15th Triennial World Congress, Barcelona, Spain

LINEAR-WAVELET MODELS FOR SYSTEM IDENTIFICATION

Roberto K. H. Galvão*, Victor M. Becerra**

**ITA, Div. Eng. Eletrônica, São José dos Campos – SP, 12228–900, Brazil*
***University of Reading, Cybernetics Department, Reading RG6 6AY, UK*

Abstract: A model structure comprising a wavelet network and a linear term is proposed for nonlinear system identification. It is shown that under certain conditions wavelets are orthogonal to linear functions and, as a result, the two parts of the model can be identified separately. The linear-wavelet model is compared to a standard wavelet network using data from a simulated fermentation process. The results show that the linear-wavelet model yields a smaller modelling error when compared to a wavelet network using the same number of regressors. *Copyright ©2002 IFAC*

Keywords: Neural–network models, System identification, Nonlinear models, Function approximation, Non–parametric identification, Fermentation processes.

1. INTRODUCTION

The wavelet network is an approach for system identification in which nonlinear functions are approximated as the superposition of dilated and translated versions of a single function, which is localized both in the space and frequency domains (Zhang and Benveniste, 1992), (Zhang, 1997). Given a sufficiently large number of network elements (called "wavelets"), any square-integrable function can be approximated to arbitrary precision (Daubechies, 1992). The following advantages of wavelet networks over similar architectures have been cited in the literature (Cannon and Slotine, 1995):

(*i*) Unlike the sigmoidal functions used in neural networks of the multi–layer perceptron type, wavelets are spacially localized. As a result, training algorithms for wavelet networks typically converge in a smaller number of iterations when compared to multi–layer perceptrons.

(*ii*) The magnitude of each coefficient in a wavelet network can be related to the local frequency content of the function being approximated. Thus, more parsimonious architectures can be achieved, by preventing the unnecessary assignment of wavelets to regions where the function varies slowly. This property is not shared by networks of conventional radial basis functions (Gaussians).

(*iii*) Efficient construction algorithms have been devised to define the structure of the wavelet network (Zhang, 1997), (Kan and Wong, 1998) and even to adapt it in real time (Cannon and Slotine, 1995).

This paper proposes a model structure comprising a wavelet network and a linear term. It is shown that, under certain conditions, wavelets are orthogonal to linear functions, which allows both parts of the model to be identified separately.

This technique would be particularly useful to model systems which are only mildly nonlinear. Moreover, in situations where a nominal linear model is already available, a wavelet term can be added to account for nonlinearities in the plant.

An example using a simulated fermentation process is used to illustrate the proposed methodology. The results show that the linear-wavelet model results in a smaller prediction error than a pure wavelet model with the same complexity.

1.1 *Notation*

Scalars are represented in italic lowercase, vectors, in boldface lowercase and matrices, in boldface capitals. $\|\mathbf{x}\|$ is the Euclidian norm of $\mathbf{x}$. The ith element of $\mathbf{x}$ is denoted by x_i. The hat symbol $\hat{}$ indicates

an estimated value. The Fourier transform of f is denoted by Ff. When the limits of an integral are not indicated, it is assumed that they are $-\infty$ and $+\infty$. The inner product of two functions $f_1, f_2 : \mathbb{R}^d \to \mathbb{R}$ is given by $\langle f_1, f_2 \rangle = \int_{\mathbb{R}^d} f_1(\mathbf{x}) f_2(\mathbf{x}) dV_x$, where $dV_x = dx_1 dx_2 \cdots dx_d$ is a volume element in $\mathbb{R}^d$. $L^2(\mathbb{R}^d)$ is the space of functions that are square-integrable in $\mathbb{R}^d$, i.e., $L^2(\mathbb{R}^d) = \{f : \mathbb{R}^d \to \mathbb{R} \text{ s.t. } \int_{\mathbb{R}^d} |f(\mathbf{x})|^2 dV_x < \infty\}$.

