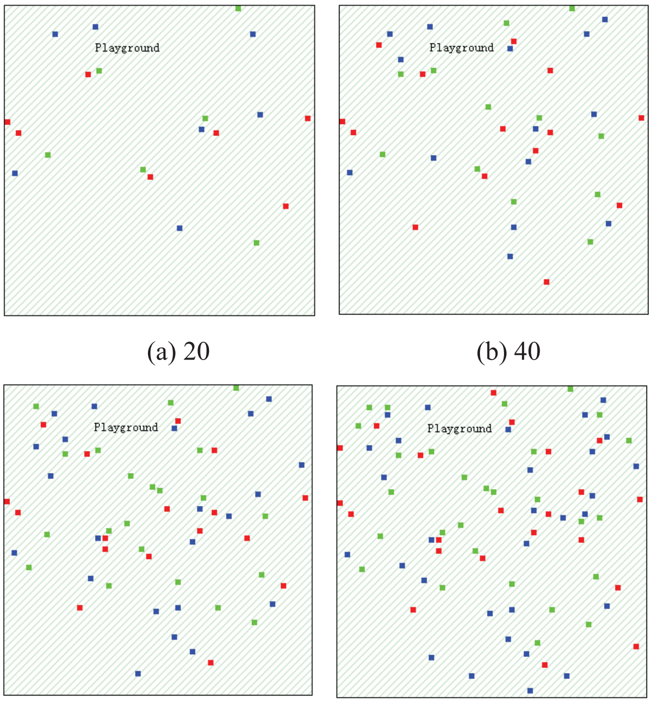
$$R_{new} = rand()/(N-1) \qquad (9)$$

The points in the grids can be indexed from the selected grids. The whole process is random, done completely by computer.

## V. Experiment

The experiment has implemented the auto generation of student agents within the playground artificial campus in the context of public health event. The playground boundary is formed by some key points known firstly. According to the daily statistical data, the personnal distribution is set reasonably.

For simple, the playground is regular rectangle, and the playground scene is divided uniformly. Here, 20, 40, 80 and 100 agents could be generated in the playground respectively given ε=σ=4, as shown in fig. 7. The red point, blue point, and green point represent three kinds of students respectively. Through computing, at the first time the number of the grids selected is 7715. After further selected, the final number of grids is decreased to 7056. The integer introduced in the process of computing the grid number results in the number reduced. Meanwhile the integer leads to error.



(a) 20

(b) 40

(c) 60

(d) 80

Figure 7. Playground Scene

As shown in fig.7, all the agents are almost generated in the playground, and there could be different density in the different district. The simulation has justified the algorithm effectively.

## VI. CONCLUSION

In the paper, we proposed a grid method, fully considered different distribution density within the actual scene, and adapted a division way to reflect agent's distribution. The experiment result showed it don't only generated agents randomly, but also ensured that the agents be ergodic and real. However, we just discussed the auto generation of multi-agents within artificial society scene from the view of 2D, and made some simplification in the process of generation. It's unavoidable that there are still some errors and infects to improve. With the development and maturity of computer and simulation technology, the future construction and implement of the artificial society must face to a wide prospect.

## ACKNOWLEDGMENT

The authors would like to thank National Nature and Science Foundation of China under Grant No. 9102403.

## REFERENCES

[1] Qiu Xiaogang, Fan Zongchen, Chen Bin, Cao Zhidong, Wang Feiyue. Requirements and Challenges of Modeling and Simulation in the Unconventional Emergency Management[J], System Simulation Technology, 2011, 3(7): 169-176.

[2] Epstein, JM R Axtell, Growing Artificial Societies: Social Science from the Bottom Up: Mit Press, 1996.

[3] Epstein, JM R Axtell. Artificial Societies and Generative Social Science. Artificial Life and Robotics, 1997, 1(1): 33-34.

[4] Security, Governor's Office of Homeland. Golden Guardian 2008 Exercise Series, 2008, http://www.ohs.ca.gov/hseep/golden_guardian/.

[5] Decesion Theater, 2010, http://www.decisiontheater.org/.

[6] Bellomo, N., C. Bianca M. Delitala. Complexity analysis and mathematical tools towards the modelling of living systems. Physics of Life Reviews, 2009, 6(3): 144-175.

[7] Shibuya, S. A multi-agent based modeling of dynamical human behavior in indoor emergency situation. Iasted: Proceedings of the Iasted International Conference on Modeling and Simulation, 2003: 413-418

[8] Yuanzheng Ge, Wei Duan, Xiaogang Qiu, Kedi Huang. Agent Based Modeling for H1N1 Influenza in Artificial Campus, IEEE ICEMMS, Beijing, 2011.

[9] Yuanzheng Ge, Liang Liu, Bin Chen, Xiaogang Qiu, Kedi Huang. Agent-Based Modeling for Influenza H1N1 in an Artificial Classroom, Systems Engineering Procedia, 2011.

[10] Yuanzheng Ge, Xiaogang Qiu, Zongchen Fan, Kedi Huang. The Network model of artificial society based on agent, The Second National Society Computing Conference, Beijing, 2010.

[11] Ding Yao. New Method to Automatically Create Polygon Node Based on Triangulation Theory[J], Journal of Chongqing Institute of Technology(Natural Science), 2008,22(3).

[12] Cai Shaohua, Qin Zhiyuan, Zhu Tao. New Method to Automatically Create Polygon Node Using MBR[J], Engineering of Surveying and Mapping,1998,7(6).

[13] Ma Xinwu, Zhao Guoqun, Wang Fang. The automatic generation of quadrilateral mesh-I: the method of domain division[J], Journal of Plasticity engineering, 2007,14(3).

[14] Ma Xinwu, Zhao Guoqun, Wang Fang. The automatic generation of quadrilateral mesh-II: the control of mesh density[J], Journal of Plasticity engineering, 2007,14(5).

[15] Jiang Ping, Liu Minshi. Improved ray method to judge the relation of point and polygon including simple curve[J], Science of Surveying and Mapping, 2009,34(5):220-222.

[16] Zou Youjian, Xiao Longxin, Chen Ding. A Contrast between Two Approaches to Find Whether the Point Being inside a Polygon[J], Surveying and Mapping of Geology and Mineral Resources, 2009,25(3).

[17] Quan Tiehan, Lu Hongwei, Yu Qifeng. The Application of Grid Method in Large Deformation Measurement[J], JOURNAL OF EXPERIMENTAL MECHANICS, 2000,15(1)：83-91.

[18] Yang Zhenhai, Zhang Guozhi. Random Number Generation[J], Symbolic Statistics and Management, 2006,25(2):244-252.

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# Robust predictor for uncertain dead time systems

M. Najafi*, F. Sheikholeslam*, S. Hosseinnia**

*Department of Electrical and Computer Engineering, Isfahan University of Technology
Isfahan, Iran, (E-mail address: majd.najafi@ec.iut.ac.ir, sheikh@cc.iut.ac.ir).

**Department of Electrical Engineering, Najafabad Branch, Islamic Azad University
Najafabad, Iran, (E-mail address: hoseinia@cc.iut.ac.ir).

*Abstract*—**This paper presents a method to stabilize uncertain dead time system based on new robust state predictor. The proposed predictor consists of a state delayed observer. Controller gain and predictor parameters are calculated by solving a nonlinear matrix inequality. More importantly, this method is extended to dead time system with a long time delay or significant uncertainty using sequential sub-predictors (SSP). This predictor composed of collection of sub predictors that each of them predicts the state for a small part of a long time delay. The number of predictors can increase attending the unstability, delay value or uncertainty where the stability condition is satisfied. Examples illustrate the capability of this method.**

***Keywords- Dead time systems, Robust Sequential sub-predictor, Robust control.***

## I. Introduction

Prediction of states or output plays a fundamental role in the control of dead time systems. This is because of delay in input impedes of stabilizing closed loop system via classical controller. To overcome this challenge, many efforts are devoted on presenting a dead time compensator (DTC) or predictor.

The famous model based predictor was presented by Smith [1] in 1957 for stable systems. Smith predictor is then modified for unstable system by Watanabe [2] (MSP) and for mixed unstable and stable systems (USP) in [3]. Moreover, [4] and [5] are applied it for stable systems with an integrator and long time delay. However, MSP may lead to unstable pole-zero cancelation for unstable systems. Therefore good approximate of distributed delay term in MSP is unavoidable to achieve the stability [6]. In addition, MSP is very sensitive to parameters and delay uncertainty. In recent years, some researchers try to improve the robustness of this method [6-12]. Also there exist a few examples that used this predictor for $H_\infty$ control of dead time systems [13-15].

Another family of predictor is classified as Finite-Spectrum Assignment [16] (FSA). The main idea of some state predictor like as Artstein reduction model [17] is much closed to FSA. [18] has been proposed a different version of FSA attending the pole-assignment methods of delay free systems. Moreover, the FSA and MSP scheme can lead to equivalent stabilize method for single input delay systems [19]. So, FSA is also sensitive to the method of distributed delay approximation [6]. This challenge will be deeper for unstable system with long time delay and uncertainty.

The main problem in these method roots in inflexibility of FSA and MSP due to uncertainty of model. This is because of the delay term of dead time systems is eliminated directly by them and they have significant challenge when face to uncertainty in system model.

To address this challenge, this paper presents a new robust state predictor that is based on state delayed observer. This predictor forecasts the state of system asymptotically. The parameters of proposed predictor can be set attending the weight of uncertainty and time delay. The state feedback is then designed applying the prediction state and the robust stability of closed loop system is proven. More importantly, proposed state predictor is extended to sequential sub-predictors (SSP) for unstable systems with a long time delay. SSP is founded on a collection of sub-predictors. The state of system is successively forecasted by each sub-predictor for a small part of delay, such that totally, SSP predicts the state for whole time delay. Consequently, the state feedback is calculated using SSP. Examples illustrate the ability of this method to stabilize dead time systems with long time delay and uncertainty.

## II. Problem statement and preliminaries

Consider linear input-delay uncertain system described by

$$\begin{cases} \dot{x}(t) = (A + \Delta A(t))x(t) + (B + \Delta B(t))u(t-d) \\ x(t) = \varphi(t) \quad \forall t \in [-d,0] \end{cases} \tag{1}$$

where $x \in \Re^n$, $u \in \Re^m$ and $d > 0$. The matrices $A$ and $B$ are known and time-varying bonded matrices $\Delta A$ and $\Delta B$ are described the uncertainty of this system where

$$[\Delta A \; \Delta B] = DJ(t)\,[E \; E_b] \tag{2}$$

$$J(t)^T J(t) \le I \tag{3}$$

The delay in the input prevents achieving stability for unstable systems with long time delay or significant uncertainty. Therefore, it is suggested to predict the state of this system to eliminate the delay in state feedback. In the other words, if $x(t+d) \approx x_p(t)$, then the delay in the controllable input can be compensated by using the predicted state instead of real state, i.e. $u(t-d) = Kx_p(t-d) \approx Kx(t)$.

The target goal in this paper is to suggest a robust control method to stabilize dead time system based on new robust predictor. Section 3 presents the simple form of this predictor and also investigates the calculation of the predictor parameters and controller gain. This method is extended to sequential sub-predictor for unstable systems with long time delay in Section 4. The predictor parameters and controller gain are also designed in this section. Examples show the capability of this method to stabilize dead time systems in Section 5. First, necessary lemma is presented as follows**.**

*Lemma 1:* [20] Given matrices $\Omega, \Gamma$ and $\Sigma$ of appropriate dimensions and with $\Omega$ symmetrical, then

$$\Omega + \Gamma J(\tau)\Sigma + \Sigma^T J(\tau)^T \Gamma^T < 0 \tag{4}$$

For all $F(\tau)$ satisfying $J(\tau)^T J(\tau) \le I$, if and only if there exists a scalar $\varepsilon > 0$ such that

$$\Omega + \varepsilon\Gamma\Gamma^T + \varepsilon^{-1}\Sigma^T\Sigma < 0 \tag{5}$$

## III. State predictor

In this section, the initial format of robust state predictor is presented as

$$\dot{\bar{x}}(t) = A\bar{x}(t) + Bu(t) + L(\bar{x}(t-d) - x(t)) \tag{6}$$

where $\bar{x} \in \Re^n$ is the predicted state that will forecast $x$ for $d$ seconds. Error is described by

$$e(t) = \bar{x}(t-d) - x(t) \tag{7}$$

The state of predictor forecasts the state of system if error converges to zero, i.e.

$$\bar{x}(t) \to x(t+d) \tag{8}$$

Now, the error dynamic equation can be calculated as

$$\dot{e}(t) = Ae(t) - \Delta Ax(t) + Le(t-d) - \Delta Bu(t-d) \tag{9}$$

The predictor matrix $L$ must be chosen such that the error converges to zero asymptotically. Following Theorem investigates the design of prediction parameter $L$ such that error equation converges to zero and calculation of the controller gain, $K$, to achieve robust closed loop stability.

*Theorem 1*: Consider system (1) with following control law.

$$\begin{cases} \dot{\bar{x}}(t) = A\bar{x}(t) + Bu(t) + L(\bar{x}(t-d) - x(t)) \\ u(t) = K\bar{x}(t) \end{cases} \tag{10}$$

Assume that $(A, B)$ is controllable. The closed-loop system is robust asymptotically stable and $\bar{x}$ predicts $x$ for $d$ second if there exist symmetric matrices $\overline{P}_0 > 0$, $\overline{P}_1 > 0$, $\overline{Q} > 0$, $\overline{S} > 0$, matrices $Y$, $U$, $M_1$, $F_0$ of appropriate dimensions, and scalar $\varepsilon$ such that the following inequality holds.

$$\begin{bmatrix} \Omega_{11} & \Omega_{12} & -\tilde{d}\overline{Y} & \tilde{d}\overline{P}\tilde{A}^T & \overline{P}\tilde{E}^T + F^T E_b{}^T \\ * & \Omega_{22} & -\tilde{d}\overline{U} & \tilde{d}M^T & 0 \\ * & * & -\tilde{d}\overline{P}\overline{S}^{-1}\overline{P} & 0 & 0 \\ * & * & * & -\tilde{d}\overline{S} + \varepsilon\tilde{D}\tilde{D}^T & 0 \\ * & * & * & * & -\varepsilon I \end{bmatrix} < 0 \tag{11}$$

where

$$\Omega_{11} = \overline{P}\tilde{A}^T + \tilde{A}\overline{P} + \overline{Y} + \overline{Y}^T + \overline{Q} + \varepsilon\tilde{D}\tilde{D}^T$$

$$\Omega_{12} = M - \overline{Y} + \overline{U}^T$$

$$\Omega_{22} = -\overline{Q} - \overline{U} - \overline{U}^T$$

And

$$\overline{P} = diag\{\overline{P}_0, \overline{P}_1\},\ \tilde{A} = diag\{A, A\},\ F = \begin{bmatrix} F_0 & 0 \end{bmatrix},\ \tilde{d} = d,$$

$$\tilde{B} = \begin{bmatrix} B \\ 0 \end{bmatrix},\ M = \begin{bmatrix} 0 & M_1 \\ 0 & M_1 \end{bmatrix},\ \tilde{D} = \begin{bmatrix} 0 \\ -D \end{bmatrix},\ \tilde{E} = \begin{bmatrix} E & -E \end{bmatrix}, \tag{12}$$

Moreover, $L$ and $K$ are given as:

$$K = P_0^{-1}F, \quad L = P_1^{-1}M_1 \tag{13}$$

*Proof*: By considering (7), the closed-loop system (1) and (10) can be rewritten as:

$$\dot{\tilde{x}}(t) = (\tilde{A} + \Delta\tilde{A} + \tilde{B}\tilde{K} + \Delta\tilde{B}\tilde{K})\tilde{x}(t) + \tilde{A}_h\tilde{x}(t - \tilde{d}) \tag{14}$$

where

$$\tilde{x} = \begin{bmatrix} \tilde{x}_1 = x + e \\ \tilde{x}_2 = e \end{bmatrix},\ \tilde{A}_h = \begin{bmatrix} 0 & L \\ 0 & L \end{bmatrix},\ \Delta\tilde{A} = \begin{bmatrix} 0 & 0 \\ -\Delta A & \Delta A \end{bmatrix},$$

$$\Delta\tilde{B} = \begin{bmatrix} 0 \\ -\Delta B \end{bmatrix},\ \tilde{K} = \begin{bmatrix} K & 0 \end{bmatrix}, \tag{15}$$

and $\tilde{A}$ and $\tilde{B}$ appear in (12). The stability of (14) is equivalent to closed loop system (1) and (10) and $\bar{x}$ approaches $x(t+d)$ if $\tilde{x}_2 = e(t)$ converge to zero asymptotically. A Lyapunov function is candidate to investigate the stability of (14) as follows.

$$V(\tilde{x}) = V_1(\tilde{x}) + V_2(\tilde{x}) + V_3(\tilde{x}) \tag{16}$$

where

$$V_1(\tilde{x}) = \tilde{x}(t)^T P\tilde{x}(t)$$

$$V_2(\tilde{x}) = \int_{-\tilde{d}}^{0} \int_{t+\beta}^{t} \dot{\tilde{x}}(\alpha)^T S\dot{\tilde{x}}(\alpha)\, d\alpha\, d\beta$$

$$V_3(\tilde{x}) = \int_{t-\tilde{d}}^{t} \tilde{x}(\alpha)^T Q\tilde{x}(\alpha) d\alpha \qquad (17)$$

By using Newton-Leibniz formula and free-weighting matrix (FWM), similar to the proof of Theorem 1 in [21], the derivative of $V(t)$ can be written as

$$\dot{V}(t) = \frac{1}{\tilde{d}} \int_{t-d}^{t} \xi(t,\alpha)^T \Lambda(\tilde{d})\xi(t,\alpha)d\alpha \qquad (18)$$

where

$$\xi(t,\alpha) = \begin{bmatrix} x(t)^T & x(t-\tilde{d})^T & x(\alpha)^T \end{bmatrix}^T$$

$$\Lambda(\tilde{d}) = \begin{bmatrix} \Xi_{11} & \Xi_{12} & -\tilde{d}Y \\ * & -Q-U-U^T+\tilde{d}\tilde{A}_h{}^T S\tilde{A}_h & -\tilde{d}U \\ * & * & -\tilde{d}S \end{bmatrix} \qquad (19)$$

where

$$\begin{aligned} \Xi_{11} &= (\tilde{A}+\Delta\tilde{A}+\tilde{B}\tilde{K}+\Delta\tilde{B}\tilde{K})^T P + P(\tilde{A}+\Delta\tilde{A}+\tilde{B}\tilde{K}+\Delta\tilde{B}\tilde{K}) \\ &\quad + d(\tilde{A}+\Delta\tilde{A}+\tilde{B}\tilde{K}+\Delta\tilde{B}\tilde{K})^T S(\tilde{A}+\Delta\tilde{A}+\tilde{B}\tilde{K}+\Delta\tilde{B}\tilde{K}) \\ &\quad + Y + Y^T + Q \end{aligned}$$

$$\Xi_{12} = P\tilde{A}_h - Y + U^T + d(\tilde{A}+\Delta\tilde{A}+\tilde{B}\tilde{K}+\Delta\tilde{B}\tilde{K})^T S\tilde{A}_h$$

Using Schur complement and Lemma 1, it is possible to show that $\Lambda(\tilde{d})$ is negative definite (i.e. $\dot{V}(t) < 0$ ) if following inequality holds.

$$\begin{bmatrix} \Psi_{11} & \Psi_{12} & -\tilde{d}Y & \tilde{d}\tilde{A}^T S & E^T + K^T E_b{}^T \\ * & \Psi_{22} & -\tilde{d}U & \tilde{d}\tilde{A}^T S & 0 \\ * & * & -\tilde{d}S & 0 & 0 \\ * & * & * & -\tilde{d}S + \varepsilon S\tilde{D}\tilde{D}^T S & 0 \\ * & * & * & * & -\varepsilon I \end{bmatrix} < 0 \qquad (20)$$

where

$$\Psi_{11} = \tilde{A}^T P + P\tilde{A} + Y + Y^T + Q + \varepsilon P\tilde{D}\tilde{D}^T P$$

$$\Psi_{12} = P\tilde{A}_h - Y + U^T$$

$$\Psi_{22} = -Q - U - U^T$$

Taking $\overline{P} = P^{-1} = diag\{\overline{P}_0, \overline{P}_1\}$, $\overline{S} = S^{-1}$, $\overline{Q} = P^{-1}QP^{-1}$, $\overline{Y} = P^{-1}YP^{-1}$, $\overline{U} = P^{-1}UP^{-1}$, pre- and post multiplying both sides of LMI (20) by $diag\{\overline{P},\overline{P},\overline{S},\overline{P},I\}$ and considering (13), it is possible to rewrite (20) as (11). □

Note that (11) is not linear due to $\overline{P}\,\overline{S}^{-1}\overline{P}$ term. The best idea to solve it without any limitation of degree of freedoms is reducing the original non-convex problem to an LMI-based nonlinear minimization problem. Then a modified cone complementarity linearization (CCL) algorithm [22] can be used to obtain a solution. This method is described in next section.

Theorem 1 may not be useful for unstable dead systems with a long time delay and weighty uncertainty (see examples in Section 5). To overcome this challenge, next section presents an extended form of this predictor called sequential sun-predictors. In this method, a series of sub-predictor is employed to each of them forecasts the state of system for small part of time delay.

## IV. SEQUENTIAL SUB -PREDICTORS

Long time delay may impede stabilizing of unstable systems with uncertainty using proposed predictor in Section 3 and it may exist no $L$ such that (11) is feasible. For this case, sequential sub-predictor (SSP) is suggested in this section. In SSP, time delay is divided to $R$ small part, and then a collection of successive sub-predictors are used to forecast the state for each small part of delay, $\overline{d}$, where

$$\overline{d} = \frac{d}{R}, \qquad R \in Z^+ \qquad (21)$$

The SSP is described by

$$\begin{cases} \dot{\overline{x}}_1(t) = A\overline{x}_1(t) + Bu(t-\overline{d}) \\ \qquad\qquad + L_1(\overline{x}_1(t-\overline{d}) - \overline{x}_2(t)) \\ \vdots \\ \dot{\overline{x}}_{R-1}(t) = A\overline{x}_{R-1}(t) + Bu(t-(R-2)\overline{d}) \\ \qquad\qquad + L_{R-1}(\overline{x}_{R-1}(t-\overline{d}) - \overline{x}_R(t)) \\ \dot{\overline{x}}_R(t) = A\overline{x}_R(t) + Bu(t-(R-1)\overline{d}) \\ \qquad\qquad + L_R(\overline{x}_R(t-\overline{d}) - x(t)) \end{cases} \qquad (22)$$

where $\overline{x}_i \in \Re^n, i = 1,\ldots,R$, Defining the prediction error as

$$\begin{cases} e_1(t) = \overline{x}_1(t-R\overline{d}) - \overline{x}_2(t-(R-1)\overline{d}) \\ \vdots \\ e_{R-1}(t) = \overline{x}_{R-1}(t-2\overline{d}) - \overline{x}_R(t-\overline{d}) \\ e_R(t) = \overline{x}_R(t-\overline{d}) - x(t) \end{cases} \qquad (23)$$

Note that $\overline{x}_1(t)$ predicts the $x(t+d)$, if all error equation converge to zero i.e.

$$e_t(t) = \overline{x}_1(t-R\overline{d}) - x(t) = e_1(t) + \cdots + e_R(t) \qquad (24)$$

The error dynamics are

$$\begin{cases} \dot{e}_1(t) = Ae_1(t) + L_1e_1(t-\overline{d}) - L_2e_2(t-\overline{d}) \\ \vdots \\ \dot{e}_{R-1}(t) = Ae_{R-1}(t) + L_{R-1}e_{R-1}(t-\overline{d}) - L_Re_R(t-\overline{d}) \\ \dot{e}_R(t) = Ae_R(t) - \Delta Ax(t) + L_Re_R(t-\overline{d}) - \Delta Bu(t-d) \end{cases} \qquad (25)$$

Following theorem investigates stability of closed loop system based on SSP. In this theorem, predictor parameters, $L_i$, and controller gain, $K$, will be calculated.

*Theorem 2*: Consider system (1) with following control law.

$$\begin{cases}\dot{\bar{x}}_1(t) = A\bar{x}_1(t) + Bu(t-\bar{d}) \\ \qquad\qquad + L_1(\bar{x}_1(t-\bar{d}) - \bar{x}_2(t)) \\ \vdots \\ \dot{\bar{x}}_R(t) = A\bar{x}_R(t) + Bu(t-(R-1)\bar{d}) \\ \qquad\qquad + L_R(\bar{x}_R(t-\bar{d}) - x(t)) \\ u(t) = K\bar{x}_1(t)\end{cases} \tag{26}$$

Assume that $(A,B)$ is controllable. The closed-loop system is robust asymptotically stable and $\bar{x}_1$ predicts $x$ for $d = R\bar{d}$ second if there exist symmetric matrices $\bar{P}_i > 0,\ i = 0,1,\cdots,R,$ $\bar{Q} > 0,\quad \bar{S} > 0$, matrices $Y,\ U,\ M_j,\ j = 1,\cdots,R,\ F_0$ of appropriate dimensions, and scalar $\varepsilon$ such that (11) holds, substituted

$$\bar{P} = diag\{\bar{P}_0, \bar{P}_1, \cdots, \bar{P}_R\},\ \tilde{A} = diag\{A, A, \cdots, A\},\ \tilde{d} = d/R,$$

$$\tilde{B} = \begin{bmatrix} B \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix},\ M = \begin{bmatrix} 0 & M_2 & 0 & \cdots & 0 \\ 0 & M_2 & -M_2 & \ddots & 0 \\ 0 & 0 & M_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & -M_R \\ 0 & 0 & 0 & \cdots & M_R \end{bmatrix},\ \tilde{D} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ -D \end{bmatrix},$$

$$F = \begin{bmatrix} F_0 & 0 & 0 & \cdots & 0 \end{bmatrix},\ \tilde{E} = \begin{bmatrix} -E & E & E & \cdots & E \end{bmatrix}. \tag{27}$$

Moreover, $L$ and $K$ are given as:

$$K = P_0^{-1}F, \quad L = P_j^{-1}M_j,\ j = 1,\cdots,R \tag{28}$$

*Proof*: By considering (23), the closed-loop system (1) and (26) can be rewritten as:

$$\dot{\tilde{x}}(t) = (\tilde{A} + \Delta\tilde{A} + \tilde{B}\tilde{K} + \Delta\tilde{B}\tilde{K})\tilde{x}(t) + \tilde{A}_h\tilde{x}(t-\tilde{d}) \tag{29}$$

where

$$\tilde{x} = \begin{bmatrix} \tilde{x}_1 = x + e_1 + \cdots + e_R \\ \tilde{x}_2 = e_1 \\ \vdots \\ \tilde{x}_R = e_R \end{bmatrix},\ \Delta\tilde{B} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ -\Delta B \end{bmatrix},$$

$$\tilde{A}_h = \begin{bmatrix} 0 & L_2 & 0 & \cdots & 0 \\ 0 & L_2 & -L_2 & \ddots & 0 \\ 0 & 0 & L_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & -L_R \\ 0 & 0 & 0 & \cdots & L_R \end{bmatrix},\ \tilde{K} = \begin{bmatrix} K & 0 & \cdots & 0 \end{bmatrix}. \tag{30}$$

$\tilde{A}$ and $\tilde{B}$ appear in (27). The stability of (29) is equivalent to closed loop system (1) and (26) and $\bar{x}_1$ approaches $x(t+d)$ if $\tilde{x}_i$, $i = 2,\cdots,R$ converge to zero asymptotically, because $e_t(t) = \tilde{x}_2 + \cdots + \tilde{x}_R$. Based the same form of (29) and (14), proof can be followed same as proof of Theorem 1. □

To solve (11), using the modified CCL, from Schur complement, it is followed that (11) holds if

$$\begin{bmatrix} \Omega_{11} & \Omega_{12} & -\tilde{d}\bar{Y} & \tilde{d}\bar{P}\tilde{A}^T & \bar{P}\tilde{E}^T + F^T E_b{}^T \\ * & \Omega_{22} & -\tilde{d}\bar{U} & \tilde{d}M^T & 0 \\ * & * & -\tilde{d}\bar{T} & 0 & 0 \\ * & * & * & -\tilde{d}\bar{S} + \varepsilon\tilde{D}\tilde{D}^T & 0 \\ * & * & * & * & -\varepsilon I \end{bmatrix} < 0 \tag{31}$$

where

$$-\bar{P}\bar{S}^{-1}\bar{P} < -\bar{T} \tag{32}$$

Taking $T = \bar{T}^{-1}$, $S = \bar{S}^{-1}$, and $P = \bar{P}^{-1}$, from Schur complement (32) becomes

$$\begin{bmatrix} T & P \\ * & S \end{bmatrix} > 0 \tag{33}$$

The existence of a solution for (31) and (33) is a sufficiently condition for the feasibility of (11), imposing some degree of conservativeness. However, using this technique, the original of non-convex problem has been casted into the following LMI based non-linear minimization problem:

Minimize $Tr(T\bar{T} + P\bar{P} + S\bar{S})$, subject to (33) and

$$\begin{bmatrix} T & P \\ * & S \end{bmatrix} > 0,\ \begin{bmatrix} \bar{T} & I \\ * & T \end{bmatrix} > 0,\ \begin{bmatrix} \bar{S} & I \\ * & S \end{bmatrix} > 0,\ \begin{bmatrix} \bar{P} & I \\ * & P \end{bmatrix} > 0 \tag{34}$$

Then, the modified CCL algorithm is used to solve it and find the maximum possible $d$ as following procedure.

*Step 1*: Solve (31) and (34) for sufficiently small initial value of $\tilde{d}_0$ and find a feasible set $\{P_0, \bar{P}_0, S_0, \bar{S}_0, T_0, \bar{T}_0, \bar{Y}_0, \bar{U}_0, \bar{Q}_0, \cdots\}$ satisfying them. Set $j = 0,\ i = 0$.

*Step 2*: Solve the LMI (31) and (34) for all variables:

Minimize $Tr(T\bar{T}_i + \bar{T}T_i + P\bar{P}_i + \bar{P}P_i + S\bar{S}_i + \bar{S}S_i)$ subject to (31) and (34). Set $T_{i+1} = T$, $\bar{T}_{i+1} = \bar{T}$, $P_{i+1} = P$, $\bar{P}_{i+1} = \bar{P}$, $S_{i+1} = S$ and $\bar{S}_{i+1} = \bar{S}$.

*Step 3*: If (11) is satisfied and $d = \tilde{d}_i$ ($d = R\tilde{d}_i$ for Theorem 2), end. If (11) is not satisfied within a specified number of iterations ($j$), then exit with no solution. If (11) is satisfied but $d > \tilde{d}_i$ ($d > R\tilde{d}_i$ for Theorem 2), set $i = i + 1$, $j = 0$, increment $\tilde{d}_i$ and go to Step 2. Otherwise, set $i = i + 1$, $j = j + 1$ and go to Step 2.

Note that the number of sub-predictors, $R$ , can be set sufficiently big to (11) becomes feasible. In the other words, if (11) is not satisfied for $d$ in a usual number of CCL iterations, then R should be increased while it becomes feasible. Therefore this method can stabilize all unstable systems with long time delay and significant uncertainty. Next section presents a few examples to illustrate the capability of this method to stabilize unstable dead time systems with time varying uncertainty.

## V. Simulation Results

In this section a few examples are presented to shows the ability SSP to closed loop stability of dead time systems due to unstability of system, time delay value and weight of uncertainty.

*Example 1*: Consider system $\dot{x} = ax + bu(t-d)$ , where $a, b$ is scalar. SSP can forecast the states of systems for maximum delay $d = Rd_m$ that is shown in Table 1 (for iteration number 40 and increment and increment delay step 0.01).

Table 1: Maximum possible delay to prediction of state

| $d = Rd_m$ | $R = 1$ | $R = 2$ | $R = 3$ | $R = 4$ | $R = 5$ |
|---|---|---|---|---|---|
| $a = -1$ | $\infty$ | $\infty$ | $\infty$ | $\infty$ | $\infty$ |
| $a = 0.2$ | 4.93 | 9.77 | 14.55 | 19.33 | 23.05 |
| $a = 0.5$ | 1.96 | 3.85 | 5.72 | 7.56 | 9.13 |
| $a = 1$ | 0.97 | 1.90 | 2.8 | 3.69 | 4.51 |
| $a = 2$ | 0.48 | 0.93 | 1.36 | 1.78 | 2.15 |

Table 1 shows that maximum possible delay to stabilize of closed loop system is directly proportional to $R$ and inversely proportional to $a$. Although the increasing the $a$, limits the bound of delay, but it is possible to compensate it by increasing the number of sub-predictor.

*Example 2*: Consider an unstable uncertain system with integrator term and non-minimum phase zero, defined as

$$\begin{cases} \dot{x}(t) = \begin{bmatrix} 0 + .02\sin(t) & 0 \\ 0 & 0.5 + .02\sin(t) \end{bmatrix} x(t) + \begin{bmatrix} 2 \\ -2 \end{bmatrix} u(t-3) \\ y(t) = \begin{bmatrix} -1 & 1 \end{bmatrix} x(t), \qquad x(0) = \begin{bmatrix} 1 & 0 \end{bmatrix}^T \end{cases}$$

Using Theorem 2, the controller gain and SSP matrices are calculated as:

$$K = \begin{bmatrix} 1.15 & 4.59 \end{bmatrix}, \; L_1 = \begin{bmatrix} -0.04 & 0.12 \\ -0.05 & -0.86 \end{bmatrix},$$

$$L_2 = \begin{bmatrix} -0.08 & 0.23 \\ -0.01 & -0.82 \end{bmatrix}, \; L_3 = \begin{bmatrix} -0.2 & 0.34 \\ -0.01 & -0.70 \end{bmatrix},$$

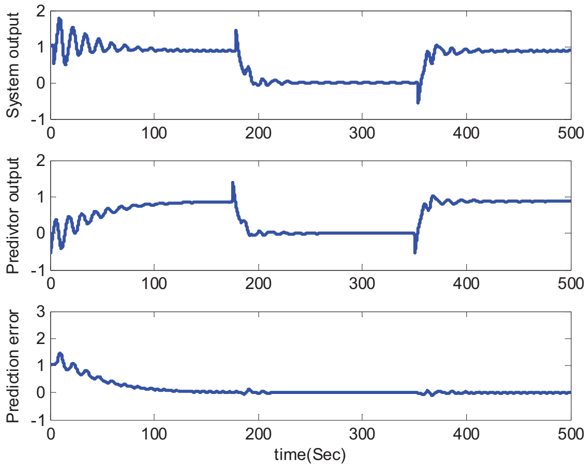


Figure.1 Pulse response of closed loop system

Note that the initial conditions of system are unknown in prediction and the initial conditions of SSP are set to zero. Figure 1 illustrates the pulse response of the closed loop system. This example shows the capability of this method to stabilize uncertain and unstable dead time systems with a long time delay.

## VI. Conclusion

This paper suggests a method to stabilize uncertain dead time systems based on a new robust predictor. Moreover, this method is applied for unstable dead time systems with a long time delay by sequential sub-predictors. The main idea in this predictor is composed of a series of sub-predictor; each of them is for a partition of long time delay. This method can improve the flexibility of predictor to overcome any weight uncertainty. This method also can apply for nonlinear systems and output feedback that will be presented in future works.

## References

[1] O. J. M. Smith, “Closer Control of Loops With Dead Time,” *Chem. Eng. Progr.*, vol. 53, pp. 217-219, 1957.

[2] K. Watanabe and M. Ito, “A Process-Model Control For linear Systems with Delay,” *Automatic Control, IEEE Transactions on*, vol. AC-26, pp. 1261-1269, 1981.

[3] Q. Zhong and G. Weiss, “A unified smith predictor based on the spectral decomposition of the plant,” *International Journal of Control*, vol. 77, p.p. 1362 – 1371, 2004.

[4] K. J. Astrom, C. Hang and B. Lim, “A New Smith Predictor For Controlling a Process with An Integrator And Long Dead-Time,” *Automatic Control, IEEE Transactions on*, vol. 39, pp. 343-345, 1994.

[5] M. Matausek and A. Micic, “A Modified Smith Predictor For Controlling a Process with an Integrator and Long Dead-Time,” *Automatic Control, IEEE Transactions on*, vol. 41, pp. 1199-1203, 1996.

[6] Q.C. Zhong, *Robust Control of Time-Delay Systems*, Springer, 2006.

[7] D. Lee, M. Lee, S. Sung and I. Lee, "Robust PID tuning for Smith predictor in the presence of model uncertainty," *Journal of Process Control*, vol. 9, iss. 1, p.p. 79–85, 1999.

[8] M. R. Stojic, F. Matijevic and L. S. Draganovic, "A Robust Smith Predictor Modified by Internal Models for Integrating Process with Dead Time," *Automatic Control, IEEE Transactions on*, vol. 46, pp. 1293-1298, 2001.

[9] R. Lozano, P. Castillo, P. Garcia and A. Dzul, "Robust Prediction-Based Control for Unstable Delay Systems: Application to the Yaw Control of a Mini-Helicopter," *Automatica*, vol. 40, pp. 603-612, 2004.

[10] S. Majhi and D. P. Atherton, "Obtaining Controller Parameters for a New Smith Predictor Using Auto tuning" *Automatica*, vol. 36, pp. 1651-1658, 2000.

[11] D. Wang, D. Zhou, Y. Jin and S. J. Qin, "A Strong Tracking Predictor For Nonlinear Processes With Input Time Delay," *Computers & chemical engineering*, vol. 28, pp. 2523-2540, 2004.

[12] D. Meng, Y. Jia, J. Du and F. Yu, "Learning Control for Time-Delay Systems with Iteration-Varying Uncertainty: a Smith Predictor-Based Approach," *IET Control Theory & Applications*, vol. 4, pp. 2707-2718, 2010.

[13] Q.C. Zhong,, "$H_\infty$ Control of Dead-Time Systems Based on Transformation," *Automatica*, vol. 39, pp. 361- 366, 2003.

[14] G. Meinsma, and H. Zwart, "On $H_\infty$ Control for Dead-Time Systems," IEEE Trans. Automat. Contr., vol. AC-45 no. 2 , pp. 272-285 , 2000.

[15] G. Tadmor, "The Standard $H_\infty$ Problem in System with a Single Input Delay," *Automatic Control, IEEE Transactions on*, vol. AC-45 no.3, pp. 382-397, 2000.

[16] W.H. Kwon and A.E. Pearson, "Feedback Stabilization of Linear System with Delayed Control," *Automatic Control, IEEE Transactions on*, vol. AC-25, pp. 266-269, 1980.

[17] Z. Artstien, "Linear Systems with Delayed Controls." *Automatic Control, IEEE Transactions on*, vol.27, no.4, pp.869-879, 1989.

[18] Q. G. Wang, T. H. Lee and K. K. Tan, *Finite Spectrum Assignment for Time Delay Systems*, Springer-Verlag , London, 1999.

[19] L. Mirkin and N. Raskin, "Every stabilizing dead-time controller has an observer-predictor-based structure," *Automatica,* vol. 39, iss.10, p.p. 1747–1754, 2003.

[20] I. R. Petersen, "A stabilization algorithm for a class of uncertain linear systems," *Syst. Contr. Let.*, vol. 8, pp. 351–357, 1987.

[21] S. Xu and J. Lam, "Improved Delay-Dependent Stability Criteria for Time Delay Systems," *Automatic Control, IEEE Transactions on*, vol. 50, no. 3, pp. 384-387, March 2005.

[22] E. L. Ghaoui, F. Oustry and M. AitRami, "A cone complementarity linearization algorithms for static output feedback and related problems," *Automatic Control, IEEE Transactions on*, vol.42, iss.10, pp. 1171-1176, 1997.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# An Improved On-Line Neuro-Identification Scheme

José A. R. Vargas, Kevin H. M. Gularte
Department of Electrical Engineering
Universidade de Brasília
Brasília, DF, Brazil.
vargas@unb.br, kevinhmg@gmail.com

Elder M. Hemerly
System and Control Department
Technological Institute of Aeronautics – Electronics
Division, São José dos Campos, SP, Brazil
hemerly@ita.br

*Abstract*— **In this paper, an on-line identification scheme is proposed to enhance the residual state error performance in face of disturbances. The proposed scheme is based on an $e_1$-modification adaptive law for the weights to approximate the unknown nonlinearities with bounded error. Besides, an identification model with feedback is introduced to improve the state error performance. The feedback is based on a bounding function to estimate an upper bound for the disturbances. Via an adaptive bounding technique and Lyapunov methods, it is proved that the residual state error performance is practically immune to disturbances. To validate the theoretical results, the identification of a four-order generalized Lü hyperchaotic system is performed.**

***Identification; neural networks; uncertain systems; Lyapunov methods; chaotic systems.***

## I. INTRODUCTION

The use of neural networks (NNs) paradigm as a powerful tool for identification of uncertain nonlinear systems has encouraged, starting from the 90s, several heuristic and theoretical studies, see for instance [1]-[8] and the references therein. This interest is motivated by the capability of the NNs to learn complex input-output mappings, since they are universal approximators, and by the inevitable presence of uncertainties in modeling problems, due to the simplification imposed by the mathematical modeling, unexpected faults, changes in operation conditions, aging of equipment, and so on. On the other hand, system neuro-identification is important not only to predict the behavior of the system, but also for providing an appealing system parameterization, which can later be used in the synthesis of control algorithms, since mathematical characterization is often a prerequisite to controller design.

Neural identification models usually employed are the dynamic ones, being their weights mainly adjusted by using gradient and backpropagation algorithms or their robust modifications [1], [3]-[8]. Most used robust modifications in neuro-identification are the $\sigma$, switching- $\sigma$, $\varepsilon_1$, parameter projection, and dead zone [3]-[8] which avoid the parameter drift. Nevertheless, to the best of our knowledge, at present most of learning algorithms for neuro-identification ensure that the residual state error is related directly to upper bounds for the approximation error, ideal weight and disturbances.

For instance in [3], the identification of a general class of uncertain continuous-time dynamical systems was proposed, and a $\sigma$ -modification adaptive law for the weights of recurrent high-order neural networks (RHONNs) was chosen to ensure that the state error converges to the neighborhood of the origin, whose radius depends directly of the approximation error and disturbances. In [4], dynamic NNs based on two-layer neural networks were used to identify a general class of uncertain nonlinear systems. It was shown that in the presence of disturbances the state error is uniformly ultimately bounded where the ultimate bound is directly proportional to an upper bound of the disturbance. In [5], the identification of delayed nonlinear system was investigated. By using identification models based on delayed neural networks with learning laws for the weights designed using a Lyapunov-Krasovskii approach, it was shown that the state error is upper bounded by a constant which depends directly of the disturbance. More recently, also others relevant works, such as [6]-[7], shown that discrete high-order neural networks and dynamic neural networks with two different time scales, respectively, can be used to identify nonlinear systems with bounded errors, which are straightforwardly related to the disturbances.

From the discussion above, observe that most of neuro-identification schemes ensure a state error performance that is directly related to the disturbance. In practice, uncertainties are inevitable, hence it is desirable to propose identification schemes with improved state error performance in face of disturbances. This is the main motivation for this paper.

Hence, in this paper we propose a neuro-identification algorithm in which the residual state error is inversely correlated with the disturbances to make the residual state error performance practically immune to disturbances. To this end, based on an adaptive bounding technique [9] and Lyapunov methods [10], a neural identification model with explicit feedback based on a bounding function is proposed. The aim is to approximate an upper bound for the disturbances, which is used in the stability proof, to make the Lyapunov derivative ($\dot{V}$ ) negative semi-definite practically in all error space, since bounding functions can be used to dominate positive terms in $\dot{V}$ and hence improve the performance.