2. WAVELET NETWORKS

A wavelet network can be regarded as a neural architecture with activation functions which are dilated and translated versions of a single function $\nu : \mathbb{R}^d \to \mathbb{R}$, where d is the input dimension (Zhang and Benveniste, 1992), (Zhang, 1997). This function, called "mother wavelet", is required to have zero mean (Daubechies, 1992). Additionally, it should be localized both in the space and frequency domains in the sense that $|\nu(\mathbf{x})|$ and $|F\nu(\omega)|$ rapidly decay to zero when $\|\mathbf{x}\| \to \infty$ and $\|\omega\| \to \infty$ respectively.

Mother wavelets for the multidimensional case (i.e., $d > 1$) can be constructed from one-dimensional functions $\psi : \mathbb{R} \to \mathbb{R}$ with fast decay in space and frequency. This can be done in several ways (Cannon and Slotine, 1995), (Zhang *et al.*, 1995), but for the purposes of this paper, the radial approach is adopted, that is, $\nu(\mathbf{x}) = \psi(\|\mathbf{x}\|)$. This is the choice made, for instance, in (Zhang, 1997).

A wavelet network model with L elements can be parameterized as

$$y(\mathbf{x}) = \sum_{j=1}^{L} w_j \nu_{a_j,\mathbf{b}_j}(\mathbf{x}) \tag{1}$$

where basis functions $\nu_{a_j,\mathbf{b}_j}$, called "daughter wavelets" (or simply wavelets), are dilated and translated versions of ν:

$$\nu_{a_j,\mathbf{b}_j}(\mathbf{x}) = a_j^{-d/2} \nu\left(\frac{\mathbf{x}-\mathbf{b}_j}{a_j}\right) \tag{2}$$

Dilation parameter $a_j \in \mathbb{R}^*$ controls the spread of the wavelet, while translation parameter $\mathbf{b}_j \in \mathbb{R}^d$ determines its central position. It can be shown (Daubechies, 1992) that, if pairs $(a_j, \mathbf{b}_j)$ are taken from the grid

$$\{(\alpha^m, \mathbf{n}\beta\alpha^m); m \in \mathbb{Z}, \mathbf{n} \in \mathbb{Z}^d\} \tag{3}$$

for convenient values of $\alpha > 1$ and $\beta > 0$ (typically $\alpha = 2, \beta = 1$), then any function in $L^2(\mathbb{R}^d)$ can be approximated by (1) to arbitrary precision, given a sufficiently large number of wavelets.

2.1 *Defining the structure of the wavelet network*

A major advantage of wavelet networks over other neural architectures is the availability of efficient construction algorithms for defining the network structure, that is, for choosing convenient values for $(m, \mathbf{n})$ in Equation (3). After the structure has been determined, weights w_j can be obtained through linear estimation techniques.

In this work, a constructive method similar to that introduced by (Zhang, 1997) is employed. It can be described as follows. Suppose that M modelling samples are available in the form of input-output pairs $(\mathbf{x}[k], y[k])$, $k = 1, \ldots, M$. Then:

1) Normalize the input data to fit within the effective support H of the mother wavelet. For radial wavelets, H is a hypersphere in $\mathbb{R}^d$ with radius R. For computational simplicity, H is approximated as a hypercube inscribed in the hypersphere with edges parallel to the coordinate axis.

2) Choose m_{min} and m_{max}, the minimum and maximum scale levels to be employed.

3) For each sample $\mathbf{x}[k]$ in the modelling set, find I_k, the index set of wavelets whose effective supports contain $\mathbf{x}[k]$:

$$\begin{aligned} I_k = \{&(m, \mathbf{n}) \text{ s.t. } \mathbf{x}[k] \in H_{m,\mathbf{n}}; \\ &m_{min} \le m \le m_{max}, \mathbf{n} \in \mathbb{Z}^d\} \end{aligned} \tag{4}$$

where $H_{m,\mathbf{n}}$ is a hypercube centered in $\mathbf{n}\beta\alpha^m$ with edges $\alpha^m R\sqrt{2}$.

4) Determine the pairs $(m, \mathbf{n})$ which appear in at least two sets I_{k1} and I_{k2}, $k_1 \neq k_2$. These are the wavelets whose effective support include at least two samples. This step is different from the algorithm described in (Zhang, 1997), which allows for wavelets with effective supports containing only one sample. In fact, such wavelets would introduce oscillations between neighbor modelling points, which might compromise the generalization ability of the model.