## II. LINEARLY PARAMETERIZED NEURAL NETWORKS

Linearly parameterized neural networks (LPNNs) can be expressed mathematically as

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

$$\rho_{nn}(W,\zeta)=W\pi(\zeta) \qquad (1)$$

where $W\in\Re^{n\times L_\rho}$, $\zeta\in\Re^{L_\zeta}$, $\pi:\Re^{L_\zeta}\mapsto\Re^{L_\rho}$ is the so-called basis function vector, which can be considered as a nonlinear vector function whose arguments are preprocessed by a scalar function $s(\cdot)$, and $n,L_\rho,L_\zeta$ are integers strictly positive. Commonly used scalar functions $s(\cdot)$ include sigmoid, tanh, gaussian, Hardy's, inverse Hardy's multiquadratic, etc [8]. However, here we are only interested in the class of LPNNs for which $\pi(\cdot)$ is bounded, since in this case we have,

$$\|\pi(\zeta)\|\le\pi_0 \qquad (2)$$

being $\pi_0$ a strictly positive constant.

The class of LPNNs considered in this work includes HONN [3], RBF networks [8], wavelet networks [11], and also others linearly parameterized approximators as Takagi-Sugeno fuzzy systems [12]. Universal approximation results in [8], [11]-[12] indicate that:

***Property 1***: Given a constant $\varepsilon_0>0$ and a continuous function $f:\Omega\mapsto\Re^n$, where $\Omega\subset\Re^{L_\zeta}$ is a compact set, there exists a weight matrix $W=W^*$ such that the output of the neural network architecture (where $L_\rho$ may depend on $\varepsilon_0$ and $f$ ) satisfies

$$\sup_{\zeta\in\Omega}\left|f(\zeta)-W^*\pi(\zeta)\right|\le\varepsilon_0 \qquad (3)$$

where $|\cdot|$ denotes the absolute value if the argument is a scalar. If the argument is a vector function in $\Re^n$ then $|\cdot|$ denotes any norm in $\Re^n$.

## III. Problem formulation

Consider the following nonlinear differential equation

$$\dot{x}=F(x,u,v,t), \qquad x(0)=x_0 \qquad (4)$$

where $x\in X$ is the *n*-dimensional state vector, $u\in U$ is a *m*-dimensional admissible input vector, $v\in V\subset\Re^q$ is a vector of time varying uncertain variables and $F:X\times U\times V\times[0,\infty)\mapsto\Re^n$ is a continuous map. In order to have a well-posed problem, we assume that $X,U,V$ are compact sets and $F$ is locally Lipschitzian with respect to *x* in $X\times U\times V\times[0,\infty)$, such that (4) has a unique solution.

We assume that the following can be established

***Assumption 1:*** On a region $X\times U\times V\times[0,\infty)$

$$\|h(x,u,v,t)\|\le h_0 \qquad (5)$$

where

$$h(x,u,v,t)=F(x,u,v,t)-f(x,u) \qquad (6)$$

$f$ is an unknown map, *h* are internal or external disturbances, and $\bar{h}_0$, such that $\bar{h}_0>h_0\ge0$, is a known constant.

Hence, except for the Assumption 1, we say that $F(x,u,v,t)$ is an unknown map and our aim is to design a NNs-based identifier for (4) to ensure that the residual state error is ultimately bounded with ultimate bound which is, practically, not affected by the disturbances.

## IV. Identification model and state error equation

We start by presenting the identification model and the definition of the relevant errors associated with the problem.

Let $\bar{f}$ be the best known approximation of *f*, $P\in\Re^{n\times n}$ a scaling matrix defined as $P=P^T>0$, $\bar{g}=P^{-1}g$, and $g(x,u)=f(x,u)-\bar{f}(x,u)$. Then, by adding and subtracting $\bar{f}(x,u)$, (4) can be rewritten as

$$\dot{x}=\bar{f}(x,u)+P\bar{g}(x,u)+h(x,u,v,t) \qquad (7)$$

***Remark 1:*** It should be noted that if the designer has no previous knowledge of *f*, then $\bar{f}$ is simply assumed as being the zero vector.

From (7), by using LPNNs, the nonlinear mapping $\bar{g}(x,u)$ can be replaced by $W^*\pi(x,u)$ plus an approximation error term $\varepsilon(x,u)$. More exactly, (7) becomes

$$\dot{x}=\bar{f}(x,u)+PW^*\pi(x,u)+P\varepsilon(x,u)+h(x,u,v,t) \quad (8)$$

where $W^*\in\Re^{nxL}$ is an "optimal" or ideal matrix, which can be defined as

$$W^*:=\arg\min_{\hat{W}\in\Gamma}\left\{\sup_{\substack{x\in X,\\ u\in U}}\left|\bar{g}(x,u)-\hat{W}\pi(x,u)\right|\right\} \qquad (9)$$

with $\Gamma=\left\{\hat{W}\ \middle|\ \|\hat{W}\|\le\alpha_{\hat{W}}\right\}$, $\alpha_{\hat{W}}$ is a strictly positive constant, $\hat{W}$ is an estimate of $W^*$, and $\varepsilon(x,u)$ is an approximation error term, corresponding to $W^*$, which can be defined as

$$\varepsilon(x,u):=\bar{g}(x,u)-W^*\pi(x,u) \qquad (10)$$

The approximation, reconstruction, or modeling error $\varepsilon$ in (10) is a quantity that arises due to the incapacity of LPNNs to match the unknown map $\bar{g}(x,u)$. Since *X*, *U* are compact sets and from (2), the following can be established

***Assumption 2:*** The Frobenius matrix norm $\left\|W^*-W_0\right\|_F$, where $W_0\in\Re^{nxL}$ is upper bounded by a known positive constant $\overline{\beta}$, such that

$$\left\| W^{*} - W_0 \right\|_F \leq \overline{\beta} \tag{11}$$

***Assumption 3:*** On a region $X \times U$ , the approximation error is upper bounded by

$$\left\| \varepsilon(x,u) \right\| \leq \varepsilon_0 \tag{12}$$

where $\bar{\varepsilon}_0$, such that $\bar{\varepsilon}_0 > \varepsilon_0 \geq 0$ , is an known constant.

***Remark 2:*** Assumption 1 is usual in identification. Assumption 2 is quite natural since $\overline{g}$ is continuous and their arguments evolve on compact sets.

***Remark 3:*** Note that any $\bar{\pi}_0 > \pi_0$ , $\bar{h}_0 > h_0$ , and $\bar{\varepsilon}_0 > \varepsilon_0$ also satisfy (2), (5), and (12). Hence, to avoid confusion, we define $\pi_0$, $h_0$, and $\varepsilon_0$ to be the smallest constants such that (2), (5), and (12) are satisfied.

***Remark 4***: It should be noted that $W^{*}$ and $\varepsilon(x,u)$ might be nonunique. However, the uniqueness of $\left\|\varepsilon(x,u)\right\|$ is ensured by (9).

***Remark 5:*** It should be noted that $W^{*}$ was defined as being the value of $\hat{W}$ that minimizes the $L_\infty$ - norm difference between $\overline{g}(x,u)$ and $\hat{W}\pi(x,u)$. The scaling matrix *P* from (7) is introduced to manipulate the magnitude of uncertainties and hence the magnitude of the approximation error. This procedure improves the performance of the identification process.

***Remark 6:*** Notice that the proposed neuro-identification scheme is a black-box methodology, hence the external disturbances and approximation error are related. Based on the system input and state measurements, the uncertain system (including the disturbances) is parameterized by a neural network model plus an approximation error term. However, the parameterization (8) is motivated by the fact that neural networks are not adequate for approximating external disturbances, since the basis function depends on the input and states, whereas the disturbances depend on the time and external variables. The aim for presenting the uncertain system in the form (8), where the disturbance *h* is explicitly considered, is also to highlight that the proposed scheme is in addition valid in the presence of unexpected changes in the systems dynamics that can emerge, for instance, due to environment change, aging of equipment or faults.

Based on structure (8) and to ensure improved state error performance, the identification model is chosen as

$$\dot{\hat{x}} = \bar{f}(x,u) + P\hat{W}\pi(x,u) - \left(l_0\hat{\psi} + \psi_0^2 P/4\right)(\hat{x} - x) \tag{13}$$

where $\hat{x}$ is the estimated state, $\hat{\psi}$ is a bounding scalar function, $l_0$ and $\psi_0$ are positive constants. It will be demonstrated that the identification model (13) used in conjunction with a convenient adjustment laws for $\hat{W}$ and $\hat{\psi}$, to be proposed in the next section, improve the residual state error performance in the presence of disturbances.

***Remark 7:*** It should be noted that in our formulation, the LPNN is only required to approximate $P^{-1}\left[f(x,u) - \bar{f}(x,u)\right]$ (whose magnitude is often small) instead of the entire function $P^{-1}\left[f(x,u)\right]$ . Hence, standard identification methods (to obtain some previous $\bar{f}$ ) can be used together with the proposed algorithm to improve performance.

By defining the state estimation error as $\tilde{x} := \hat{x} - x$ , from (8) and (13), we obtain the state estimation error equation

$$\dot{\tilde{x}} = P\tilde{W}\pi(x,u) - \left(l_0\hat{\psi} + \psi_0^2 P/4\right)\tilde{x} - P\varepsilon(x,u) - h(x,u,v,t) \tag{14}$$

where $\tilde{W} := \hat{W} - W^{*}$ .

## V. ADAPTIVE LAWS AND STABILITY

Before presenting the main theorem, we state a fact, which will be used in the stability analysis.

***Fact 1:*** Let $W^{*}, W_0, \hat{W}, \tilde{W} \in \Re^{n \times L_p}$ . Then, with the definition of $\tilde{W} = \hat{W} - W^{*}$ , the following equalities are true:

$$\begin{aligned} 2tr\left[\tilde{W}^T\left(\hat{W} - W_0\right)\right] = \left\|\tilde{W}\right\|_F^2 \\ + \left\|\left(\hat{W} - W_0\right)\right\|_F^2 - \left\|\left(W^{*} - W_0\right)\right\|_F^2 \end{aligned} \tag{15}$$

We now state and prove the main theorem of the paper.

***Theorem 1:*** Consider the class of nonlinear systems described by (4) and the Assumptions 1-3. Let the identification model be given by (13) with

$$\dot{\hat{W}} = -\gamma_W [\gamma_0 (\hat{W} - W_0) \left\|\tilde{x}\right\| + 2\tilde{x}\pi^T(x,u)] \tag{16}$$

and

$$\dot{\hat{\psi}} = -\gamma_\psi \left[\psi_0\hat{\psi}\left\|\tilde{x}\right\| - \psi_0\left\|\tilde{x}\right\| - \psi_2 l_0 \left\|\tilde{x}\right\|^2\right] , \hat{\psi}(0) > 0 \tag{17}$$

where

$$\gamma_\psi > 0, \quad \psi_2 = 2\lambda_{\min}\left(P^{-1}\right), \tag{18}$$

$$\psi^* = 1 - \frac{1}{\psi_0}\left[\psi_0^2 - 2\psi_0\left(2\bar{\varepsilon}_0 + 2\bar{h}_0\left\|P^{-1}\right\|_F + \gamma_0\bar{\beta}^2/2\right)\right]^{\frac{1}{2}}$$

Then, the signal errors $\tilde{W}, \tilde{\psi}$ are uniformly bounded and the state error $\tilde{x}$ is uniformly ultimately bounded.

***Proof:*** Consider the Lyapunov function candidate

$$V = \tilde{x}^T P^{-1}\tilde{x} + tr\left(\tilde{W}^T \gamma_W^{-1}\tilde{W}\right)/2 + \gamma_\psi^{-1}\tilde{\psi}^2/2 \qquad (19)$$

where $\tilde{W} = \hat{W} - W^*$ and $\tilde{\psi} = \hat{\psi} - \psi^*$.

By evaluating (19) along the trajectories of (14), (16) and (17), we obtain

$$\begin{aligned}\dot{V} &= 2\tilde{x}^T\tilde{W}\pi - 2l_0\hat{\psi}\tilde{x}^T P^{-1}\tilde{x} - 2\tilde{x}^T\varepsilon - 2\tilde{x}^T P^{-1}h \\ &- \gamma_0 \| \tilde{x} \| tr[\tilde{W}^T(\hat{W} - W_0)] - 2tr(\tilde{W}^T\tilde{x}\pi^T) \\ &- \psi_0 \| \tilde{x} \| \tilde{\psi}\hat{\psi} + \psi_0\tilde{\psi} \| \tilde{x} \| + \psi_2 l_0\tilde{\psi} \| \tilde{x} \|^2 - \psi_0^2\left\|\tilde{x}\right\|/2\end{aligned} \qquad (20)$$

Furthermore, by using the representations $tr\left(\tilde{W}^T\tilde{x}\pi^T\right) = \tilde{x}^T\tilde{W}\pi$ and $2\tilde{\psi}\hat{\psi} = \tilde{\psi}^2 + \hat{\psi}^2 - \psi^{*2}$, the fact 1 and the definition (18), (20) can be upper bounded as

$$\begin{aligned}\dot{V} \le &- l_0\psi_2\hat{\psi} \| \tilde{x} \|^2 + (2\bar{\varepsilon}_0 + 2\bar{h}_0 \| P^{-1} \|_F) \| \tilde{x} \| \\ &- \gamma_0 \| \tilde{x} \| \| \tilde{W} \|_F^2/2 + \gamma_0 | W^* - W_0 \|_F^2 \| \tilde{x} \|/2 \\ &- \psi_0 \| \tilde{x} \| (\tilde{\psi}^2 + \hat{\psi}^2 - \psi^{*2})/2 + \psi_0\tilde{\psi} \| \tilde{x} \| \\ &+ \psi_2 l_0\tilde{\psi} \| \tilde{x} \|^2 - \psi_0^2 \| \tilde{x} \|^2/2\end{aligned} \qquad (21)$$

By employing the definitions of $\psi_2$ and $\psi^*$, see (18), and recalling that $\tilde{\psi} = \hat{\psi} - \psi^*$, (21) implies

$$\dot{V} \le -\psi_2 l_0\psi^*\left\|\tilde{x}\right\|^2 + \psi_0\hat{\psi}\left\|\tilde{x}\right\| - \psi_0\left\|\tilde{x}\right\|\hat{\psi}^2/2 - \psi_0^2\left\|\tilde{x}\right\|^2/2 \qquad (22)$$

Since $\hat{\psi} \le \hat{\psi}^2/2 + 1/2$, we arrive at

$$\dot{V} \le -(\psi_2 l_0\psi^* + \frac{\psi_0^2}{2})\left\|\tilde{x}\right\|^2 + \psi_0\left\|\tilde{x}\right\|/2 \qquad (23)$$

Hence, $\dot{V} < 0$ as long as

$$\left\|\tilde{x}\right\| > \frac{\psi_0}{2\psi_2 l_0\psi^* + \psi_0^2} := \alpha \qquad (24)$$

Thus, since α is constant, by using Lyapunov arguments [10], we concluded that $\tilde{x}$ are uniformly ultimately bounded, with ultimate bound α. Based on (16) and (17) $\tilde{W}, \tilde{\psi}$ are also bounded. Note that if, by any reason, $\| \tilde{x} \|$ escapes of the residual set Ω, where $\Omega = \{ \tilde{x} \mid \| \tilde{x} \| \quad \alpha \}$, $\dot{V}$ becomes negative definite again, and force the convergence of the state error to the ball of radius $\alpha$.

□

***Remark 8:*** The existence of $\psi^*$ is guaranteed as long as $\psi_0 \ge 2\bar{\varepsilon}_0 + 2\bar{h}_0\left\|P^{-1}\right\|_F + \gamma_0\bar{\beta}/2$. However, it is a mild condition, since any increase of $\psi_0$ has only a positive impact on the residual state error, as can be seen in (24).

***Remark 9:*** Since the ultimate bound $\alpha$ is inversely proportional to $\psi^*$, which depends on an upper bound for the disturbances, see (18), the performance of the proposed method cannot be adversely affected by the increase of disturbances.

***Remark 10:*** It should be noted that $\psi^*$ might be nonunique. In fact,

$$\psi^* = 1 + \frac{1}{\psi_0}\left[\psi_0^2 - 2\psi_0\left(2\bar{\varepsilon}_0 + 2\bar{h}_0\left\|P^{-1}\right\|_F + \gamma_0\bar{\beta}^2/2\right)\right]^{\frac{1}{2}} \qquad (25)$$

also satisfy (22). However, note from (24) that the ultimate bound for the residual state error is practically of order $1/\psi_0$ for large $\psi_0$. Hence, the residual state error is, practically, not affected by disturbances, as long as the design constants are adequately selected. The above mentioned peculiarity, to the best of our knowledge, is the main advantage of the proposed scheme in comparison with the literature.

## IV. Simulations

To illustrate the application of the proposed scheme, we consider a generalized Lü hyperchaotic system described by [13], [14]

$$\begin{aligned}\dot{\bar{x}} &= a\left(y - \bar{x}\right) + u_{\bar{x}} + d_{\bar{x}} \\ \dot{y} &= b\bar{x} - k\bar{x}z + \omega + u_y + d_y \\ \dot{z} &= -cz + l\bar{x}^2 + u_z + d_z \\ \dot{\omega} &= -d\bar{x} + u_\omega + d_\omega\end{aligned} \qquad (26)$$

where $a$, $b$, $c$, $d$, $l$ and $k$ are constant parameters, $u_{\bar{x}}, u_y, u_z$ and $u_\omega$ are control inputs, and $d_{\bar{x}}, d_y, d_z$ and $d_\omega$ are unknown disturbances. It was considered that $a$=10, $b$=40, $c$=2.5, $d$=10.6, $k$=1 and $l$=4. Notice that system (26) satisfies Assumptions 1-3, since the state variables evolved on compact sets.

To identify the uncertain system (26) the proposed identification model (13) and the adaptive laws (16) and (17) were implemented. The initial conditions for the hyperchaotic system and the identification model were $\bar{x}(0)=-4,\ y(0)=-8,\ z(0)=-6,\quad \omega(0)=12$ and $\hat{x}(0)=0,$ in order to evaluate the performance of the proposed algorithm under adverse initial conditions.

The others design parameter were chosen as $u$=0,

$$\gamma_W=9,\ \gamma_0=2.5, \psi_0=50, \psi_1=10,$$
$$lo=10,\ s(\cdot)=10/[1+\exp-0.5(\cdot)],$$

$$\pi=\left[s(\bar{x}); s(y); s(z); s(\omega); s^2(\bar{x}); s^2(y); s^2(z); s^2(\omega)\right],$$

$$P=\begin{bmatrix}100 & 0 & 0 & 0\\ 0 & 50 & 0 & 0\\ 0 & 0 & 100 & 0\\ 0 & 0 & 0 & 100\end{bmatrix},$$

and $P=\begin{bmatrix}1 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0\end{bmatrix}.$

By keeping all design parameters as before, we introduced disturbances at $t=0.1$ in order to check the robustness for the proposed method. Two cases are considered:

a) $h_1(x,t)=\eta\left[e^{0.1t}\quad 2\sin(5t)\quad e^{0.1t}\cos(t)\quad 0.1\log(10+2\eta)\right]^T$

b) $h_2(x,t)=\eta[\sin(t)\quad 1.2\sin(2t)\quad \cos(4t)\quad 1.5\sin(t)]^T,$

where $\eta=0.5\|x\|$ and $x=[\bar{x}\quad y\quad z\quad \omega]^T$.

It should be noted that the last disturbance $h_1$ is unbounded as $t\to\infty$. However, it was considered here in order to evaluate the residual state error performance in the presence of severe disturbances.

The performances in the estimation of the states $\bar{x}, y, z$ and $\omega$ when disturbance $h_1$ is present are shown in Figures 1-5, and when disturbance $h_2$ is present is shown in Figure 6. We can see that the simulations confirm the theoretical results, that is, the algorithm is stable and the residual state error was, practically, not affected by the disturbance in $t=0.1\,\text{s}$.

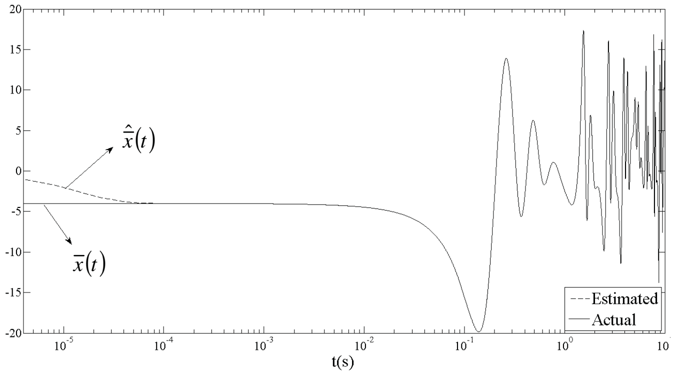


Figure 1. Actual and estimated state $\bar{x}$.

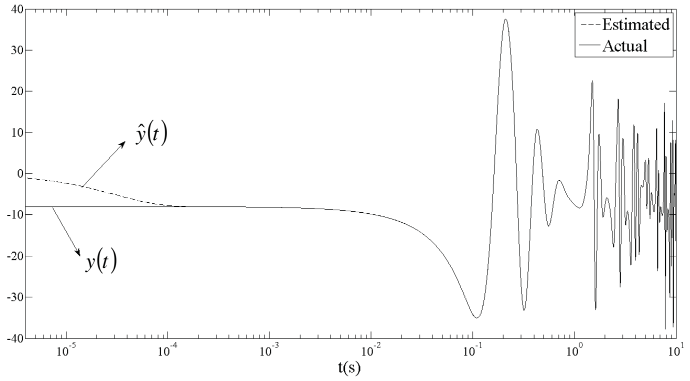


Figure 2. Actual and estimated state $y$.

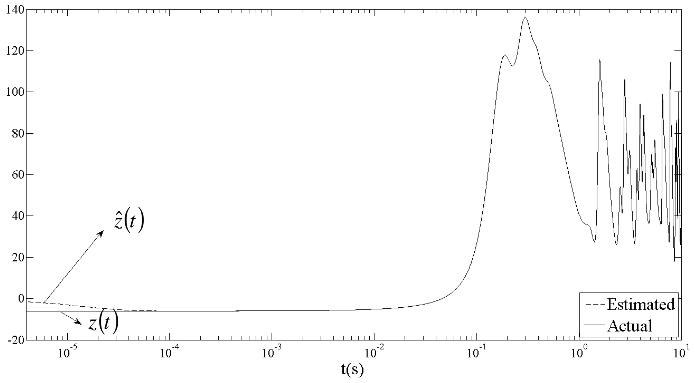


Figure 3. Actual and estimated state $z$.

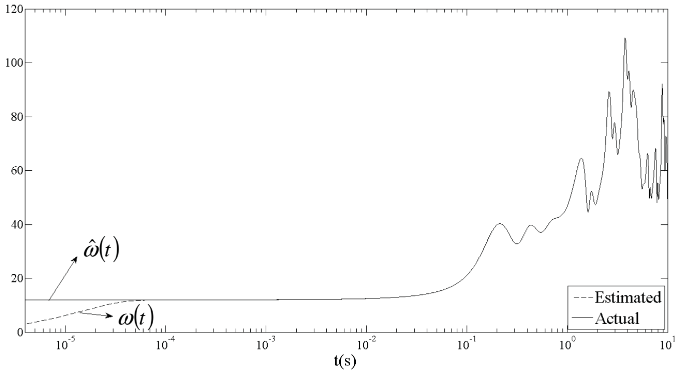


Figure 4. Actual and estimated state $\omega$.

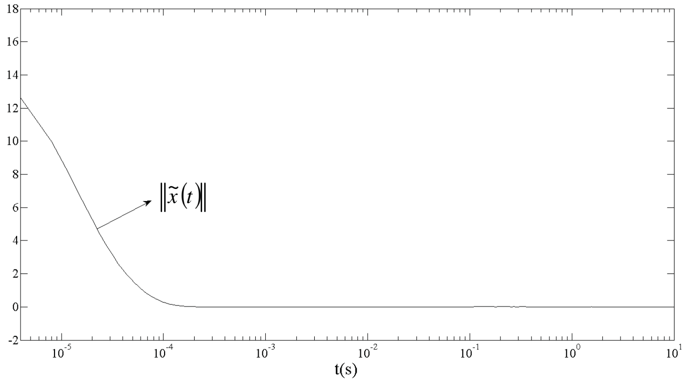


Figure 5. State error norm performance for disturbance $h_1$.

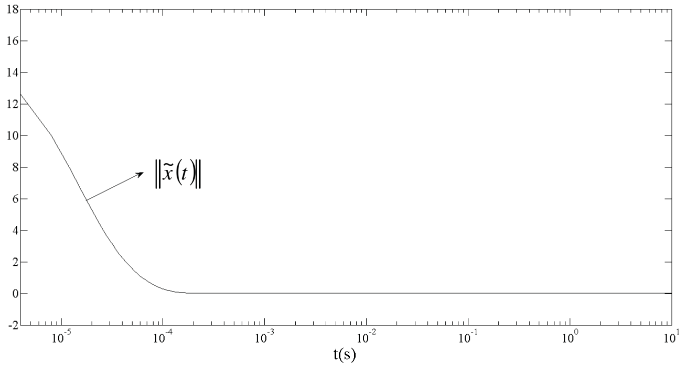


Figure 6. State error norm performance for disturbance $h_2$.

## V. Conclusions

In this work, by using Lyapunov analysis and an adaptive bounding technique, we have proposed an on-line identification scheme which presents an improved tolerance in face of disturbances. The proposed algorithm is based on a $\varepsilon_l$-modification and uses an explicit feedback on the identification model to improve residual state error performance. A simulation example showed the effectiveness of the proposed method.

## References

[1] K. S. Narendra, K. Parthasarathy, Identification and control of dynamical systems using neural networks, IEEE Trans. Neural Networks, vol. 1, no. 1, pp. 4-27, 1990.

[2] F. O. Souza, R. M. Palhares, Interval time-varing delay stability for neural networks, Neurocomputing, vol. 73, pp. 2789-2792, 2010.

[3] E. B. Kosmatopoulos, M. M. Polycarpou, M. A. Christodoulou, P. A. Ioannou, High-order neural network structures for identification of dynamical systems, IEEE Trans. Neural Networks, vol. 6. no. 2, pp. 422-431, 1995.

[4] R. R. Selmic, F. L. Lewis, Multimodel neural networks identification and failure detection of nonlinear systems, Proc. 2001 CDC, Orlando, Florida USA, pp. 3128-3133, 2001.

[5] J. J. Rubio and W. Yu. Stability analysis of nonlinear system identification via delayed neural networks, ," IEEE Trans. on Circ. Syst. II: Express Briefs, vol. 54, no. 2, pp. 161-165, 2007.

[6] A. Y. Alanis, E. N. Sanchez, A. G. Loukianov, E. A. Hernandez, Discrete-time recurrent high order neural networks for nonlinear identification, Journal of the Franklin Institute, vol. 347, pp. 1253-1265, 2010.

[7] X. Han, W. F. Xie, Z. Fu and W. Luo. Nonlinear system identification using multi-time scale neural networks, Neurocomputing, vol. 74, no. 17, pp. 3428-3439, 2011.

[8] S. S. Ge, C. C. Hang, T. H. Lee, T. Zhang, Stable adaptive neural network control, Kluwer academic publishers, 2002.

[9] M. Corless, Control of uncertain nonlinear systems, Trans. ASME J. Dyn. Sys., Meas., Control, vol. 115, pp. 362-372, 1993.

[10] P.A. Ioannou, J. Sun, Robust adaptive control, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1996.

[11] Q. Zhang and A. Benveniste, Wavelet networks, IEEE Trans. Neural Networks, vol. 3, no. 6, pp. 889-898, 1992.

[12] L. X. Wang, Adaptive fuzzy system and control: design and stability analysis, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1994.

[13] A. Loría, "Master–Slave Synchronization of Fourth-Order Lü Chaotic Oscillators via Linear Output Feedback," IEEE Trans. on Circ. Syst. II: Express Briefs, vol. 57, no. 3, pp. 213-217, 2010.

[14] F. Dou, J. Sun, K.Lü and W.Duan, "Controlling hyperchaos in the new hyperchaotic system," Commun. Nonlinear Sci. Numer. Simul., vol. 14, no. 2, pp. 552–559, Feb. 2009.

**UKACC International Conference on Control 2012**
**Cardiff, UK, 3-5 September 2012**

# Output Consensus of Linear Multi-agent Systems

Xin-Rong Yang
Center for Control Theory and Guidance Technology
Harbin Institute of Technology
Harbin, 150001, China
Email: yxr19841123@163.com

Guo-Ping Liu
Faculty of Advanced Technology
University of Glamorgan
Pontypridd, CF37 IDL, UK
Email: gpliu@glam.ac.uk

*Abstract*—**This study concerns output consensus problems of linear multi-agent systems with fixed topologies and agents described by homogeneous or heterogeneous linear systems. Based on generalized inverse, graph and linear system theory, conditions are investigated for output consensusability with respect to a set of admissible consensus protocols. The given output consensus conditions depend on both topologies within linear multi-agent systems and structure properties of each agent's dynamic. The designed output feedback control law can guarantee output consensusability of the considered linear multi-agent systems with respect to a given admissible set. A provided example illustrates applicability of the proposed approach.**

## I. Introduction

In recent years, consensus problems in multi-agent systems have received intensive attention from many fields including physics, biology and control theory and engineering due to their numerous and various applications in these areas [1]–[3]. The consensus problem is one of the most fundamental distributed coordination control problems of multi-agent networks. Roughly speaking, consensus means that multiple agents reach an agreement on a common value which might be, for example, the altitude in multi-spacecraft alignment, heading direction in flocking behavior, or average in distributed computation [4]. Numerous results have been obtained in consensus problems from different perspectives [5]–[14]. For phase transition of a group of self-driven particles, a simple and popular model has been proposed by using graph theory [2]. Numerically it has been demonstrated that all agents move in the same direction eventually. Attitude alignment has been studied for the network of agents with an undirected graph in which each agent has a discrete-time integrator dynamic [5]. Moreover, [5] has provided a theoretical explanation for the model in [2]. Many researchers have applied reinforcement learning to multi-agent systems, and these results show that reinforcement learning can perform well in multi-agent systems [6]. For first-order discrete-time multi-agent systems with time-varying topologies and stochastic communication noises, average-consensus conditions for distributed stochastic approximation type protocols have been proposed by using probability limit theory and algebraic graph theory [7]. Several dynamic consensus algorithms for second-order multi-agent systems and sufficient conditions for state consensus of the system have been proposed in [8] and [9]. Based on matrix theory, algebraic graph theory and Lyapunov control approach, [10] has derived some sufficient conditions to achieve second-order consensus for multi-agent systems with directed topology and nonlinear dynamic. A class of nonlinear high-order multi-agent systems have been studied and achieved consensus even though the communication graph which has no spanning tree [11]. To solve the consensus problem of multi-agent systems with a time-invariant communication topology and agents described by linear time-invariant systems, the necessary and sufficient condition has been given, which can guarantee the existence of an observer-type protocol solving the consensus problem, meanwhile, under this condition, an unbounded consensus region has been yielded if and only if each agent is both stabilizable and detectable [12]. [15] has provided necessary and sufficient state consensus conditions with respect to a set of admissible consensus protocols for descriptor multi-agent systems with fixed topologies.

In fact, most of existing results have investigated in consensus problems for multi-agent systems described by homogeneous agents. Comparing the discussion of consensus problems for homogeneous multi-agent systems with that for heterogeneous ones, it is easy to see that the later one has more significant due to heterogeneous multi-agent systems have more extensive description in the real world. However, rare works have been published to deal with consensus problems for multi-agent systems consisting of heterogeneous agents [16]–[18]. [16] has given state consensus conditions using tools of graph, algebra and descriptor linear system theory for descriptor multi-agent systems with agents described by homogeneous or heterogeneous descriptor systems. For heterogeneous and nonlinear multi-agent systems with topologies changing in an intermittent and arbitrary way, the necessary and sufficient condition of cooperative controllability has been proposed by using matrix-theoretical approach [17]. [18] has studied output consensus problems for a class of heterogeneous multi-agent systems consisting of uncertain linear SISO systems. Thus, this paper investigates output consensus problem for a class of multi-agent systems consisting of homogeneous or heterogeneous linear systems. Based on graph, generalized inverse and linear system theory, output consensus conditions are proposed. And designing the output feedback control law can guarantee that the studied system is output consensusable with respect to an admissible set.

*Notation:* For the given vector or matrix $X$, $\|X\|$ represents the Euclidean norm of $X$. Let $\sigma(A)$ be the set of all eigenvalues of the square matrix $A$. $\mathbb{C}^-$ represents the open left-half

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

complex plane.

## II. Preliminaries and Problem Formulation

In general, information exchange between agents in a multi-agent system can be modeled by directed or undirected graphs [19]. Similar to [20], let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$ be a weighted digraph with the set of vertices $\mathcal{V} = \{1, 2, \cdots, N\}$ and the set of edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. In $\mathcal{G}$, the $i$-th vertex represents the $i$-th agent, and a directed edge from $i$ to $j$ is denoted as an ordered pair $(i, j) \in \mathcal{E}$, which means that agent $j$ can directly receive information from agent $i$. In this case, the vertex $i$ is called the parent vertex and the vertex $j$ is called the child vertex. The set of neighbors of the $i$-th agent is denoted by $N_i = \{j \in \mathcal{V} | (j, i) \in \mathcal{E}\}$. And the weighted adjacency matrix $\mathcal{A} = [a_{ij}] \in \mathbb{R}^{N\times N}$ with nonnegative elements, and $a_{ii} = 0, a_{ij} > 0 \Leftrightarrow j \in N_i$, otherwise $a_{ij} = 0$.

**Definition** *2.1:* [21] Let $A \in \mathbb{C}^{m\times n}$. Then the matrix $X \in \mathbb{C}^{n\times m}$ satisfying the following four equations (usually called the Penrose conditions)

$$AXA = A,\ XAX = X,\ (AX)^* = AX,\ (XA)^* = XA$$

is called the Moore-Penrose inverse of $A$, and is denoted by $X = A^\dagger$.

**Lemma** *2.1:* [21] Let $A \in \mathbb{C}^{m\times n}$. The generalized inverse $X$ satisfying the Penrose conditions is existent and unique. Moreover, if rank$A = n$, then $A^\dagger A = I_n$; if rank$A = m$, then $AA^\dagger = I_m$.

Consider a linear multi-agent system consisting of $N$ agents indexed by $1, 2, \cdots, N$, respectively. The dynamics of the $i$-th agent is described by homogeneous or heterogeneous linear systems

$$\dot{x_i}(t) = A_i x_i(t) + B_i u_i(t), \tag{1a}$$

$$y_i(t) = C_i x_i(t), \tag{1b}$$

$i = 1,\ 2, \cdots\ N$, where $x_i(t) \in \mathbb{R}^{n_i}$, $u_i(t) \in \mathbb{R}^{r_i}$ and $y_i(t) \in \mathbb{R}^m$ are the state, control input and control output of the $i$-th agent, respectively; $A_i \in \mathbb{R}^{n_i\times n_i}$, $B_i \in \mathbb{R}^{n_i\times r_i}$, $C_i \in \mathbb{R}^{m\times n_i}$, and $\mathrm{rank} C_i = n_i$. System (1) is said to be homogeneous if $A_1 = A_2 = \ldots = A_N$, $B_1 = B_2 = \ldots = B_N$ and $C_1 = C_2 = \ldots = C_N$, otherwise, system (1) is said to be heterogeneous. Regarding the above $N$ agents as vertices, the topology relationship among them can be conveniently described by a digraph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$ with $\mathcal{V} = \{1, 2, \cdots, N\}$ and $\mathcal{A} = [a_{ij}] \in \mathbb{R}^{N\times N}$.

Comparing output feedback with state feedback, the former one would have more practical significant in consensus protocols. Because of constraints on measurement or economic costs in practice, it is sometimes hard to directly measure the relative information of all states. However, only the relative information of all outputs is available. Consensus protocols based on outputs or states are equivalent if all the states are measurable. Hence, adopt two kinds of output consensus protocols with the output feedback case as follows, respectively.

**Case One:** Since each agent has limited capability of collecting information, output consensus protocols for each agent in multi-agent systems are distributed and only depend on the information of the agent itself and its neighbors. Adopt output consensus protocols of the $i$-th agent are given by

$$u_1(t) = \sum_{j\in N_1} a_{1j} K_{1j}[y_j(t) - y_1(t)], \tag{2a}$$

$$\begin{aligned} u_i(t) =& D_i(C_1 A_1 C_1^\dagger - C_i A_i C_i^\dagger) y_i(t) \\ &+ \sum_{j\in N_i} a_{ij} K_{ij}[y_j(t) - y_i(t)], \\ & t \geq 0, i = 2, 3, \cdots, N, \end{aligned} \tag{2b}$$

where $D_i \in \mathbb{R}^{r_i\times m}$ such that $C_i B_i D_i = I$ and $K_{ij} \in \mathbb{R}^{r_i\times m}$, $i, j = 1, 2, \cdots, N$, are weighted constant matrices which will be designed.

Applying the property of $a_{ij}$, (2) is equivalent to

$$u_1(t) = \sum_{j=1}^{N} a_{1j} K_{1j}[y_j(t) - y_1(t)], \tag{3a}$$

$$\begin{aligned} u_i(t) =& D_i(C_1 A_1 C_1^\dagger - C_i A_i C_i^\dagger) y_i(t) \\ &+ \sum_{j=1}^{N} a_{ij} K_{ij}[y_j(t) - y_i(t)], \\ & t \geq 0, i = 2, 3, \cdots, N. \end{aligned} \tag{3b}$$

Let

$$u(t) = [u_1^T(t)\ u_2^T(t)\ \cdots\ u_N^T(t)]^T.$$

Define an admissible control set:

$$\begin{aligned} \mathcal{U}_1 =& \{u(t) : [0, \infty) \to \mathbb{R}^{rN} | u_1(t) = \sum_{j=1}^{N} a_{1j} K_{1j}[y_j(t) \\ & - y_1(t)],\ u_i(t) = D_i(C_1 A_1 C_1^\dagger - C_i A_i C_i^\dagger) y_i(t) \\ & + \sum_{j=1}^{N} a_{ij} K_{ij}[y_j(t) - y_i(t)], \\ & i = 2, 3, \cdots, N,\ t \geq 0,\ K_{ij} \in \mathbb{R}^{r\times m}, \\ & i, j = 1, 2, \cdots, N\}. \end{aligned}$$

**Case Two:** Adopt output consensus protocols of the $i$-th agent are given by

$$\begin{aligned} u_i(t) = \sum_{j\in N_i} a_{ij} K_{ij}[y_j(t) - y_i(t)], \\ t \geq 0, i = 1, 2, \cdots, N, \end{aligned} \tag{4}$$

where $K_{ij} \in \mathbb{R}^{r_i\times m}$, $i, j = 1, 2, \cdots, N$, are weighted constant matrices which will be designed.

Similar to Case One, define the following admissible control set:

$$\begin{aligned} \mathcal{U}_2 =& \{u(t) : [0, \infty) \to \mathbb{R}^{rN} | u_i(t) = \sum_{j=1}^{N} a_{ij} K_{ij}[y_j(t) \\ & - y_i(t)],\ i = 1, 2, \cdots, N, t \geq 0,\ K_{ij} \in \mathbb{R}^{r\times m}, \\ & i, j = 1, 2, \cdots, N\}. \end{aligned}$$

The admissible control set $\mathcal{U}_k (k = 1, 2)$ covers a relatively large class of distributed output consensus protocols. Therefore, we want to know under what conditions, the multi-agent system consisting of homogeneous or heterogeneous linear time-invariant systems is output consensusable with respect to such a kind of admissible control set. To solve this problem, a definition of the output consensusability of a multi-agent system with respect to an admissible control set $\mathcal{U}_k (k = 1, 2)$ is given as follows.

**Definition** *2.2:* For linear multi-agent system (1), if there exists a $u(t) \in \mathcal{U}_1$ ($\mathcal{U}_2$) such that for any initial value $x_i(0)$,

$$\lim_{t\to\infty} \|y_j(t) - y_i(t)\| = 0,\ i, j = 1, 2, \cdots, N,$$

then the system (1) is said to be output consensusable with respect to $\mathcal{U}_1$ ($\mathcal{U}_2$).

**Remark** *2.1:* Due to heterogeneity of the agents, to achieve the state consensus (i.e., $\lim\limits_{t\to\infty} \|x_j(t) - x_i(t)\| = 0$) is impossible generally.

**Remark** *2.2:* If $A_1 = A_i$ in the output consensus protocols (3), it is not necessary to assume that $C_iB_iD_i = I$.

**Remark** *2.3:* Definition 2.2 is different from the state consensus definition in [22], where the states of all the agents are required to converge to the same one constant value. Similar to [20], here only the output differences between different agents are required to tend to zero, no matter whether the outputs themselves converge or not.

**Remark** *2.4:* When $\sigma(A_i) \subseteq \mathbb{C}^-,\ i = 1, 2, \cdots, N$, take $K_{i,j} = 0,\ i, j = 1, 2, \cdots, N$ in output consensus protocols (3) or (4). Then it follows that $y_i,\ i = 1, 2, \cdots, N$, converges to zero exponentially. Hence multi-agent system (1) is naturally output consensusable with respect to $\mathcal{U}_k (k = 1, 2)$. Thus, assume that the eigenvalues of the matrices $A_i,\ i = 1, 2, \cdots, N$ are not all in the open left-half plane in this paper.

## III. Output Consensusability Conditions for Linear Multi-agent Systems

**Lemma** *3.1:* [23]For $A, B \in \mathbb{R}^{n\times n}$, if $(A, B)$ is stabilizable, then Riccati equation

$$A^TP + PA - PBB^TP + I_n = 0$$

has a unique nonnegative definite solution $P$, furthermore, $\sigma(A - BB^TP) \subseteq \mathbb{C}^-$.

**Theorem** *3.1:* System (1) is output consensusable with respect to $\mathcal{U}_1$ if the following conditions hold:

(i) $(C_1A_1C_1^\dagger,\ C_1B_1 + C_2B_2)$ and $(C_1A_1C_1^\dagger,\ C_iB_i)$, $i = 3, 4, \cdots, N$, are stabilizable;

(ii) $a_{12} \neq 0$, $a_{21} \neq 0$ and there exists at least one $j \in \{1, 2, \cdots, i-1\}$ such that $a_{ij} \neq 0,\ i = 3, 4, \cdots N$.

*Proof:*

Using Definition 2.2, system (1) is output consensusable with respect to $\mathcal{U}_1$ if and only if there exist matrices $K_{ij} \in \mathbb{R}^{r\times m}$ $i, j = 1, 2, \cdots N$, and output consensus protocols (3) such that for any $i \neq j$,

$$\lim_{t\to\infty} \|y_j(t) - y_i(t)\| = 0.$$

Combining (1a) and (1b) with $\mathrm{rank} C_i = n_i$, it follows that

$$\begin{aligned} \dot{y}_i(t) &= C_i\dot{x}_i(t) \\ &= C_iA_iC_i^\dagger C_ix_i(t) + C_iB_iu_i(t) \\ &= C_iA_iC_i^\dagger y_i(t) + C_iB_iu_i(t). \end{aligned} \tag{5}$$

Then output consensusability of system (1) is equivalent to state consensusability of system (5) with respect to $\mathcal{U}_1$, which implies there exist matrices $K_{ij} \in \mathbb{R}^{r\times m}$ $i, j = 1, 2, \cdots N$, and output consensus protocols (3) such that for any $i \neq j$,

$$\lim_{t\to\infty} \|y_j(t) - y_i(t)\| = 0 \tag{6}$$

for system (5).

Let $\delta_i(t) \triangleq y_1(t) - y_i(t),\ i = 2, 3, \cdots, N$. One obtains (6) is equivalent to $\lim\limits_{t\to\infty} \|\delta_i(t)\| = 0,\ i = 2, 3, \cdots, N$.