5) Let L be the number of wavelets obtained above. For simplicity of notation, replace the double index $(m, \mathbf{n})$ by a single index $j = 1, \ldots, L$.

6) Apply the L wavelets to the M modelling samples and gather the results in matrix form as

$$\mathbf{V} = \begin{bmatrix} \nu_1(\mathbf{x}[1]) & \nu_1(\mathbf{x}[2]) & \cdots & \nu_1(\mathbf{x}[M]) \\ \nu_2(\mathbf{x}[1]) & \nu_2(\mathbf{x}[2]) & \cdots & \nu_2(\mathbf{x}[M]) \\ \vdots & \vdots & \cdots & \vdots \\ \nu_L(\mathbf{x}[1]) & \nu_L(\mathbf{x}[2]) & \cdots & \nu_L(\mathbf{x}[M]) \end{bmatrix}_{L \times M} \tag{5}$$

Notice that each sample is now represented by L wavelet outputs (a column of $\mathbf{V}$).

If the M values of the output variable y are stacked in a row vector $\mathbf{y} = [y[1]\, y[2] \cdots y[M]]$ then least-squares regression can be used to estimate the row vector of network weights $\mathbf{w} = [w_1\, w_2 \cdots w_L]$ as

$$\hat{\mathbf{w}} = \mathbf{y}\mathbf{V}^T(\mathbf{V}\mathbf{V}^T)^{-1} \tag{6}$$

provided $\mathbf{V}\mathbf{V}^T$ is non-singular. If necessary, QR decomposition (Lawson and Hanson, 1974) or Principal Component Analysis (Naes and Mevik, 2001) can be used to deal with ill-conditioning.

Since many wavelets resulting from steps *1*) to *4*) may be redundant, a convenient subset of wavelets must be selected to improve model parsimony. Thus, the next step consists of determining which rows of $\mathbf{V}$ are the most relevant for the estimation task. For this purpose, rows from $\mathbf{V}$ are selected in a stepwise manner, starting from the one which displays the largest correlation with $\mathbf{y}$ and adding a new row at each iteration. This procedure can be described as follows.

a) Let $\mathbf{v}_j$ be the *jth* row of $\mathbf{V}$, that is, $\mathbf{v}_j = [v_j(\mathbf{x}[1])\ v_j(\mathbf{x}[2])\ \cdots\ v_j(\mathbf{x}[M])]$.

b) (Preliminary pruning) Eliminate all vectors $\mathbf{v}_j$ whose norm is smaller than a fixed threshold δ. Normalize all remaining vectors to unit norm.

c) (First selection) For each vector $\mathbf{v}_j$, evaluate the correlation index r_j as

$$r_j = \frac{|\mathbf{v}_j \mathbf{y}^T|}{\|\mathbf{v}_j\|\|\mathbf{y}\|} \tag{7}$$

Let $\mathbf{h}_1$ be the vector with the largest correlation index. Let also $i = 1$.

d) (Projections) Replace $\mathbf{y}$ and all vectors $\mathbf{v}_j$ by their projections onto the subspace orthogonal to $\mathbf{h}_i$, that is

$$\mathbf{v}_j \leftarrow \mathbf{v}_j(\mathbf{I} - \mathbf{P}_i) \quad \text{and} \quad \mathbf{y} \leftarrow \mathbf{y}(\mathbf{I} - \mathbf{P}_i) \tag{8}$$

where $\mathbf{P}_i = \mathbf{h}_i^T(\mathbf{h}_i\mathbf{h}_i^T)^{-1}\mathbf{h}_i$ and $\mathbf{I}$ is the identity matrix.

e) (Selection) For each vector $\mathbf{v}_j$, evaluate the index ρ_j defined as

$$\rho_j = r_j\|\mathbf{v}_j\| \tag{9}$$

Let $\mathbf{h}_{i+1}$ be the vector with the largest value for ρ_j.

f) Let $i = i+1$ and return to step *d*).