Notice that

$$\begin{aligned} \dot{\delta}_i(t) =& C_1A_1C_1^\dagger\delta_i(t) - C_1B_1\sum_{j=1}^N a_{1j}K_{1j}\delta_j(t) \\ &+ C_iB_i\sum_{j=1}^N a_{ij}K_{ij}[\delta_j(t) - \delta_i(t)] \\ =& (C_1A_1C_1^\dagger - C_iB_i\sum_{j=1}^N a_{ij}K_{ij})\delta_i(t) \\ &+ \sum_{j=1}^N (C_iB_ia_{ij}K_{ij} - C_1B_1a_{1j}K_{1j})\delta_j(t) \end{aligned}$$

$i = 2, 3, \cdots, N$. Let

$$\delta(t) = [\delta_2^T(t)\ \delta_3^T(t)\ \cdots\ \delta_N^T(t)]^T.$$

One obtains

$$\dot{\delta}(t) = \bar{A}\delta(t), \tag{7}$$

where

$$\bar{A} = \begin{pmatrix} \bar{A}_{22} & \bar{A}_{23} & \cdots & \bar{A}_{2N} \\ \bar{A}_{32} & \bar{A}_{33} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \bar{A}_{(N-1)N} \\ \bar{A}_{N2} & \cdots & \bar{A}_{N(N-1)} & \bar{A}_{NN} \end{pmatrix},$$

$$
\begin{aligned}
&\bar{A}_{22} = C_1A_1C_1^{\dagger} - C_2B_2(\sum_{j=1}^{N} a_{2j}K_{2j}) - a_{12}C_1B_1K_{12},\\
&\bar{A}_{23} = a_{23}C_2B_2K_{23} - a_{13}C_1B_1K_{13},\\
&\bar{A}_{2N} = a_{2N}C_2B_2K_{2N} - a_{1N}C_1B_1K_{1N},\\
&\bar{A}_{32} = a_{32}C_3B_3K_{32} - a_{12}C_1B_1K_{12},\\
&\bar{A}_{33} = A_1 - C_3B_3(\sum_{j=1}^{N} a_{3j}K_{3j}) - a_{13}C_1B_1K_{13},\\
&\bar{A}_{(N-1)N} = a_{(N-1)N}C_{N-1}B_{N-1}K_{(N-1)N}\\
&\qquad - a_{1N}C_1B_1K_{1N},\\
&\bar{A}_{N2} = a_{N2}C_NB_NK_{N2} - a_{12}C_1B_1K_{12},\\
&\bar{A}_{N(N-1)} = a_{N(N-1)}C_NB_NK_{N(N-1)}\\
&\qquad - a_{1(N-1)}C_1B_1K_{1(N-1)},\\
&\bar{A}_{NN} = C_1A_1C_1^{\dagger} - C_NB_N(\sum_{j=1}^{N} a_{Nj}K_{Nj})\\
&\qquad - a_{1N}C_1B_1K_{1N}.
\end{aligned}
$$

In order to prove that the system (1) is output consensusable with respect to $\mathcal{U}_1$, it suffices to show that there exist matrices $K_{ij},\ i,j = 1,2,\cdots,N$, such that system (7) is stable which is equivalent to $\sigma(\bar{A}) \subseteq \mathbb{C}^-$.

In output consensus protocols (3), choose

$$K_{11} = K_{13} = K_{14}\cdots = K_{1N} = 0,$$

$$K_{ij} = 0,\ j \geq i,\ i = 2,3,\cdots,N.$$

Then $\bar{A}$ turns into

$$\bar{A} = \begin{pmatrix} \tilde{A}_{22} & 0 & \cdots & 0 \\ \tilde{A}_{32} & \tilde{A}_{33} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \tilde{A}_{N2} & \cdots & \tilde{A}_{N(N-1)} & \tilde{A}_{NN} \end{pmatrix},$$

where

$$
\begin{aligned}
&\tilde{A}_{22} = C_1A_1C_1^{\dagger} - a_{21}C_2B_2K_{21} - a_{12}C_1B_1K_{12},\\
&\tilde{A}_{32} = a_{32}C_3B_3K_{32} - a_{12}C_1B_1K_{12},\\
&\tilde{A}_{33} = C_1A_1C_1^{\dagger} - C_3B_3(a_{31}K_{31} + a_{32}K_{32}),\\
&\tilde{A}_{N2} = a_{N2}C_NB_NK_{N2} - a_{12}C_1B_1K_{12},\\
&\tilde{A}_{N(N-1)} = a_{N(N-1)}C_NB_NK_{N(N-1)},\\
&\tilde{A}_{NN} = C_1A_1C_1^{\dagger} - C_NB_N(\sum_{j=1}^{N-1} a_{Nj}K_{Nj}).
\end{aligned}
$$

In this case, system (7) is stable if and only if there exist $K_{12}$ and $K_{ij},\ j < i$ such that

$$\sigma\left(C_1A_1C_1^{\dagger} - a_{21}C_2B_2K_{21} - a_{12}C_1B_1K_{12}\right) \subseteq \mathbb{C}^-,$$

$$\sigma\left(C_1A_1C_1^{\dagger} - C_iB_i(\sum_{j=1}^{i-1} a_{ij}K_{ij})\right) \subseteq \mathbb{C}^-,\ i = 3,4\cdots,N.$$

Since condition (ii) holds, it is concluded from Lemma 3.1 that

$$
\begin{aligned}
&(C_1A_1C_1^{\dagger})^TX + X^T(C_1A_1C_1^{\dagger})\\
&\quad - X^T(C_1B_1 + C_2B_2)(C_1B_1 + C_2B_2)^TX + I_n = 0
\end{aligned}
\tag{8}
$$

has the unique admissible solution $P_2$ and

$$
\begin{aligned}
&(C_1A_1C_1^{\dagger})^TX_i + X_i^T(C_1A_1C_1^{\dagger})\\
&\quad - X_i^TC_iB_i(C_iB_i)^TX_i + I_n = 0
\end{aligned}
\tag{9}
$$

has the unique admissible solution $P_i,\ i = 3,4,\cdots,N$. Hence

$$\sigma\left(C_1A_1C_1^{\dagger} - (C_1B_1 + C_2B_2)(C_1B_1 + C_2B_2)^TP_2\right) \subseteq \mathbb{C}^-,$$

$$\sigma\left(C_1A_1C_1^{\dagger} - C_iB_i(C_iB_i)^TP_i\right) \subseteq \mathbb{C}^-,\ i = 3,4,\cdots,N.$$

Let

$$a_{12}K_{12} = a_{21}K_{21} = (C_1B_1 + C_2B_2)^TP_2 \tag{10}$$

and

$$\sum_{j=1}^{i-1} a_{ij}K_{ij} = (C_iB_i)^TP_i,\ i = 3,4,\cdots,N. \tag{11}$$

Under the condition (iii), it is obtained that equations (10) and (11) have at least one group solutions $K_{12}$ and $K_{ij},\ j < i,\ i = 2,3,\cdots,N$, such that

$$\sigma\left(C_1A_1C_1^{\dagger} - a_{21}C_2B_2K_{21} - a_{12}C_1B_1K_{12}\right) \subseteq \mathbb{C}^-,$$

$$\sigma\left(C_1A_1C_1^{\dagger} - C_iB_i(\sum_{j=1}^{i-1} a_{ij}K_{ij})\right) \subseteq \mathbb{C}^-,\ i = 3,4\cdots,N.$$

Combining the above results, one obtains multi-agent system (1) is output consensusable with respect to $\mathcal{U}_1$. ■

**Theorem** *3.2:* System (1) is output consensusable with respect to $\mathcal{U}_2$ if the following conditions hold:

(i) $C_1A_1C_1^{\dagger} = C_2A_2C_2^{\dagger} = \cdots = C_NA_NC_N^{\dagger}$;

(ii) $(C_1A_1C_1^{\dagger},\ C_1B_1 + C_2B_2)$ and $(C_1A_1C_1^{\dagger},\ C_iB_i)$, $i = 3,4,\cdots,N$, are stabilizable;

(iii) $a_{12} \neq 0$, $a_{21} \neq 0$ and there exists at least one $j \in \{1,2,\cdots,i-1\}$ such that $a_{ij} \neq 0,\ i = 3,4,\cdots N$.

*Proof:* Here, this proof is omitted since it is very similar to that of Theorem 3.1. ■

Based on Theorem 3.1, give the following algorithm to design the output feedback control law $K_{ij},\ i,j = 1,2,\cdots N$, which can guarantee that system (1) is output consensusable with respect to $\mathcal{U}_1$ under the precondition of Theorem 3.1.

**Algorithm** *3.1:* Input: the matrices $A_i \in \mathbb{R}^{n_i\times n_i}$, $B_i \in \mathbb{R}^{n_i\times r_i}$, $C_i \in \mathbb{R}^{m\times n_i}$, $i = 1,2,\cdots,N$, and $\mathcal{A} = [a_{ij}] \in \mathbb{R}^{N\times N}$;

Output: the gain matrix $K_{ij},\ i,j = 1,2,\cdots,N$.

(a) Compute Moore-Penrose inverse of $C_i$, and it is denoted by $C_i^{\dagger},\ i = 1,2,\cdots,N$;

(b) Solve Ricatti equations (8) and (9), and the admissible solutions are denoted by $P_2$ and $P_i,\ i = 3,4,\cdots,N$, respectively;

(c) Choose $K_{11} = K_{13} = K_{14} \cdots = K_{1N} = 0,\ K_{ij} = 0,\ j \geq i,\ i = 2, 3, \cdots, N$;

(d) Solve an arbitrary group of solutions of matrix equations (10) and (11), and denote them by $K_{12}$ and $K_{ij},\ j < i,\ i = 2, 3, \cdots, N$, respectively.

Thus, the required matrices $K_{ij},\ i, j = 1, 2, 3$, are designed.

## IV. A Numerical Example

A numerical example is provided to demonstrate the application of Algorithm 3.1. Consider linear multi-agent system (1) consisting of $N = 3$ heterogeneous agents with

$$A_1 = \begin{bmatrix} -2 & 0.5 \\ 1 & -1 \end{bmatrix},\ A_2 = \begin{bmatrix} 0.5 & 0 \\ -1 & 1 \end{bmatrix},$$

$$A_3 = \begin{bmatrix} 0 & 1 \\ -2 & 1 \end{bmatrix},\ B_1 = \begin{bmatrix} 0.6 & 1 \\ -5 & 1 \end{bmatrix},$$

$$B_2 = \begin{bmatrix} 4 & 1.5 \\ -1 & 0.9 \end{bmatrix},\ B_3 = \begin{bmatrix} 0.3 & -1 \\ 2 & 5 \end{bmatrix},$$

$$C_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},\ C_2 = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix},\ C_3 = \begin{bmatrix} 1 & 3 \\ 1 & 2 \end{bmatrix},$$

and the topology $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$, where $\mathcal{V} = \{1, 2, 3\}$,

$$\mathcal{E} = \{(1, 2),\ (1, 3),\ (2, 1),\ (2, 3),\ (3, 1),\ (3, 2)\}$$

and

$$\mathcal{A} = \begin{bmatrix} 0 & 0.5 & 2 \\ 0.2 & 0 & 3 \\ 1 & 0.4 & 0 \end{bmatrix}.$$

According to the steps in Algorithm 3.1, the following can be obtained:

(a)

$$C_1^{\dagger} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},\ C_2^{\dagger} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix},$$

$$C_3^{\dagger} = \begin{bmatrix} -2 & 3 \\ 1 & -1 \end{bmatrix};$$

(b)

$$P_2 = \begin{bmatrix} 0.2 & 0.1 \\ 0.1 & 0.2 \end{bmatrix},\ P_3 = \begin{bmatrix} 0.1 & 0.1 \\ -0.1 & 0.2 \end{bmatrix};$$

(c)

$$K_{11} = K_{13} = K_{22} = K_{23} = K_{33} = 0_{n \times n};$$

(d)

$$K_{12} = \begin{bmatrix} 0.9 & -1.3 \\ 1.0 & 1.1 \end{bmatrix},\ K_{21} = \begin{bmatrix} 2.2 & -3.2 \\ 2.5 & 2.7 \end{bmatrix},$$

$$K_{31} = \begin{bmatrix} 0.7 & 1 \\ -1 & 2 \end{bmatrix},\ K_{32} = \begin{bmatrix} -1.1 & -1.7 \\ 4.1 & -3.4 \end{bmatrix}.$$

Then the matrices $K_{ij},\ i, j = 1, 2, 3$, are designed.

In this case, the closed-loop system made of (1) and (2) is formulated:

$$\dot{x}(t) = A_c x(t), \tag{12a}$$

$$y(t) = Cx(t), \tag{12b}$$

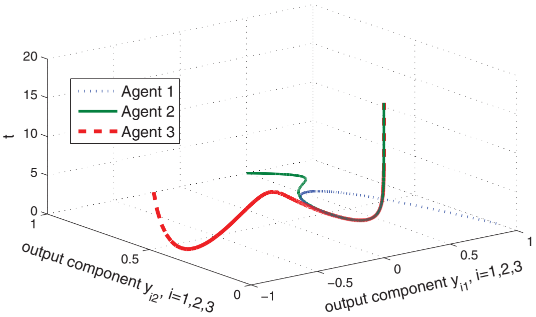


Fig. 1. the output responses of system (12)

where

$$A_c = \begin{bmatrix} A_{11} & a_{12}B_1K_{12}C_2 & a_{13}B_1K_{13}C_3 \\ a_{21}B_2K_{21}C_1 & A_{22} & a_{31}B_2K_{31}C_3 \\ a_{31}B_3K_{31}C_1 & a_{32}B_3K_{32}C_2 & A_{33} \end{bmatrix},$$

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix},\ y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix},\ C = \begin{bmatrix} C_1 & 0 & 0 \\ 0 & C_2 & 0 \\ 0 & 0 & C_3 \end{bmatrix},$$

$$A_{11} = A_1 - a_{12}B_1K_{12}C_1 - a_{13}B_1K_{13}C_1,$$

$$A_{22} = B_2D_2C_1A_1C_1^{\dagger}C_2 - a_{21}B_2K_{21}C_2 - a_{31}B_2K_{31}C_2,$$

$$A_{33} = B_3D_3C_1A_1C_1^{\dagger}C_3 - a_{31}B_3K_{31}C_3 - a_{32}B_3K_{32}C_3.$$

Choose the initial states of system (1) arbitrarily as follows:

$$x_1(0) = \begin{bmatrix} 1 \\ 0.1 \end{bmatrix},\ x_2(0) = \begin{bmatrix} 1.5 \\ 1 \end{bmatrix},\ x_3(0) = \begin{bmatrix} 3.4 \\ -1.2 \end{bmatrix}.$$

Denote

$$y_i = \begin{bmatrix} y_{i1} \\ y_{i2} \end{bmatrix},\ i = 1, 2, 3.$$

The simulation of system (12) is presented in Figure 1. It shows that the output responses of the system (12) indicates multi-agent system (1) is output consensusable with respect to the given admissible set $\mathcal{U}_1$.

## V. Conclusion

For linear multi-agent systems with fixed topologies and agents consisting of general linear systems, the conditions of output consensusability have been provided using generalized inverse, graph and linear system theory. The designed output feedback control law can guarantee that the studied multi-agent systems are output consensusable with respect to a given admissible set. Moreover, the simulation results have successfully demonstrated the applicability of the proposed approach in this paper.

The study of output consensus conditions for linear multi-agent systems with fixed topologies is a basic problem which only serves as a stepping stone to study more complicated agent dynamics. Future research will be on multi-agent systems with time delays, switching topology or time-varying topology, and so on.

# REFERENCES

[1] C. W. Reynolds, Flocks, herds, and schools: A distributed behavioral model, *in Proc. Comp. Graphics ACM SIGGRAPH'87 Conf.*, pp. 25–34, 1987.

[2] T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen and O. Schochet, Novel type of phase-transition in a system of self-driven particles, *Phys. Rev. Lett.,* vol. 75, no. 6, pp. 1226–1229, 1995.

[3] L. Xiao and S. Boyd, Fast linear iterations for distributed averaging, *Syst. Control Lett.,* vol. 53, pp. 65–78, 2004.

[4] P. Lin and Y. Jia, Robust $H_\infty$ consensus analysis of a class of second-order multi-agent systems with uncertainty, *IET Control Theory Appl.,* vol. 4, no. 3, pp. 487–498, 2010.

[5] A. Jadbabaie, J. Lin and A. Morse, Coordination of groups of mobile autonomous agents using nearest neighbor rules, *IEEE Trans. Autom. Control,* vol. 48, no. 6, pp. 988–1001, 2003.

[6] M. Kaya and R. Alhajj, Modular fuzzy-reinforcement learning approach with internal model capabilities for multiagent systems, *IEEE Trans. Syst., Man, Cybern. B, Cybern.,* vol. 34, no. 2, pp. 1210–1223, 2004.

[7] T. Li and J. Zhang, Consensus conditions of multi-agent system with time-varying topologies and stochastic communication noises, *IEEE Trans. Autom. Control,* vol. 55, no. 9, pp. 2043–2057, 2010.

[8] W. Ren, On consensus algorithms for double-integrator dynamics, *IEEE Trans. Autom. Control,* vol. 53, no. 6, pp. 1503–1509, 2008.

[9] P. Lin and Y. Jia, Further results on decentralized coordination in networks of agents with second-order dynamics, *IET Control Theory Appl.,* vol. 3, no. 7, pp. 957–970, 2009.

[10] W. W. Yu, G. R. Chen, M. Cao and J. Kurths, Second-order consensus for multiagent systems with directed topologies and nonlinear dynamics, *IEEE Trans. Syst., Man, Cybern. B, Cybern.,* vol. 40, no. 3, pp. 881–891, 2010.

[11] Z. Qu, J. Wang and R. Hull, Cooperative control of dynamical systems with application to autonomous vehicles, *IEEE Trans. Autom. Control,* vol. 53, no. 4, pp. 894–911, 2008.

[12] Z. K. Li, Z. S. Duan and G. R. Chen, Consensus of multiagent systems and synchronization of complex networks: A unified viewpoint, *IEEE Trans. Circuits Syst. I: Regular Papers,* vol. 57, no. 1, pp. 213–224, 2010.

[13] K. S. Barber, T. H. Liu and S. Ramaswamy, Conflict detection during plan integration for multi-agent systems, *IEEE Trans. Syst., Man, Cybern. B, Cybern.,* vol. 31, no. 4, pp. 616–628, 2001.

[14] M. Kaya and R. Alhajj, Fuzzy olap association rules mining-based modular reinforcement learning approach for multiagent systems, *IEEE Trans. Syst., Man, Cybern. B, Cybern.,* vol. 35, no. 2, pp. 326–338, 2005.

[15] X. R. Yang and G. P. Liu, Necessary and sufficient consensus conditions of descriptor multi-agent systems, *IEEE Trans. Circuits Syst. I: Regular Papers*, DOI: 10.1109/TCSI.2012.2190663.

[16] X. R. Yang and G. P. Liu, Consensus Conditions of Linear Descriptor Multi-agent Systems, *in Proc. 2nd International Conf. Intelligent Control and Information Processing*, pp. 69–774, 2011.

[17] Z. Qu, J. Chunyu and J. Wang, Nonlinear cooperative control for consensus of nonlinear and heterogeneous systems, *in Proc. 46th IEEE Conf. Decision Control*, pp. 2301–2308, 2007.

[18] H. Kim, H. Shim and J. H. Seo, Output consensus of heterogeneous uncertain linear multi-agent systems, *IEEE Trans. Autom. Control,* vol. 56, no. 1, pp. 200–206, 2011.

[19] R. E. Skelon, T. Iwasaki and K. Grigoriadis, *A United Algebra Approach to Linear Control Design*, Bristol, PA: Taylor and Francis, 1998.

[20] C. Q. Ma and J. F. Zhang, Necessary and sufficient conditions for consensusability of linear multi-agent systems, *IEEE Trans. Autom. Control,* vol. 55, no. 55, pp. 1263–1268, 2010.

[21] A. Ben-Israel and T. N. E. Greville, *Generalized Inverses: Theorey and Applications*, New York: John Wiley, 1974.

[22] Z. Y. Lin, B. Francis and M. Maggiore, Necessary and sufficient graphical conditions for formation control of unicycles, *IEEE Trans. Autom. Control,* vol. 50, no. 1, pp. 121–127, 2005.

[23] Z. Cheng and S. Ma, *Linear System Theory*, Bei Jing, China: Science Press, 2006.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Consensusability of Discrete-Time Linear Networked Multi-Agent Systems

Chong Tan
School of Astronautics
Harbin Institute of Technology
Harbin 150001, China
Email: tc20021671@126.com

Guo-Ping Liu
School of Astronautics
Harbin Institute of Technology, Harbin 150001, China
Faculty of Advanced Technology, University of Glamorgan
Pontypridd CF37 IDL, U.K.
Email: gpliu@glam.ac.uk

***Abstract*—The consensusability problem of discrete-time linear networked multi-agent systems with a communication delay is investigated in this paper. Based on the networked predictive control scheme and dynamic output feedback control, a novel protocol is proposed to compensate for communication delay actively. For discrete-time linear networked multi-agent systems with a directed topology and a constant network delay, necessary and/or sufficient conditions of consensusability with respect to a set of admissible consensus protocols are given. A simulation result demonstrates the effectiveness of theoretical results.**

***Index Terms*—Consensusability, networked multi-agent systems, networked predictive control.**

## I. Introduction

Recently, consensus problem has received significant attention as a fundamental research topic in decentralized control of networks of dynamic agents, due to its broad applications in cooperative control of unmanned aerial vehicles, scheduling of automated highway systems, formation control of satellite clusters, distributed optimization of multiple robotic systems, etc.

The most of existing works [1] about consensus problems focused on how consensus protocols are designed to achieve good performances. However, the consensusability problem of networked multi-agent systems is lack of enough attention, which is concerned about the existence of consensus protocols, and important in both synthesis and implementation of the protocols [2], [3].

Since the communication among agents in the networked multi-agent systems (NMASs) is achieved by a network, it is inevitable that the network-induced delay will occur while exchanging data among devices connected to the shared network, due to the limited bandwidth of the communication channels and the finite transmission speed. However, time delay can degrade the performance of control systems and even destabilize the system [4], [5]. Besides, due to economic costs or constraints on measurement in practice, it is often difficult or even unavailable to get the information of all the agents' states. Therefore, the consensusability problem of discrete-time linear networked multi-agent systems with a communication delay is investigated in this paper. By exploiting the networked predictive control scheme proposed by Liu [6], [7], a novel distributed protocol is put forward to overcome the effect of network delay actively rather than passively. For the NMAS consisting of uniform discrete-time linear time-invariant dynamical nodes with a directed topology and a constant network delay, delay-independent necessary and/or sufficient conditions of consensusability with respect to a set of admissible consensus protocols are established.

The paper is organized as follows. Some preliminaries of graph theory are briefly reviewed in Section II. Main results are given in Section III. To illustrate the theoretical results, a numerical simulation is provided in Section IV. Finally, Section V concludes the paper.

## II. Preliminaries

In this context, some necessary notations are introduced to make readers to easily understand. Let $\mathbb{R}$ and $\mathbb{C}$ be real and complex number fields, respectively. $M_{m,n}(\mathbb{F})$ denotes the set of all $m$-by-$n$ matrices over a field $\mathbb{F}$, and $M_{n,n}(\mathbb{F})$ is abbreviated to $M_n(\mathbb{F})$. For $A \in M_{m,n}(\mathbb{C})$, $A^{\mathrm{T}}$ denotes the transpose of $A$. Specially $m = n$, $A$ is said to be Schur if $\sigma(A) \subseteq U_0$, where $\sigma(A)$ represents the spectrum of matrix $A$, and $U_0$ denotes an open unit disk centered at the origin. $\otimes$ stands for the Kronecker product of matrices. $\mathbf{1}_N$ denotes a $N$-dimension column vector with all entries equal to one. 0 represents zero matrix with an appropriate dimension. $\|\cdot\|$ represents $l^2$ norm on vectors or its induced norm on matrices. "With respect to" is short for w.r.t..

First of all, some basic concepts and notations in graph theory are briefly introduce, which is very important and helpful in the analysis of NMASs. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$ be a weighted digraph of order $N$, where the set of nodes $\mathcal{V} = \{1, 2, \cdots, N\}$, set of edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, and a nonnegative weighted adjacency matrix $\mathcal{A} = [a_{ij}] \in M_N(\mathbb{R})$. An edge from $i$ to $j$ is denoted by $e_{ij} = (i, j)$ and adjacency element $a_{ji}$ associated with edge $e_{ij}$ is positive, i.e., $e_{ij} \in \mathcal{E} \Leftrightarrow a_{ji} > 0$. Moreover, assume that $a_{ii} = 0$ for all $i \in \mathcal{V}$. The set of neighbors of the node $i$ is denoted by $N_i = \{j \in \mathcal{V} : (j, i) \in \mathcal{E}\}$. The set of all reachable nodes to node $i$ is denoted by $N_i^*$. The Laplacian matrix $\mathcal{L} = [l_{ij}]_{n \times n}$ of weighted digraph $\mathcal{G}$ is defined as $\mathcal{L} = \mathcal{D} - \mathcal{A}$, where $\mathcal{D} = \mathrm{diag}(d_{in}(1), d_{in}(2), \cdots, d_{in}(N))$ and $d_{in}(i) = \sum\limits_{j=1, j \neq i}^{N} a_{ij}$, $i = 1, 2, \cdots, N$. Obviously, all the row-sums of $\mathcal{L}$ are zero, which implies that $\mathcal{L}$ has always an eigenvalue zero corresponding the right eigenvector $\mathbf{1}_N$. For a comprehensive restatement of the graph theory, the reader is referred to [8].

## III. Consensusability Based on the Networked Predictive Control Scheme

Consider an NMAS composed of $N$ agents, where the dynamics of agent $i$ are described by a discrete-time linear time-invariant system as follows:

$$\begin{aligned} x_i(t+1) = &\ Ax_i(t) + B_i u_i(t), \\ y_i(t) = &\ Cx_i(t), \\ i = &\ 1, 2, \cdots, N, \ t = 0, 1, 2, \cdots, \end{aligned} \tag{1}$$

where $x_i \in M_{n,1}(\mathbb{R})$, $u_i \in M_{m,1}(\mathbb{R})$ and $y_i \in M_{l,1}(\mathbb{R})$ are the state, control input and measured output of the agent $i$, respectively. $A \in M_n(\mathbb{R})$, $B_i \in M_{n,m}(\mathbb{R})$, $C \in M_{l,n}(\mathbb{R})$ are constant matrices.

Regarding the above $N$ agents as nodes of a graph, the communication relationship among agents can be conveniently represented by a weighted digraph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$ with the set of nodes $\mathcal{V} = \{1, 2, \cdots, N\}$, set of edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, and a nonnegative weighted adjacency matrix $\mathcal{A}$. The directed edge $e_{ij} \in \mathcal{E}$ means that agent $j$ can receive the information from agent $i$.

In this paper, it is considered that the information exchanged among all agents is achieved by a network with a time delay $\tau$. And the following assumptions can reasonably be made:

*Assumption 1:* (A1) Network delay $\tau$ is a constant and known positive integer.

(A2) States of all agents are not available but the outputs of them can be measured.

(A3) Each agent can receive information from itself and all reachable nodes to it, i.e. for $\forall j \in \{i\} \cup N_i^*$, agent $i$ can receive information from agent $j$, where $i \in \mathcal{V}$.

Because agent $i$ receives the information from agent $j$ ($j \in \{i\} \cup N_i^*$) with time delay $\tau$, in order to overcome the effect of the network delay, based on the output data of agent $j$ up to time $t - \tau$, the state predictions of agent $j$ from time $t - \tau$ to $t$ are constructed as:

$$\begin{aligned} \hat{x}_j(t-\tau+1|t-\tau) = &\ A\hat{x}_j(t-\tau|t-\tau-1) \\ &+ B_j u_j(t-\tau) + G_j\left[y_j(t-\tau)\right. \\ &\left.- C\hat{x}_j(t-\tau|t-\tau-1)\right], \end{aligned} \tag{2a}$$

$$\begin{aligned} \hat{x}_j(t-\tau+2|t-\tau) = &\ A\hat{x}_j(t-\tau+1|t-\tau) \\ &+ B_j u_j(t-\tau+1), \\ \hat{x}_j(t-\tau+3|t-\tau) = &\ A\hat{x}_j(t-\tau+2|t-\tau) \\ &+ B_j u_j(t-\tau+2), \\ &\vdots \\ \hat{x}_j(t|t-\tau) = &\ A\hat{x}_j(t-1|t-\tau) + B_j u_j(t-1), \\ &\ j \in \{i\} \cup N_i^*, \end{aligned} \tag{2b}$$

where $\hat{x}_j(t-\tau+1|t-\tau) \in M_{n,1}(\mathbb{R})$ and $u_j(t-\tau) \in M_{m,1}(\mathbb{R})$ are the one-step ahead state prediction and the input of the observer at time $t - \tau$, respectively, and $G_j \in M_{n,l}(\mathbb{R})$ can be designed using observer design approaches, $\hat{x}_j(t-\tau+d|t-\tau) \in M_{n,1}(\mathbb{R})$ is a state prediction of agent $j$ at time $t - \tau + d$ on the basis of information up to time $t - \tau$, and $u_j(t-\tau+d-1) \in M_{m,1}(\mathbb{R})$ is the input at time $t-\tau+d-1$, $d = 2, 3, \cdots, \tau$, $j \in \{i\} \cup N_i^*$.

For agent $i$ of NMAS (1), the following protocol based on the dynamic output feedback is designed:

$$\begin{aligned} z_i(t+1) &= \hat{A}_i z_i(t) + \hat{H}_i \zeta_i(t|t-\tau), \\ u_i(t) &= \hat{C}_i z_i(t) + \hat{F}_i \zeta_i(t|t-\tau), \\ i &= 1, 2, \cdots, N, \ t = 0, 1, 2, \cdots, \end{aligned} \tag{3}$$

where $z_i \in M_{\tilde{n},1}(\mathbb{R})$ is the protocol state, and

$$\zeta_i(t|t-\tau) = \sum_{j \in N_i} a_{ij}(\hat{y}_j(t|t-\tau) - \hat{y}_i(t|t-\tau)) \tag{4}$$

is the weighted sum of output prediction differences between agent $i$ and its neighboring ones, $\hat{y}_i(t|t-\tau) = C\hat{x}_i(t|t-\tau)$ is the output prediction of agent $i$ at time $t$ based on the output data of agent $i$ up to time $t - \tau$. $\mathcal{A} = [a_{ij}] \in M_{\scriptscriptstyle N}(\mathbb{R})$ is the weighted adjacency matrix of digraph $\mathcal{G}$. $\hat{A}_i$, $\hat{C}_i$, $\hat{H}_i$, $\hat{F}_i$ are matrices to be designed.

Let $u(t) = \left[\begin{array}{cccc} u_1^{\mathrm{T}}(t) & u_2^{\mathrm{T}}(t) & \cdots & u_{\scriptscriptstyle N}^{\mathrm{T}}(t) \end{array}\right]^{\mathrm{T}}$. The following admissible control set is considered:

$$\begin{aligned} \mathscr{U}^* = &\ \{u(t) : [0, +\infty) \to M_{\scriptscriptstyle Nm,1}(\mathbb{R}) \,\big|\, u_i(t) \text{ satisfies} \\ &\ (3), i = 1, 2, \cdots, N\}. \end{aligned} \tag{5}$$

*Definition 1:* NMAS (1) is said to be consensusable w.r.t. $\mathscr{U}^*$, if there exists a $u(t) \in \mathscr{U}^*$ such that for any initial value $x_i(0)$, $z_i(0)$ and $e_i(t)$, $t = -\tau, -\tau+1, \cdots, -1, 0$, $i \in \mathcal{V}$, the following conditions hold:

(1) $\lim\limits_{t \to \infty} \|x_i(t) - x_j(t)\| = 0, \ \forall\, i, j \in \mathcal{V}$,

(2) $\lim\limits_{t \to \infty} z_i(t) = 0, \ \forall\, i \in \mathcal{V}$,

(3) $\lim\limits_{t \to \infty} e_i(t) = 0, \ \forall\, i \in \mathcal{V}$,

where $e_i(t) = \hat{x}_i(t|t-1) - x_i(t)$ is the estimate error satisfying

$$\begin{aligned} e_i(t+1) = (A - G_i C)e_i(t), \ i \in \mathcal{V}, \\ t = -\tau, -\tau+1, \cdots, -1, 0, 1, \cdots. \end{aligned} \tag{6}$$

Let

$$\begin{aligned} \delta_i(t) &= x_1(t) - x_i(t), \ i = 1, 2, \cdots, N, \\ \delta(t) &= \left[\begin{array}{cccc} \delta_2^{\mathrm{T}}(t) & \delta_3^{\mathrm{T}}(t) & \cdots & \delta_{\scriptscriptstyle N}^{\mathrm{T}}(t) \end{array}\right]^{\mathrm{T}}, \\ x(t) &= \left[\begin{array}{cccc} x_1^{\mathrm{T}}(t) & x_2^{\mathrm{T}}(t) & \cdots & x_{\scriptscriptstyle N}^{\mathrm{T}}(t) \end{array}\right]^{\mathrm{T}}, \\ z(t) &= \left[\begin{array}{cccc} z_1^{\mathrm{T}}(t) & z_2^{\mathrm{T}}(t) & \cdots & z_{\scriptscriptstyle N}^{\mathrm{T}}(t) \end{array}\right]^{\mathrm{T}}, \\ e(t) &= \left[\begin{array}{cccc} e_1^{\mathrm{T}}(t) & e_2^{\mathrm{T}}(t) & \cdots & e_{\scriptscriptstyle N}^{\mathrm{T}}(t) \end{array}\right]^{\mathrm{T}}. \end{aligned}$$

For NMAS (1) with a fixed and directed topology, along with a constant network delay, a necessary and sufficient condition of consensusability of NMAS (1) w.r.t. $\mathscr{U}^*$ will be presented as follows.

*Theorem 1:* Consider NMAS (1) with a directed topology $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$ and a network delay $\tau$. NMAS (1) is consensusable w.r.t. $\mathscr{U}^*$ if and only if

$$(A, C) \text{ is detectable,} \tag{7}$$

and there exist $\hat{A}_i \in M_{\tilde{n}}$, $\hat{C}_i \in M_{m,\tilde{n}}$, $\hat{H}_i \in M_{\tilde{n},l}$ and $\hat{F}_i \in M_{m,l}$, $i = 1, 2, \cdots, N$, such that

$$\Theta \triangleq \begin{bmatrix} I_{N-1} \otimes A + RB_{D}\hat{F}_{D}(\mathcal{L}_2 \otimes C) & RB_{D}\hat{C}_{D} \\ \hat{H}_{D}(\mathcal{L}_2 \otimes C) & \hat{A}_{D} \end{bmatrix} \text{ is Schur,} \tag{8}$$

where $R = \begin{bmatrix} \mathbf{1}_{N-1} & -I_{N-1} \end{bmatrix} \otimes I_n$, $\mathcal{L}_2 = \mathcal{L} \begin{bmatrix} 0 & I_{N-1} \end{bmatrix}^{\mathrm{T}}$ and $\mathcal{L}$ is the Laplacian matrix of the graph $\mathcal{G}$.

*Proof:* From (2),

$$\begin{aligned} \zeta_i(t|t-\tau) = {} & C(\tilde{l}_i \otimes I_n)\delta(t) \\ & - CA^{\tau-1}(l_i \otimes I_n)e(t-\tau+1), \end{aligned} \tag{9}$$

where

$$\tilde{l}_i = \begin{bmatrix} l_{i2} & l_{i3} & \cdots & l_{iN} \end{bmatrix}$$

and

$$l_i = \begin{bmatrix} l_{i1} & \tilde{l}_i \end{bmatrix}, \ i = 1, 2, \cdots, N.$$

Substituting (9) into (3) derives

$$\begin{aligned} u_i(t) &= \hat{C}_i z_i(t) + \hat{F}_i \zeta_i(t) \\ &= \hat{C}_i z_i(t) + \hat{F}_i C(\tilde{l}_i \otimes I_n)\delta(t) \\ &\quad - \hat{F}_i CA^{\tau-1}(l_i \otimes I_n)e(t-\tau+1), \\ i &= 1, 2, \cdots, N. \end{aligned} \tag{10}$$

Hence, the closed-loop systems subjected to protocol (3) have the following forms:

$$\begin{aligned} x_i(t+1) &= Ax_i(t) + B_i u_i(t) \\ &= Ax_i(t) + B_i\hat{C}_i z_i(t) + B_i\hat{F}_i C(\tilde{l}_i \otimes I_n)\delta(t) \\ &\quad - B_i\hat{F}_i CA^{\tau-1}(l_i \otimes I_n)e(t-\tau+1) \end{aligned}$$

and

$$\begin{aligned} z_i(t+1) &= \hat{A}_i z_i(t) + \hat{H}_i \zeta_i(t) \\ &= \hat{A}_i z_i(t) + \hat{H}_i C(\tilde{l}_i \otimes I_n)\delta(t) \\ &\quad - \hat{H}_i CA^{\tau-1}(l_i \otimes I_n)e(t-\tau+1), \\ i &= 1, 2, \cdots, N. \end{aligned}$$

Then the following compact form can be presented:

$$\begin{aligned} x(t+1) &= (I_{N} \otimes A)x(t) + B_{D}\hat{C}_{D}z(t) \\ &\quad + B_{D}\hat{F}_{D}(\mathcal{L}_2 \otimes C)\delta(t) \\ &\quad - B_{D}\hat{F}_{D}[\mathcal{L} \otimes (CA^{\tau-1})]e(t-\tau+1) \end{aligned}$$

and

$$\begin{aligned} z(t+1) &= \hat{A}_{D}z(t) + \hat{H}_{D}(\mathcal{L}_2 \otimes C)\delta(t) \\ &\quad - \hat{H}_{D}[\mathcal{L} \otimes (CA^{\tau-1})]e(t-\tau+1). \end{aligned}$$

It should be noted that

$$\delta(t) = Rx(t)$$

and

$$e_i(t) = (A - G_iC)e_i(t-1), \ i = 1, 2, \cdots, N.$$

Therefore, the generalized closed-loop system can be described as

$$\xi(t+1) = \Omega\xi(t), \tag{11}$$

where

$$\begin{aligned} \xi(t) &= \begin{bmatrix} \delta^{\mathrm{T}}(t) & z^{\mathrm{T}}(t) & e^{\mathrm{T}}(t-\tau+1) \end{bmatrix}^{\mathrm{T}}, \\ \Omega &= \begin{bmatrix} \Omega_1 & RB_{D}\hat{C}_{D} & \Omega_2 \\ \hat{H}_{D}(\mathcal{L}_2 \otimes C) & \hat{A}_{D} & \Omega_3 \\ 0 & 0 & \Omega_4 \end{bmatrix}, \\ \Omega_1 &= I_{N-1} \otimes A + RB_{D}\hat{F}_{D}(\mathcal{L}_2 \otimes C), \\ \Omega_2 &= -RB_{D}\hat{F}_{D}[\mathcal{L} \otimes (CA^{\tau-1})], \\ \Omega_3 &= -\hat{H}_{D}[\mathcal{L} \otimes (CA^{\tau-1})], \\ \Omega_4 &= I_{N} \otimes A - G_{D}(I_{N} \otimes C). \end{aligned}$$

From Definition 1, NMAS (1) is consensusable w.r.t. $\mathscr{U}^*$ if and only if system (11) is asymptotically stable or, equivalently, $\Omega$ is Schur. So it follows from (7) and (8) that the conclusion holds. ■

Theorem 1 provides a necessary and sufficient condition of the consensusability of NMAS (1) w.r.t. $\mathscr{U}^*$. Based on it, a sufficient condition of the consensusability will be presented.

*Corollary 1:* Consider NMAS (1) with a directed topology $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$ and a network delay $\tau$. If there exist $\hat{A}_i \in M_{\tilde{n}}$ and $\hat{F}_i \in M_{m,l}$, $i = 1, 2, \cdots, N$, satisfying that the following conditions (i) – (iii) hold, then NMAS (1) is consensusable w.r.t. $\mathscr{U}^*$.

(i) $(A, C)$ is detectable.
(ii) $\hat{A}_i$ is Schur, $i = 1, 2, \cdots, N$.
(iii) $I_{N-1} \otimes A + RB_{D}\hat{F}_{D}(\mathcal{L}_2 \otimes C)$ is Schur,

where $R$ and $\mathcal{L}_2$ are defined in Theorem 1.

*Proof:* By choosing $\hat{C}_i$ to satisfy $B_i\hat{C}_i = 0$, or choosing $\hat{H}_i$ to satisfy $\hat{H}_iC = 0$, $i = 1, 2, \cdots, N$, it is sufficient that $\hat{A}_{D}$ and $I_{N-1} \otimes A + RB_{D}\hat{F}_{D}(\mathcal{L}_2 \otimes C)$ are Schur. Hence, from Theorem 1, NMAS (1) is consensusable w.r.t. $\mathscr{U}^*$. The proof is completed. ■

## IV. Simulation

In this section, a numerical simulation is presented to illustrate the effectiveness of the proposed theoretical results.

*Example 1:* Consider an NMAS with a network delay $\tau = 3$ and four agents indexed by 1, 2, 3 and 4, respectively. The dynamics of agent $i$ ($i = 1, 2, 3, 4$) are described by (1), where

$$A = \begin{bmatrix} 1 & 1.5 & 0.6 \\ -0.5 & -0.8 & 0.5 \\ 0 & 0.65 & 0.05 \end{bmatrix},$$

$$C = \begin{bmatrix} 1 & -1 & 1 \\ 2 & 1 & -1 \\ 1 & 1 & 1 \end{bmatrix},$$

$$B_1 = B_2 = B_3 = B_4 = \begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 1 & 0 \end{bmatrix}.$$

The interconnection among four agents is described by $\mathcal{G}$ in Fig. 1 with the adjacent matrix

$$\mathcal{A} = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

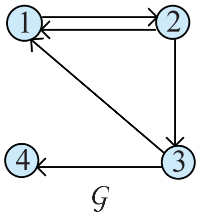


Fig. 1. Fixed topology.