Index ρ_j used in step *e*) reflects both the amount of useful information in $\mathbf{v}_j$ (measured by r_j) and its lack of collinearity with the vectors already selected. In fact, if vector $\mathbf{v}_{j1}$ is highly collinear to vector $\mathbf{v}_{j2}$, then the projection of $\mathbf{v}_{j1}$ onto the subspace orthogonal to $\mathbf{v}_{j2}$ will have a small norm. Collinearity avoidance is important to achieve a model with good generalization ability (Naes and Mevik, 2001).

Statistical criteria such as minimum description length (Rissanen, 1978) and generalized cross-validation (Zhang, 1997) can be used to select the best number of wavelets to include in the model.

3. LINEAR-WAVELET MODELS

The model proposed here is of the form

$$y(\mathbf{x}) = f_1(\mathbf{x}) + f_2(\mathbf{x}) \tag{10}$$

where $f_1(\mathbf{x}) = \sum_{i=1}^{d} \theta_i x_i$, $\theta_i \in \mathbb{R}$ is a linear term and $f_2(\mathbf{x}) = \sum_{j=1}^{L} w_j v_{a_j,\mathbf{b}_j}(\mathbf{x})$ is a wavelet network.

It will be now be proved that f_1 is orthogonal to wavelets $v_{a_j,\mathbf{b}_j}$ and, as a result, to f_2. Initially, a proof will be given for the one-dimensional case, in which $v(x) = \psi(x)$. Then, the multidimensional case, in which $v(\mathbf{x}) = \psi(\|\mathbf{x}\|)$, will be considered. Notice that these proofs involve inner products de£ned for a continuous input $\mathbf{x}$. The orthogonality property can be extended for the sampled-data case provided the spatial sampling of each wavelet is suf£ciently £ne.

3.1 *One-dimensional case*

Definition 1. Function $\psi : \mathbb{R} \to \mathbb{R}$ is said to have n vanishing moments if

$$\int \psi(x)x^m dx = 0,\ m = 0, 1, ..., n-1. \tag{11}$$

Theorem 1. If $\psi : \mathbb{R} \to \mathbb{R}$ has two vanishing moments, then functions $\psi_{a,b}$ defined for $a \in \mathbb{R}^*$, $b \in \mathbb{R}$ as

$$\psi_{a,b}(x) = a^{-1/2}\psi\left(\frac{x-b}{a}\right) \tag{12}$$

are orthogonal to functions $f : \mathbb{R} \to \mathbb{R}$ of the form $f(x) = \theta x$, $\theta \in \mathbb{R}$.

Proof. Assume that ψ has two vanishing moments, that is

$$\int \psi(x)dx = 0 \tag{13}$$

$$\int \psi(x)xdx = 0 \tag{14}$$

The inner product of $\psi_{a,b}$ and f is then given by

$$\langle \psi_{a,b}, f \rangle = a^{-1/2} \int \psi\left(\frac{x-b}{a}\right)\theta x dx \tag{15}$$

By defining a new variable z as $z = (x-b)/a$, Equation (15) can be rewritten as

$$\langle \psi_{a,b}, f \rangle = a^{-1/2} \int \psi(z)\theta(az+b)adz =$$
$$\theta b a^{1/2} \int \psi(z)dz + \theta a^{3/2} \int z\psi(z)dz = 0 \tag{16}$$

where the two last terms equal zero due to equations (13) and (14). □

Notice that mother wavelets are already required to have at least one vanishing moment (Daubechies, 1992). The second vanishing moment is a property of many mother wavelets usually employed, such as the functions *dbN* of the Daubechies family for $N > 1$ (Daubechies, 1992), as well as any wavelet that is symmetrical around $x = 0$, such as the Mexican Hat function $\psi(x) = (1-x^2)e^{-0.5x^2}$.

It is worth noting that Gaussians employed in conventional radial basis functions networks do not satisfy (13), whereas sigmoidal functions used in multi–layer perceptron neural networks do not satisfy (14). As a result, if such network structures are used to synthesize f_2 in Equation (10), orthogonality to linear functions is not granted.

3.2 *Multidimensional Case*

Theorem 2. Let $v : \mathbb{R}^d \to \mathbb{R}$ be a radial mother wavelet obtained from $\psi : \mathbb{R} \to \mathbb{R}$ as

$$v(\mathbf{x}) = \psi(\|\mathbf{x}\|) \tag{17}$$

If the mean of v is zero, that is

$$\int_{\mathbb{R}^d} v(\mathbf{x}) dV_x = 0 \tag{18}$$

then wavelets $v_{a,\mathbf{b}}$ defined for $a \in \mathbb{R}^*$, $\mathbf{b} \in \mathbb{R}^d$ as

$$v_{a,\mathbf{b}}(\mathbf{x}) = a^{-d/2} v\left(\frac{\mathbf{x}-\mathbf{b}}{a}\right) \tag{19}$$

are orthogonal to functions $f : \mathbb{R}^d \to \mathbb{R}$ of the form $f(\mathbf{x}) = \sum_{i=1}^{d} \theta_i x_i, \theta_i \in \mathbb{R}$.

Proof. The inner product of $v_{a,\mathbf{b}}$ and f is given by

$$\langle v_{a,\mathbf{b}}, f \rangle = a^{-d/2} \int_{\mathbb{R}^d} v\left(\frac{\mathbf{x}-\mathbf{b}}{a}\right) \sum_{i=1}^{d} \theta_i x_i dV_x \tag{20}$$

By defining a new variable $\mathbf{z}$ as $\mathbf{z} = (\mathbf{x}-\mathbf{b})/a$, Equation (20) can be rewritten as

$$\langle v_{a,\mathbf{b}}, f \rangle = a^{d/2} \int_{\mathbb{R}^d} v(\mathbf{z}) \sum_{i=1}^{d} \theta_i (b_i + a z_i) dV_z =$$

$$a^{d/2} \sum_{i=1}^{d} \theta_i \left(b_i \int_{\mathbb{R}^d} v(\mathbf{z}) dV_z + a \underbrace{\int_{\mathbb{R}^d} z_i v(\mathbf{z}) dV_z}_{I} \right) \tag{21}$$

The first term inside brackets in the last line of Equation (21) vanishes due to the zero-mean hypothesis on v. Integral I in the second term can be evaluated as

$$I = \int_{\mathbb{R}^d} z_i \psi(\|\mathbf{z}\|) dV_z =$$

$$\iint \cdots \underbrace{\left[\int z_i \psi\left(\sqrt{\sum_{i=1}^{d} z_i^2} \right) dz_i \right]}_{I_i} \prod_{\substack{j=1 \\ j \neq i}}^{d} dz_j \tag{22}$$

By letting $C = \sum_{\substack{k=1 \\ k \neq i}}^{d} z_k^2$ and $F(z_i) = z_i \psi\left(\sqrt{z_i^2 + C}\right)$, integral I_i becomes

$$I_i = \int_{-\infty}^{\infty} F(z_i) dz_i \tag{23}$$

Since $F(z_i) = -F(-z_i), \forall z_i \in \mathbb{R}$, it follows that $\int_{-\tau}^{\tau} F(z_i) dz_i = 0, \forall \tau \in \mathbb{R}$ and, as a result, $I_i = 0$. Then, from (22), $I = 0$ and thus $\langle v_{a,\mathbf{b}}, f \rangle = 0$ □

The requirement of zero mean for v can be restated for the function ψ used in its generation. Suffice it to rewrite the integral in Equation (18) using the hyperspherical coordinates $(r, \gamma_1, \gamma_2, \ldots, \gamma_{d-1})$ defined as

$$x_1 = r \cos\gamma_1 \tag{24}$$

$$x_i = r \left(\prod_{j=1}^{i-1} \sin\gamma_j \right) \cos\gamma_i, \; i = 2, \ldots, d-1 \tag{25}$$

$$x_d = r \prod_{j=1}^{d-1} \sin\gamma_j \tag{26}$$

where $r > 0$, $0 \leq \gamma_i \leq \pi$, $i = 1, 2, \ldots, d-2$ and $-\pi \leq \gamma_{d-1} \leq \pi$. It follows that

$$\int_{\mathbb{R}^d} v(\mathbf{x}) dV_x = \int_{\mathbb{R}^d} \psi(\|\mathbf{x}\|) dV_x =$$
$$= \int_{\gamma_1=0}^{\pi} \int_{\gamma_2=0}^{\pi} \cdots \int_{\gamma_{d-1}=-\pi}^{\pi} \left(\int_{r=0}^{\infty} \psi(r) r^{d-1} dr \right)$$
$$\prod_{j=1}^{d-2} (\sin\gamma_j)^{d-j-1} d\gamma_1 d\gamma_2 \cdots d\gamma_{d-1} \tag{27}$$