It is obvious that $(A, C)$ is detectable. By choosing $Q = 6I_3$ and using MATLAB, a solution of Riccati equation

$$APA^{\mathrm{T}} - P - APC^{\mathrm{T}}(I + CPC^{\mathrm{T}})^{-1}CPA^{\mathrm{T}} + Q = 0$$

can be obtained as:

$$P = \begin{bmatrix} 7.1235 & -0.2844 & 0.4230 \\ -0.2844 & 6.2245 & -0.1377 \\ 0.4230 & -0.1377 & 6.2029 \end{bmatrix}.$$

Then a gain matrix of the observer (2a) can be got

$$G = \begin{bmatrix} -0.2752 & 0.1567 & 0.9671 \\ 0.2981 & -0.3275 & -0.1394 \\ -0.2913 & -0.0057 & 0.3157 \end{bmatrix}.$$

Take

$$\hat{A}_1 = \begin{bmatrix} -0.2 & 0 & 0 \\ 0 & 0.6 & 0 \\ 0 & 0 & 0.5 \end{bmatrix}, \ \hat{A}_2 = \begin{bmatrix} -0.3 & 0 & 0 \\ 0 & 0.8 & 0 \\ 0 & 0 & 0.1 \end{bmatrix},$$

$$\hat{A}_3 = \begin{bmatrix} 0.1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0.3 \end{bmatrix}, \ \hat{A}_4 = \begin{bmatrix} -0.5 & 0 & 0 \\ 0 & 0.2 & 0 \\ 0 & 0 & 0.85 \end{bmatrix},$$

$$\hat{C}_1 = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 4 & 1 \end{bmatrix}, \ \hat{C}_2 = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 2 \end{bmatrix},$$

$$\hat{C}_3 = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}, \ \hat{C}_4 = \begin{bmatrix} 0 & 0 & 0 \\ 0.5 & -2 & 1 \end{bmatrix},$$

$$\hat{H}_1 = \begin{bmatrix} -1.4980 & -1.2360 & 2.3740 \\ -3.2956 & -2.7192 & 5.2228 \\ 1.4980 & 1.2360 & -2.3740 \end{bmatrix},$$

$$\hat{H}_2 = \begin{bmatrix} 1.4980 & 1.2360 & -2.3740 \\ -2.3519 & -1.9405 & 3.7272 \\ -4.4940 & -3.7080 & 7.1220 \end{bmatrix},$$

$$\hat{H}_3 = \begin{bmatrix} -0.4494 & -0.3708 & 0.7122 \\ 0 & 0 & 0 \\ -1.4980 & -1.2360 & 2.3740 \end{bmatrix},$$

$$\hat{H}_4 = \begin{bmatrix} 0 & 0 & 0 \\ 3.5203 & 2.9046 & -5.5789 \\ -1.4980 & -1.2360 & 2.3740 \end{bmatrix},$$

$$\hat{F}_1 = \begin{bmatrix} -0.0997 & 0.0449 & -0.1473 \\ -0.1898 & 0.0896 & 0.0370 \end{bmatrix},$$

$$\hat{F}_2 = \begin{bmatrix} -0.0806 & 0.0453 & -0.1508 \\ -0.2257 & 0.1075 & 0.0389 \end{bmatrix},$$

$$\hat{F}_3 = \begin{bmatrix} -0.0935 & 0.0532 & -0.1715 \\ -0.0332 & 0.0140 & 0.2052 \end{bmatrix},$$

$$\hat{F}_4 = \begin{bmatrix} -0.3333 & 0.0667 & 0.4000 \\ 0.8333 & 0.3333 & -0.5000 \end{bmatrix}.$$

It is easy to verify that

$$\sigma(\Theta) = \{0.2, 0.6, 0.5, 0.3, 0.8, 0.1, 0.1, 0.2, 0.3, 0.85, \\ 0.3803, 0.2861, 0.0705, 0.6245, 0.6245, 0.8272, \\ 0.8272, 0.7703, 0.7703, 0, 0.5\} \subseteq U_0.$$

Hence, NMAS (1) is consensusable w.r.t. $\mathscr{U}^*$ by Theorem 1.

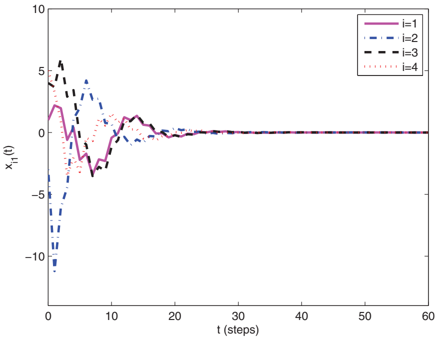


Fig. 2. Trajectories of the first state variables.

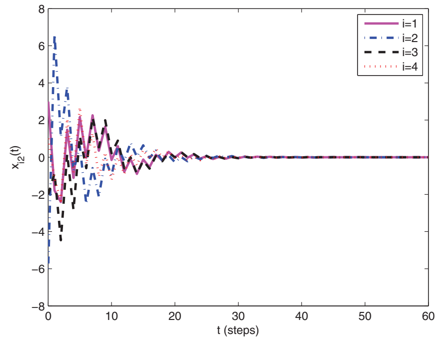


Fig. 3. Trajectories of the second state variables.

Initial conditions of NMAS (1), protocol (3) are chosen as

$$x_1(0) = \begin{bmatrix} 1 & 3 & -2 \end{bmatrix}^{\mathrm{T}}, \quad x_2(0) = \begin{bmatrix} -3 & -6 & 1 \end{bmatrix}^{\mathrm{T}},$$
$$x_3(0) = \begin{bmatrix} 4 & -2 & 2 \end{bmatrix}^{\mathrm{T}}, \quad x_4(0) = \begin{bmatrix} 5 & 1 & -1 \end{bmatrix}^{\mathrm{T}},$$
$$z_1(0) = \begin{bmatrix} -1 & 1 & 2 \end{bmatrix}^{\mathrm{T}}, \quad z_2(0) = \begin{bmatrix} 1 & 4 & 1 \end{bmatrix}^{\mathrm{T}},$$
$$z_3(0) = \begin{bmatrix} 4 & 6 & 2 \end{bmatrix}^{\mathrm{T}}, \quad z_4(0) = \begin{bmatrix} -10 & 5 & -1 \end{bmatrix}^{\mathrm{T}},$$

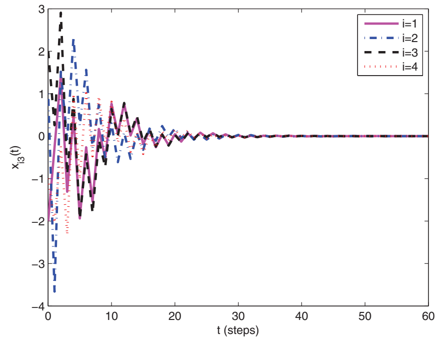


Fig. 4. Trajectories of the third state variables.

Initial conditions of estimate error (6) are chosen as

$$\begin{aligned}
e_1(0) &= \begin{bmatrix} 0.1 & -0.1 & -0.1 \end{bmatrix}^{\mathrm{T}},\\
e_2(0) &= \begin{bmatrix} 0.1 & 0.3 & 0.5 \end{bmatrix}^{\mathrm{T}},\\
e_3(0) &= \begin{bmatrix} 1 & 0 & -0.5 \end{bmatrix}^{\mathrm{T}},\\
e_4(0) &= \begin{bmatrix} 0.6 & 0 & -0.8 \end{bmatrix}^{\mathrm{T}},\\
e_1(-1) &= -\begin{bmatrix} 0.1 & 0.1 & 0.1 \end{bmatrix}^{\mathrm{T}},\\
e_2(-1) &= -\begin{bmatrix} 0.5 & 0.1 & 0.1 \end{bmatrix}^{\mathrm{T}},\\
e_3(-1) &= -\begin{bmatrix} 1 & 0.1 & 0.1 \end{bmatrix}^{\mathrm{T}},\\
e_4(-1) &= -\begin{bmatrix} 0.6 & 1 & 0.1 \end{bmatrix}^{\mathrm{T}},\\
e_1(-2) &= \begin{bmatrix} 0.1 & 0.3 & -0.1 \end{bmatrix}^{\mathrm{T}},\\
e_2(-2) &= \begin{bmatrix} -0.1 & 0.2 & 0.4 \end{bmatrix}^{\mathrm{T}},\\
e_3(-2) &= \begin{bmatrix} -0.6 & 0.5 & 0.1 \end{bmatrix}^{\mathrm{T}},\\
e_4(-2) &= \begin{bmatrix} 0.2 & 0.1 & -0.6 \end{bmatrix}^{\mathrm{T}}.
\end{aligned}$$

The state trajectories of NMAS (1) are shown in Figures 2–4, respectively, which demonstrates that states of NMAS (1) achieve consensus under protocol (3).

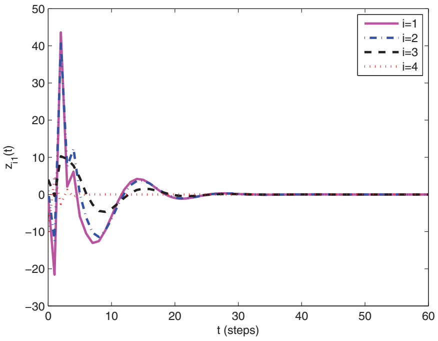


Fig. 5. Trajectories of the first protocol state variables.

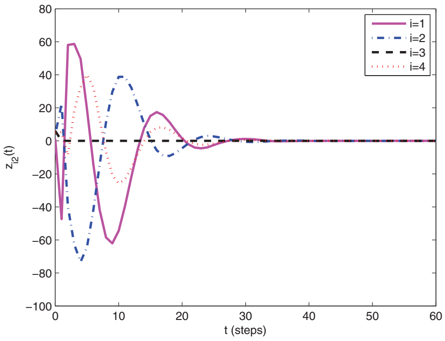


Fig. 6. Trajectories of the second protocol state variables.

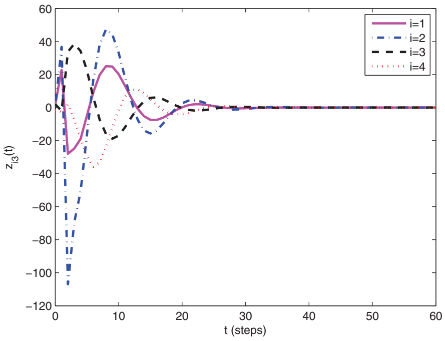


Fig. 7. Trajectories of the third protocol state variables.

The state trajectories of protocol (3) are shown in Figures 5–7, respectively. It is thus clear that protocol states asymptotically converge to zero.

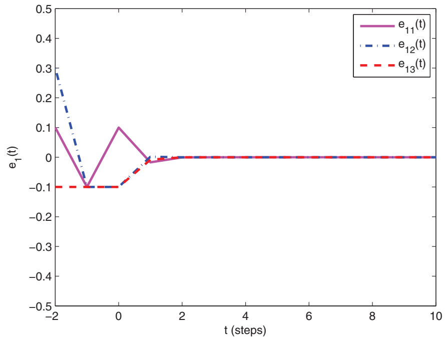


Fig. 8. The error trajectory $e_1(t)$.

The estimate error trajectories of observers are shown in Figures 8–11, respectively. It is obvious that states of observers track ones of NMAS (1).

## V. Conclusion

The consensusability problem of discrete-time linear NMASs with uniform dynamical agents and a communication delay has been investigated. A new distributed protocol is proposed by using the networked predictive control scheme. For discrete-time linear NMASs with a directed topology and

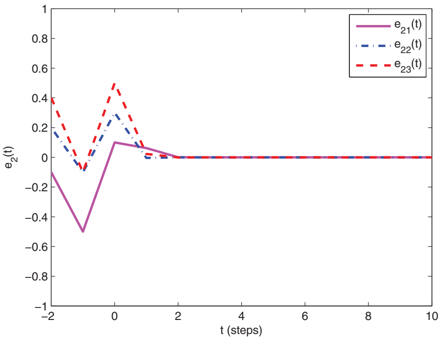


Fig. 9. The error trajectory $e_2(t)$.

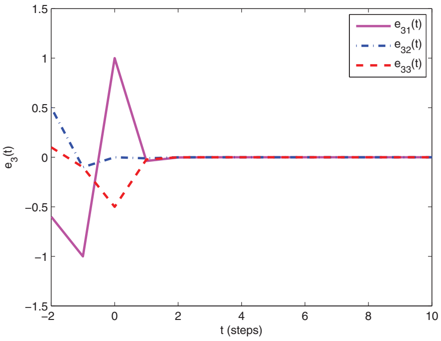


Fig. 10. The error trajectory $e_3(t)$.

a network delay, delay-independent necessary and/or sufficient criteria of consensusability have been obtained. A numerical example is provided to demonstrate the effectiveness of theoretical results.

## References

[1] R. Olfati-Saber and R. M. Murray, "Consensus problems in networks of agents with switching topology and time-delays," *IEEE Trans. Autom. Control*, vol. 49, no. 9, pp. 1520–1533, 2004.

[2] C. Q. Ma and J. F. Zhang, "Necessary and sufficient conditions for consensusability of linear multi-agent systems," *IEEE Trans. Autom. Control*, vol. 55, no. 5, pp. 1263–1268, 2010.

[3] K. Y. You and L. H. Xie, "Network topology and communication data rate for consensusability of discrete-time multi-agent systems," *IEEE Trans. Autom. Control*, vol. 56, no. 10, pp. 2262–2275, 2011.

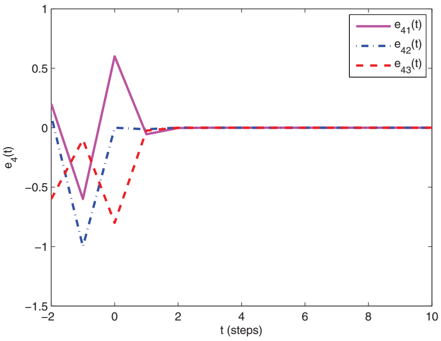


Fig. 11. The error trajectory $e_4(t)$.

[4] X. W. Liu, W. L. Lu, and T. P. Chen, "Consensus of multi-agent systems with unbounded time-varying delays," *IEEE Trans. Autom. Control*, vol. 55, no. 10, pp. 2396–2401, 2010.

[5] J. H. Qin, H. J. Gao, and W. X. Zheng, "Second-order consensus for multi-agent systems with switching topology and communication delay," *Syst. Control Lett.*, vol. 60, no. 6, pp. 390–397, 2011.

[6] G. P. Liu, Y. Q. Xia, D. Rees, and W. S. Hu, "Design and stability criteria of networked predictive control systems with random network delay in the feedback channel," *IEEE Trans. Syst. Man Cybern. Part C-Appl. Rev.*, vol. 37, no. 2, pp. 173–184, 2007.

[7] G. P. Liu, Y. Q. Xia, J. Chen, D. Rees, and W. S. Hu, "Networked predictive control of systems with random network delays in both forward and feedback channels," *IEEE Trans. Ind. Electron.*, vol. 54, no. 3, pp. 136–140, 2007.

[8] C. Godsil and G. Royle, *Algebraic Graph Theory*, ser. Graduate Texts in Mathematics. New York: Springer-Verlag, 2001, vol. 207.

UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

# Dynamical Analysis of a Duolever Suspension System

Ciro Moreno Ramírez
School of Engineering and
Mathematical Sciences,
City University
London, United Kingdom
CiroMoreno@city.ac.uk

M. Tomás-Rodríguez
School of Engineering and
Mathematical Sciences,
City University
London, United Kingdom
Maria.Tomas-Rodriguez.1@city.ac.uk

Simos A. Evangelou
Electrical and Electronic and
Mechanical Engineering Departments
Imperial College
London, United Kingdom
S.evangelou@imperial.ac.uk

*Abstract* — **The authors investigate the dynamical behaviour of a Duolever type of suspension on a standard sports motorcycle. The paper contains the modelling aspects of it, as well as the optimization process followed in order to obtain the suspension parameters and geometry arrangements. Head angle, wheelbase and normal trail are studied as indicators of the handling properties of the suspension system. Matlab optimization toolbox was used to design a mathematical model of a duolever front suspension system which keeps its normal trail constant during the full suspension travel. By using VehicleSim software, non-linear simulations were performed on motorcycle model that includes a duolever suspension. By a quasi-static variation of the forward speed of the motorcycle, the time histories of the system's states were obtained. The corresponded root locus to the linearized model were plotted and compared to those of the original motorcycle model without duolever system. A modal analysis was performed in order to get a deeper understanding of the different modes of oscillation and how the duolever system affects them. The results show that whilst a satisfactory anti-dive effect is achieved with this suspension system, it has a destabilizing effect on pitch and wobble modes.**

***Keywords- Modelling; motorcycle; weave; wobble; suspension; Hossack; Duolever***

## I. Introduction

One of the most important factors on motorcycle stability is the front end. It links the front wheel with the main frame and has two main functions: the suspension of the front wheel and the steering of the motorcycle. Up to this date several suspension systems have been developed to reach the best behaviour of the front end, being the telescopic fork the most extended one. The Hossack/Fior (marketed as Duolever), decouples the suspension and steering functions. One of its advantages is that it can be designed to achieve a desirable performance when suspension action takes place in terms of wheelbase, trail and head angle. The purpose of this paper is to study the effect of a Duolever suspension system on the dynamical properties of high performance motorcycles. Making use of Duolever's configurable properties in terms of wheelbase, head angle and trail, an eventual alternative front suspension is designed. This is done making use of the mathematical modelling and simulation of a motorbike. It will predict the behaviour of the various systems and help to decide which one is the most appropriate as base of the alternative front suspension system. The authors base this work on an existing high fidelity model of a Suzuki GSX-R1000, extensively used and validated in previous research (see [1], [2] and [3]),. The suspension system is designed by using algebraic methods to ensure as a first approach that similar properties and parameters to the original design are kept so that they can be compared under equal conditions. This is; similar head angle, trail, masses and inertia, etc. Later on, parameters such as mass or inertia will be varied -always within the limits of engineering constrictions- to study their influence on the motorcycle's dynamical properties.

The structure of this paper is as follows: Section II introduces the high-fidelity motorbike mathematical model which forms the basis of this work including a description of the modelling software VehicleSim. Section III contains an explanation on the Duolever system. Parametrization methodology, optimization of the parameters and suspension behaviour are also included. Section IV discusses on the oscillation modes and stability issues arising from the Duolever suspension. Finally, the results are discussed in section V and some future research ideas are presented.

## II. Model Description

The model used is based on an existing model of a Suzuki GSX-R1000 used in the past for several contributions in the field of motorcycle dynamics and stability analysis (see [4], [5], [6], [7], [8]). It consist of seven bodies: rear wheel, swinging arm, main frame (comprising rider's lower body, engine and chassis), rider's upper-body, steering frame, telescopic fork suspension and front wheel assembly. It involves three translational and three rotational freedoms of the main frame, a steering freedom associated with the rotation of the front frame relative to the main frame and spinning freedoms of the road wheels. The road tires are treated as wide, flexible in compression and the migration of both contact points as the machine rolls, pitches and steers is tracked dynamically. The tyre's forces and moments are generated from the tyre's

978-1-4673-1560-9/12/$31.00 ©2012 IEEE

camber angle relative to the road, the normal load and the combined slip using Magic Formulae models [9] and [10]. This model is applicable to motorcycle tires operating at roll angles of up to 60$^{o}$.The aerodynamic drag/lift forces and pitching moment are modelled as forces applied to the aerodynamic centre and they are proportional to the square of the motorcycle's forward speed. In order to maintain steady-state operating conditions, the model contains a number of control systems, which mimic the rider's control action. These systems control the throttle, the braking and braking distribution between the front and rear wheels, and the vehicle's steering. For a detailed description of the complete model the reader is referred to [3]. It has been developed using VehicleSim [11], it is a set of LISP macros, enabling the description of mechanical multi-body systems. The outputs from VehicleSim are a simulation program based on "C"  language with the implementation of the equations of motion and a Matlab [12] file containing the model's linear state-space equations. VehicleSim commands are used to describe the components of the motorcycle multi-body system in a parent-child relationship according to their physical constraints and joints. Once the VehicleSim code generates the simulation program, this is capable of computing general motions corresponding to specified initial conditions and external forcing inputs.

## III. Duolever Suspension System

Following the scheme of double wishbone car's suspension systems, the Duolever suspension for motorcycles consists of two wishbones, one upright and a steering linkage. The wishbones can rotate around transverse axes and the upright is now a fork in which the front wheel is attached. In the car version the wheel spins in a perpendicular axis due to the position of the system which is placed in the side of the car. In the bike case, the system is placed in the front, so the wheel has to be rotated 90 degrees with respect to the car wheel. The connection of the fork with the two wishbones is made by ball joints which allow the wishbones rotate and the fork turns in the steering axis. The steering axis is defined by the ball joints centres. The steering linkage connects the handlebar with the fork. It is a system of two levers, connected by an axis, which can be compressed or elongated in order to reach the length between the handlebar and the fork.  See [13] and [14] for more detailed information about Duolever systems. Fig. 1 shows a schematic CAD design for a standard motorcycle fitted with a Duolever system: the different structural points of the duolever and the parameters defining its geometry have been marked in red. The spring-damper unit has not been included to help a clearer view.

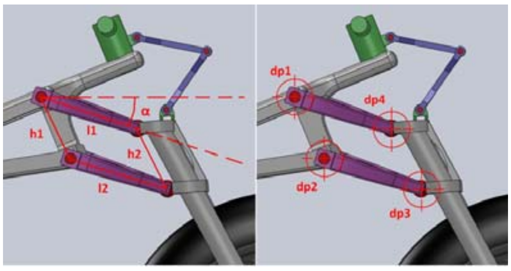


Figure 1. 3D kinematic components of a Duolever system. Parameters and points defining the Duolever geometry.

### A. Parametrization

The position of all the points is calculated in order to keep the model as close as possible to the configuration of the original motorbike described before. First of all, the parameters which must be considered in the design of the system the Duolever must be defined. These parameters are *l1*, *l2*, *h1*, *h2* and *α*. Where *l1* and *l2* are the lengths of the upper and lower wishbones, *h1* is the distance between the attachment points of the upper and lower wishbone, *h2* is the distance between the tips of the upper and lower wishbones and *α* is the nominal angle formed between the upper wishbone and the horizontal. With these parameters and the head angle the Duolever system is defined. The question is to find the attachment point to the main frame. To simplify this task the model of the motorbike is reduced to four main bodies: rear frame, front frame and two wheels. Two axes are considered: the rear-axis is the axis from the rear wheel attachment point to the point of attachment of the conventional front fork and, the front-axis, starting at this same point and forming the head angle with the vertical. The main points defined are:

*dp1: attachment point of upper wishbone in main frame.*

*dp2: attachment point of lower wishbone in main frame.*

*dp3: tip of the lower wishbone.*

*dp4: tip of the upper wishbone.*

*dp5: spring-damper unit in lower wishbone.*

*dp6: spring-damper unit in main.*

*pts: point located at the origin of the twist body in GSX-R1000 model when telescopic fork suspension was used. Now it is an auxiliary point located at the same position.*

In order to not modify the steering axis of the original model, *dp3* and *dp4* should be located on the front-axis and *dp1* is placed in the rear-axis to keep the delta-box configuration. Fig. 2 shows these points in the geometrical model.

### B. Optimization

#### 1) Suspension behaviour:

There exist four main parameters that mainly affect motorcycles´ handling. These are the wheelbase, the head angle, the trail and the normal trail. Wheelbase is the distance between the front wheel contact point and the rear wheel contact point. The head angle is the angle existing between the steering axis and the vertical axis. The trail is the distance between the front wheel contact point and the intersection of the steering axis with the road's plane. Finally, the normal trail is the distance between the front wheel contact point to the steering axis; it depends directly on the head angle and is just a perpendicular projection of the trail:

$$ntrail = trail \cdot cos(H_{ang})$$

For a Duolever suspension system the behaviour of the trails, wheelbase and head angle under suspension actuation depends on its design.