A necessary and sufficient condition for the integral that spans the two last lines of Equation (27) to equal zero is

$$\int_{r=0}^{\infty} \psi(r) r^{d-1} dr = 0 \tag{28}$$

Example. Consider the unidimensional Mexican Hat function given by

$$\psi(x) = (c - x^2) e^{-0.5x^2} \tag{29}$$

where c is a parameter which needs to be adjusted to ensure that $\psi(\|\mathbf{x}\|)$ has zero mean. By introducing the above expression for $\psi(x)$ in (28), it follows that

$$\int_0^{\infty} (c - r^2) e^{-0.5r^2} r^{d-1} dr = 0$$
$$\Rightarrow c = \frac{\int_0^{\infty} e^{-0.5r^2} r^{d+1} dr}{\int_0^{\infty} e^{-0.5r^2} r^{d-1} dr} = \frac{I_{d+1}}{I_{d-1}} \tag{30}$$

By letting $\eta = e^{-0.5r^2}$ and $d\xi = r^{d-1} dr$, I_{d-1} can be integrated by parts, yielding

$$I_{d-1} = \eta\xi|_{r=0}^{\infty} - \int_{r=0}^{\infty} \xi d\eta =$$

$$\underbrace{\frac{e^{-0.5r^2} r^d}{d}\bigg|_{r=0}^{\infty}}_{0} + \frac{1}{d} \int_{r=0}^{\infty} r^{d+1} e^{-0.5r^2} dr = \frac{I_{d+1}}{d} \tag{31}$$

By using this result in (30), it follows that $c = d$.

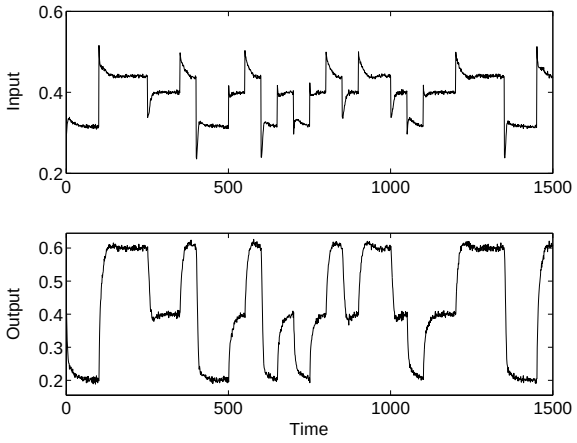

Fig. 1. Input and output data from the simulation.

4. NUMERICAL EXAMPLE

Consider a fermentation process described by the following Monod model (D'Ans *et al.*, 1972), (Aborhey and Williamson, 1978):

$$\frac{dC}{dt} = g\frac{CS}{S+p} - Cu \tag{32}$$

$$\frac{dS}{dt} = -qg\frac{CS}{S+p} + (S_{in} - S)u \tag{33}$$

where C = microbial concentration, S = substrate concentration (process output), u = dilution rate (process input), g = maximum growth rate, p = saturation parameter, q = yield factor, S_{in} = inlet substrate concentration. Values for the model constants were taken from (Zhang, 1997) as $g = 0.55$, $p = 0.15$, $q = 2$, $S_{in} = 0.8$. Suppose that S is observed at discrete time instants such that $y[k] = S(kT_s) + \varepsilon[k]$, where T_s is the sampling period and $\varepsilon[k]$ is the measurement noise.

The system was simulated in closed loop, with input u being provided by a PI controller with proportional gain $K_p = 0.5$ and integral gain $K_i = 0.05$. The set point for S was changed between three values: 0.2, 0.4 and 0.6. The measurement noise was simulated using a zero-mean white Gaussian noise process with standard deviation of 0.005. The sampling period adopted was $T_s = 1.0$ time unit. The resulting input ($u[k]$) and output ($y[k]$) signals can be seen in Figure 1.

The first 750 samples were employed for modelling, and the remaining data, for validation. The means of the input and output signals were removed during the identification procedures.

For the purpose of illustration, assume that it is desired to obtain a nonlinear ARX (autoregressive with exogenous input) model of the form

$$y[k] = f(y[k-1], y[k-2], u[k-1]) + e[k] \tag{34}$$

as proposed in (Zhang, 1997), where f is a nonlinear function to be estimated from the input-output data and $e[k]$ is the modelling residual. Notice that the input to the model is $\mathbf{x}[k] = [y[k-1]\, y[k-2]\, u[k-1]]^T$, so $d = 3$.

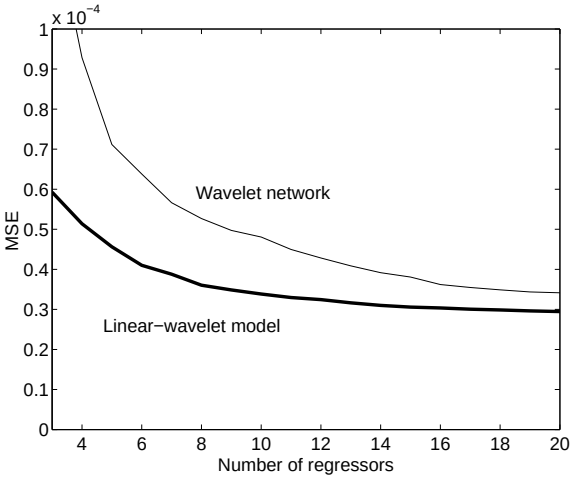

Fig. 2. Mean-square-error for the linear-wavelet model and the wavelet network.

To obtain a linear-wavelet model, a linear ARX model of the form

$$y[k] = \theta_1 y[k-1] + \theta_2 y[k-2] + \theta_3 u[k-1] + e^{lin}[k] \tag{35}$$

was initially identified by a standard least-squares procedure. A wavelet network was then built according to the method described in subsection 2.1, by using the residue of the linear identification $\mathbf{e} = [e^{lin}[1]\, e^{lin}[2] \cdots e^{lin}[750]]$ in the place of $\mathbf{y}$. A radial Mexican Hat mother wavelet obtained from (29) was employed. Its effective support can be taken as $R = 5$. The other parameters of the construction algorithm were adopted as $\alpha = 2$, $\beta = 1$, $m_{min} = 0$ and $m_{max} = 2$. Steps *1*) and *3*) resulted in 785 wavelets, a number which was reduced to 545 by Step *4*). In step *b*) of the selection process, 191 wavelets were discarded by using a norm threshold $\delta = 10^{-3}$.

A similar process was carried out to directly identify function f in (34) using a wavelet network.

Figure 2 compares the linear-wavelet and the wavelet network models in terms of the mean-square-error of modelling *MSE* defined as

$$MSE(n) = \frac{1}{M}\sum_{k=1}^{M}\left[y[k] - \hat{f}_n(\mathbf{x}[k])\right]^2 \tag{36}$$

where M is the number of modelling points and $\hat{f}_n$ is an estimate of f generated using n regressors. For the wavelet network, each wavelet corresponds to a regressor (a column of matrix $\mathbf{V}$ in (5)). In the case of the linear-wavelet model, the first three regressors are related to the linear part.

Figure 2 reveals that, for a given number of regressors (which indicate the complexity of the model), the linear-wavelet model yields a smaller mean-square-error than the wavelet network. Conversely, it can be stated that, for a given degree of approximation accuracy, the linear-wavelet model is more parsimonious than the wavelet network.