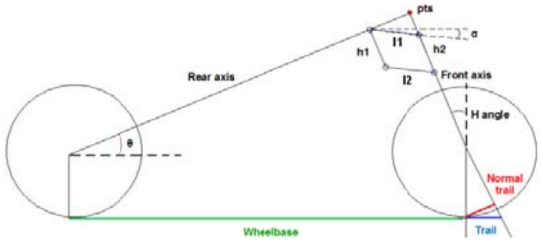


Figure 2. Points and angles defining the models geometry. Trail, head angle and wheelbase shown.

Whilst for a telescopic fork suspension system it is not possible to modify this behaviour, in a Duolever system the parameters *l1*, *l2*, *h1*, *h2* and *α* can be optimized in order to modify it. Fig. 3 shows this concept. For the (simplest) case of a parallelogram structure, the same result as for the conventional fork can be obtained due to the steering axis remains parallel to its initial direction along the full suspension travel.

*2) Parameters variations:*

The starting point for the optimization of the Duolever is to study the variation of the wheelbase and the normal trail with the suspension action. A geometrical model has been implemented on Matlab so that it allows tracking the eventual position of all the points in the assembly motorbike-Duolever along with the suspension travel, once the nominal position is known. The nominal position of the points *dp1*, *dp2, dp3* and *dp4* are given by the values of geometrical parameters. Considering the geometrical limitation of the motorcycle under study we took as a good approach the following values: *l1=170mm*, *l2=170mm*, *h1=120mm*, *h2=120mm* and *α=0.1rads*. A variation between the maximum and minimum *α* values (which have been calculated in order to produce an equivalent displacement of the motorbike as the conventional fork does) is performed, obtaining the geometrical position of all the points along the suspension travel. In Fig. 5 it is shown (dashed blue line) the behaviour of the wheelbase and the normal trail with the vertical suspension travel using the nominal set of parameters. Finally, in order to see how this behaviour changes according to Duolever parameters' variation, an external function has been developed. It takes an initial parameters vector and varies in a loop the parameter selected by the user. This loop calculates and stores the values of wheelbase and normal trail along the suspension travel for every value of the parameter varied. As an example, the 3D meshes representing the results obtained from the variation of geometrical parameter *α*, *l1* and *l2* are shown in Fig. 4. An *xz* reference axis that shows the nominal configuration is included in the figures.

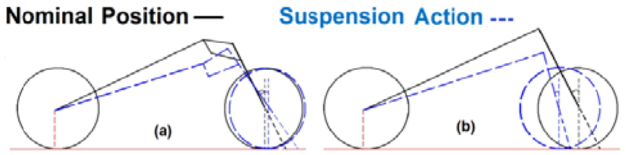


Figure 3. a) Duolever suspension system, head angle trail and wheelbase increase with the travel of the suspension. b) Telescopic fork suspension. Head angle, trail and wheelbase decrease with the travel of the suspension.

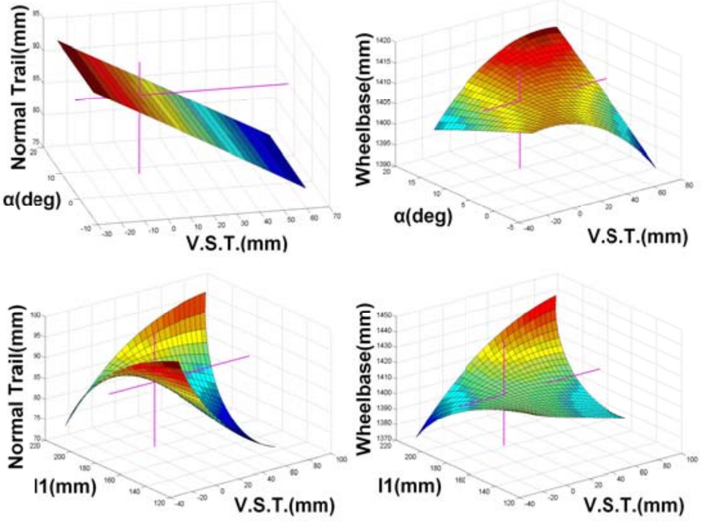


Figure 4. Behaviour variation with vertical suspension travel when parameters *α* and *l1* are modified.

As it can be seen, the variation with the parameters change is complicated enough to dissuade us to attempt a manual setting of them if we want to get constant trail or wheelbase. An automated optimization process is clearly needed to resolve this task.

*3) Optimization process:*

The goal is to find an optimal Duolever's parameter set such that the front suspension keeps the normal trail (so the trail and head angle) as constant as possible. A target function is defined and minimized by using Matlab optimization toolbox. This target function is the maximum difference -for a full suspension travel- between the nominal normal trail and the new normal trail depending on the set of parameters.

$$target = max(abs(ntrail-ntrail0))$$

Fig. 5 shows the behaviour of the wheelbase and normal trail for the optimized set of parameters in a solid green line. The values of these parameters are *l1=171mm*, *l2=182mm*, *h1=105mm*, *h2=124mm* and *α=0rad*. The nominal values of normal trail and wheelbase are plotted in dotted red lines. It can be seen how the lines for the optimized and standard Duolever cross each other at the initial value of the normal trail and the wheelbase but then their behaviour change completely. It is clear that the optimized set of parameter reduces almost to zero the variation in normal trail of the Duolever system.

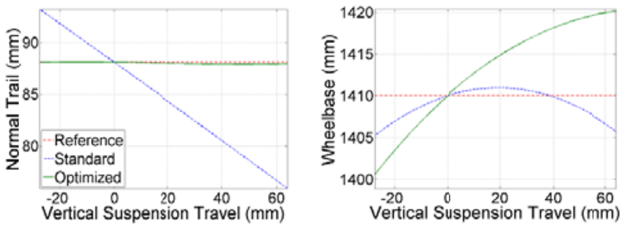


Figure 5. Wheelbase and normal trail behaviour with the vertical suspension travel for the geometrical standard and optimized models.

However, because we have used as target function the variation of the normal trail, the wheelbase is not kept constant but increases with the vertical suspension travel. Nevertheless this variation represents less than 1% and is not considered to be representative enough.

## IV. Oscillatory Modes and Stability Issues

One possible consequence of introducing new features and geometry changes in a motorcycle is that the stability of the system can be severely compromised. Motorbikes are nonlinear, oscillating complex systems that can represent a risk if they are not well damped. The modes under study are wobble and weave (see [4] for more details on these two modes). In this section the stability of the motorcycle model fitted with a Duolever type of suspension is analyzed by means of root locus diagrams in a similar manner as previous works such as [5], [6], [7] and [8]. Once the model with the optimized Duolever is built in VehicleSim, nonlinear simulations under different running conditions are performed and the linearized state space matrices of the system are fed with the nonlinear simulations states values in order to study the evolution of the eigenvalues over the operating envelope. Fig. 6 represents the root locus for the GSX-R1000 model with a telescopic fork (red +), a standard (blue x) and an optimized (green o) Duolever suspension systems. The roll angle for these simulations is 0 degrees and the swept variable is the forward speed ranging from 10 up to 80m/s. The stability properties of three characteristic modes will be analyzed: weave, wobble and pitch.

There are two main differences between the Duolever and the telescopic fork root locus plot. The wobble mode becomes unstable at medium speeds when a Duolever suspension is fitted in the model. Also, it can be seen that a "new" eigenvalue appears. This corresponds to the pitch mode. In the case of the telescopic fork it did not appear as it was well damped and greatly displaced on the left hand side of the complex plane. It lightly differs from the optimized and the standard Duolever models.

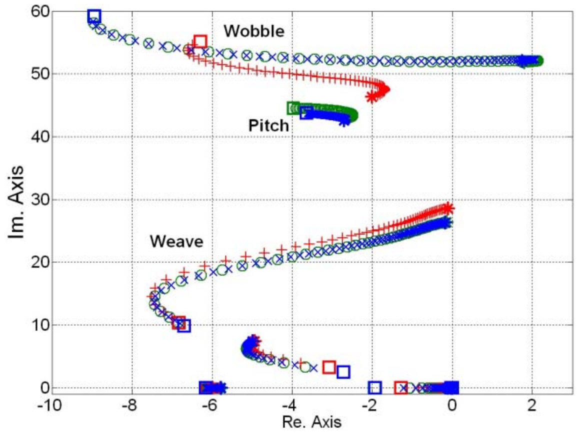


Figure 6. Root locus of the motorbike fitted with a telescopic fork (red+), a standard (blue x) and an optimized Duolever (green o) for 0 degrees of roll angle and a speed going from 10 (squares) up to 80 m/s (stars).

In order to find what eigenvalue it was and where it came from, the eigenvectors of the model with the Duolever system were compared with the eigenvectors of the model with telescopic fork system. As the eigenvalues for the optimized and standard Duolever models are very similar the comparison between eigenvectors has been done only for the optimized Duolever model. The comparison in Fig. 7 shows the modulus of the components of the eigenvectors. Only the generalized speeds are shown on the bar diagram. On the left side for each component, the value for the optimized Duolever is shown in green and on the right; the correspondent value to the telescopic fork is shown in red. The components of each eigenvector are labelled as follows:

| | |
|---|---|
| *XT, YT, ZT:* | *Translation of main body.* |
| *XR, YR, ZR:* | *Rotation of main body (Roll, Pitch, Yaw).* |
| *RSP and FSP:* | *Compression of rear and front springs.* |
| *RW and FW:* | *Rotation of rear and front wheel.* |
| *UBR:* | *Rotation of riders´ upper body.* |
| *STR:* | *Rotation of steer axis.* |
| *TWS:* | *Rotation of twist axis.* |

It has to be noted that the Duolever mathematical model was defined without flexibility, that means that it will have not twist degree of freedom, hence this coordinate will only appear for the telescopic fork model. It can be seen that there exist a high symmetry between the modes of the telescopic fork and the optimized Duolever models. There is a similar pattern in their components except for the twist generalized coordinate. For the Duolever case the twist degree of freedom has not been included, leaving this for the next step of this research.

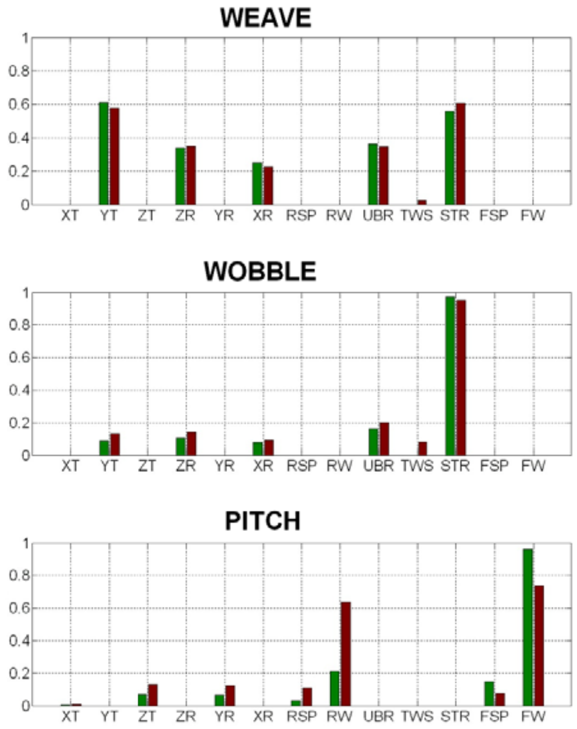


Figure 7. Eigenvector components for weave, wobble and pitch modes. Duolever suspension in green on the left, telescopic fork in red on the right.

Weave and wobble are out-of-plane modes. For both Duolever and telescopic fork cases, it is shown that the contribution of the various degrees of freedom to their eigenvectors is similar. On the other hand, pitch is an in-plane mode, the oscillation takes place in the symmetry plane of the motorbike, but for the Duolever case, the front wheel contribution becomes more relevant than in the fork suspension case whilst the contribution of the rear wheel is less. Also the front suspension coordinate increases its relevance and rear suspension decreases it. Finally the amplitude for the rotation in *y* and translation *z* (which implies the pitching of the main body) is reduced. Considering this, we can think of an oscillation about the front wheel which cannot be damped effectively by the front suspension. In order to check this, several simulations have been performed introducing various values of front tire damping coefficient. Fig. 8 shows these results for various values of damping. The weave and wobble modes appear as in Fig. 6 for both the telescopic and Duolever cases. The pitch mode appearing for the Duolever case changes according to various values of front tyre damping coefficient.

In the light of results shown in Fig. 8, it can be seen how the Duolever suspension does not damp pitch oscillations as effectively as the fork suspension does. This is a consequence of the Duolever's geometry and the anti-dive effect that it provides, reason why a Duolever suspension does not dive whilst performing braking action. The front assembly has a main role in the motorbike dynamic and in the case of the Duolever model its design becomes relevant for the pitch mode. In order to illustrate this, a straight running, front wheel braking simulation was carried out. The vertical suspension travel is shown in Fig. 9.a and the pitch rotation of the main body is shown in Fig. 9.b. The force applied in the front brake was calculate to provide the same deceleration of 1.5m/s$^2$ for all the three cases: telescopic fork (red), standard (blue) and optimized Duolever (green).

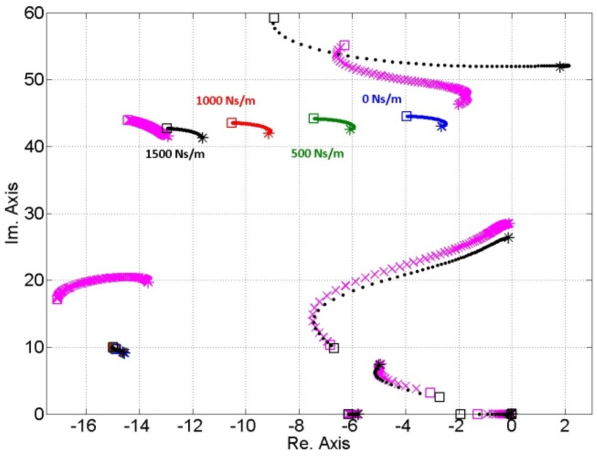


Figure 8. Root locus for the model of the motorcycle fitted with a telescopic fork (magenta +) and an optimized Duolever (blue ▪, green ▪, red ▪ and black ▪) for 0 degrees of roll angle and speed being swept from 10 (squares) up to 80m/s (stars). The damping of the front wheel is varied from 0 Ns/m up to 1500 Ns/m.

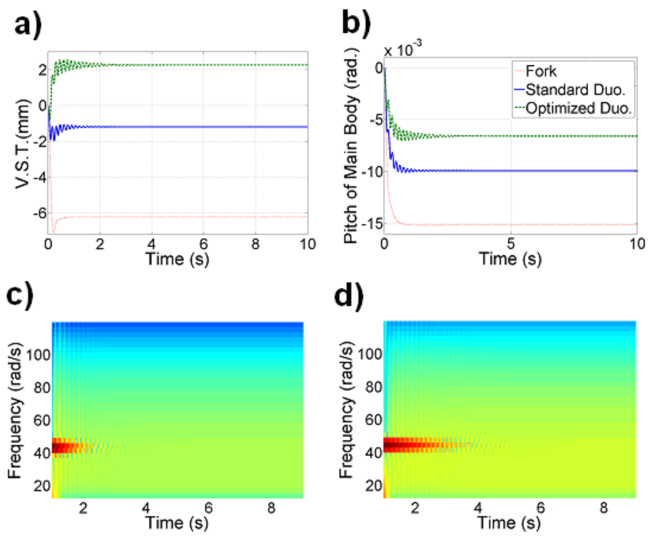


Figure 9. a) Vertical Suspension Travel for a braking simulation of 1.5m/s$^2$, b) Pitch of the Main Body for a braking simulation of 1.5m/s$^2$, c) Spectrogram of the Vertical Suspension Travel for the standard Duolever model during the braking simulation, d) Spectrogram of the Vertical Suspension Travel for the optimized Duolever model during the braking simulation.

It can be seen how whilst the front fork dives about 6mm, the standard Duolever does it only less than 1.2mm and the optimized Duolever does not dive but rises about 2.2mm. This behaviour appears due to the Duolever geometry which was optimized to get a constant trail. Other effect seen in this figure is the oscillation for Duolever systems, being higher and larger in time for the optimized one. In order to get a better understanding a spectrogram of the signal was done. It was used a 2 seconds time window with an overlapping of 99% to get good compromise between time and frequency resolutions. The low and high frequency components were neglected in this plot. The results are shown in Fig. 9.c for the standard and Fig. 9.d for the optimized Duolever. Due to the size of the window (2 secs.) the spectrograms for the first and the last seconds cannot be displayed. However, in both plots, oscillations about 43rad/s can be clearly recognized, they propagate reducing their amplitudes until they disappear. It can be seen how for the optimized Duolever the oscillation is sustained up to 4 seconds, whilst for standard Duolever model it disappears about 2.5 seconds.

The root locus plots showed that the frequency of the pitch mode is around 43rad/s at 80m/s, which is the initial speed of the motorcycle in the braking simulation case. This mode that becomes less stable with the Duolever front suspension system is prone to affect the behaviour of the motorcycle, representing a handicap for these type of suspension systems.

From these simulations it is clear to see that fitting a Duolever suspension system produces instability in the wobble mode. Wobble mode depends mainly on three factors that need to be taken into account: the mass and inertia of the front assembly and the damping ratio of the steering damper. If a high value of damping ratio is used, a more stable steering will be found at high speeds but it will be much less manoeuvrable at low speeds. Also, as it has shown in [6], increasing the steering damping coefficient the weave mode becomes less stable. Several commercial motorcycles with telescopic fork suspensions include steering dampers whose damping coefficients are variable with the speed. At the moment, the authors are investigating the possible benefits of including a speed dependant steering damper in the case of a Duolever suspension type. These results will be presented in a separate report.

## V. Conclusions

The mathematical model used for this study corresponds to a Suzuki GSX-R1000. This motorbike is not fitted with a Duolever, it is designed to make use of a telescopic fork. The mathematical model was modified with a carefully designed new suspension system model based on reasonable assumptions. Some dynamical properties about this type of suspension system have been studied.

The Duolever suspension can be designed in order to get a determined behaviour of the wheelbase, the head angle or the trail and the normal trail. In this study, a configuration which provides a constant normal trail along all the suspension travel for a Duolever system was obtained.

In general, a Duolever suspension system provides an anti-dive effect due to tyre's contact patch curvilinear trajectory. One of the consequences of the optimization of the Duolever is the increased anti-dive effect that appears compared to the standard Duolever suspension with a parallelogram design.

The anti-dive effect would represent in most cases beneficial characteristics but, in terms of oscillating behaviour, the pitch mode becomes clearly less damped compared to the case of standard telescopic fork suspension, representing in this way a possible risk issue under certain running conditions.

The advantages of the Duolever suspension system are meant to be the comfort, the manoeuvrability and the better performance of the front suspension, keeping the trails almost constant for all the suspension travel and presenting a relevant anti-dive effect. This allows the suspension to be fully operative on braking. However, it has been shown that less stable pitch modes are associated to this system.

It has also been shown how after including this suspension system in the model of a motorcycle which has not been designed to fit this type of suspensions, the wobble mode becomes unstable at high roll angles and medium-moderate speeds. In order to get wobble stability for the Duolever case, possibly a more complex steering damper unit depending on the speed should be design, or an inerter could be included. These possible solutions are currently under investigation by the authors.

## References

[1] S. Evangelou and D. J. N. Limebeer, “Lisp Programing of the ‘Sharp 1971’ Motorcycle Model,” 2000.

[2] S. Evangelou and D. J. N. Limebeer, “Lisp Programing of the ‘Sharp 1994’ Motorcycle Model,” 2000.

[3] R. S. Sharp, S. Evangelou, and D. J. N. Limebeer, “Advances in the Modelling of Motorcycle Dynamics,” *Multibody System Dynamics*, vol. 12, no. 3, pp. 251–283, 2004.

[4] D. J. N. Limebeer and R. S. Sharp, “Bicycles, motorcycles, and models,” *Control Systems, IEEE*, vol. 26, no. 5, pp. 34–61, 2006.

[5] S. Evangelou, D. J. N. Limebeer, R. S. Sharp, and M. C. Smith, “Control of motorcycle steering instabilities,” *Control Systems, IEEE DOI - 10.1109/MCS.2006.1700046*, vol. 26, no. 5, pp. 78–88, 2006.

[6] S. Evangelou and D. J. N. Limebeer, “Mechanical Steering Compensators for High-Performance Motorcycles,” *Journal of Applied Mechanics*, vol. Volume 74, no. Issue 2, p. 15, 2007.

[7] S. A. Evangelou, D. J. N. Limebeer, and M. Tomas-Rodriguez, “Suppression of burst oscillations in racing motorcycles,” in *Decision and Control (CDC), 49th IEEE Conference on*, 2010, pp. 5578–5585.

[8] S. Evangelou and M. Tomas-Rodriguez, “Influence of Road Camber on Motorcycle Stability,” *Journal of Applied Mechanics*,2008, pp. 231–236.

[9] H. B. Pacejka and E. Society of Automotive, *Tire and vehicle dynamics*. [Warrendale, PA ]: SAE International, 2006.

[10] E. J. H. de Vries and H. B. Pacejka, “MOTORCYCLE TYRE MEASUREMENTS AND MODELS,” *Vehicle System Dynamics*, vol. 29, no. sup1, pp. 280–298, 1998.

[11] http://www.carsim.com.

[12] http://www.mathworks.co.uk.

[13] Y. Watanabe and M. Sayers, “The Effect of Nonlinear Suspension Kinematics on the Simulated Pitching and Cornering Behavior of Motorcycles.” SAE Technical Paper, 2011.

[14] B. Mavroudakis and P. Eberhard, “Analysis of alternative front suspension systems for motorcycles,” *Vehicle System Dynamics*, vol. 44, no. sup1, pp. 679–689, 2006.