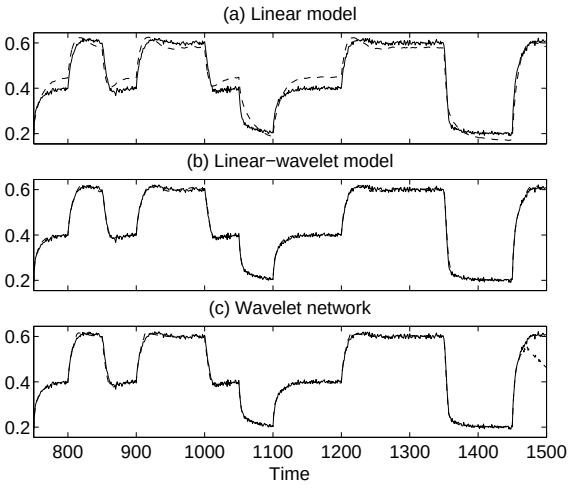

Fig. 3. Validation results. The model prediction and the actual plant output are represented by dashed and continuous lines respectively.

After identification, the models were used to predict the validation data recursively, that is, $\hat{y}[k] = \hat{f}(\hat{y}[k-1], \hat{y}[k-2], u[k-1])$, starting from the initial conditions $\hat{y}[751] = y[751]$, $\hat{y}[752] = y[752]$. For illustration, 20 regressors were employed for both the linear-wavelet model and the wavelet network.

Figure 3 displays the validation results for the linear model, the linear-wavelet model and the wavelet network. By comparing Figures 3a and 3b, it can be concluded that the use of wavelets improved the prediction ability of the linear model considerably. The advantage of the linear-wavelet model over the wavelet network becomes apparent at the end of the validation window, when the wavelet network prediction largely deviates from the actual process output.

5. CONCLUSION

This paper proposed a model structure comprising a wavelet network and a linear term. It was shown that, under certain conditions, linear functions are orthogonal to wavelets, which means that the two parts of the model can be identified separately. In the one-dimensional case, the mother wavelet is required to have two vanishing moments. For multidimensional wavelets of the radial type, suffice it to enforce the zero-mean condition. An example using a simulated fermentation process showed that the proposed model yields a better approximation than a conventional wavelet network using the same number of regressors.

The proposed technique can be used to improve the quality of existing linear models, by adding wavelet terms to account for nonlinearities. Conversely, the use of linear functions may help improve the generalization ability of a wavelet network, by providing interpolation over regions of the input space in which no modelling samples are available. One possible drawback is that the wavelet network will not be able to replace missing linear terms in the event that the linear part of the model has been under-parameterized.

Though not discussed here, linear-wavelet models could also be applied for data compression, particularly in the case of signals with linear trends. For such signals, the orthogonality property shows that linear de-trending can be followed by a wavelet identification of the residual oscillations.

Work is being carried out to use linear-wavelet models for predictive control. At each step, the linear term will be employed to generate an initial solution for the sequence of control movements. This solution will then be used as the starting point for an optimization algorithm that takes the whole model into account.

6. ACKNOWLEDGEMENT

The first author acknowledges the financial support of FAPESP under grant 00/09390-6.

7. REFERENCES

Aborhey, S. and D. Williamson (1978). State and parameter estimation of microbial growth processes. *Automatica* **14**(5), 493–498.

Cannon, M. and J.-J. E. Slotine (1995). Space-frequency localized basis function networks for nonlinear system estimation and control. *Neurocomputing* **9**, 293–342.

D'Ans, G., D. Gottlieb and P. Kokotovic (1972). Optimal control of bacterial growth. *Automatica* **8**, 729–736.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. SIAM. Philadelphia.

Kan, K.C. and K.W. Wong (1998). Self-construction algorithm for synthesis of wavelet networks. *Electronic Letters* **34**(20), 1953–1955.

Lawson, C. L. and R. J. Hanson (1974). *Solving Least Squares Problems*. Prentice-Hall. Englewood Cliffs.

Naes, T. and B. H. Mevik (2001). Understanding the collinearity problem in regression and discriminant analysis. *Journal of Chemometrics* **15**(4), 413–426.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica* **14**, 465–471.

Zhang, J., G. G. Walter, Y. Miao and W. N. W. Lee (1995). Wavelet neural networks for function learning. *IEEE Trans. Signal Processing* **43**(6), 1485–1496.

Zhang, Q. (1997). Using wavelet network in nonparametric estimation. *IEEE Trans. Neural Networks* **8**(2), 227–236.

Zhang, Q. and A. Benveniste (1992). Wavelet networks. *IEEE Trans. Neural Networks* **3**(6), 889–898.