NOISE REMOVAL IN A SINGLE SPEECH CHANNEL THROUGH CODING BY INDEPENDENT COMPONENT ANALYSIS

Allan Kardec Barros, Natália Abreu

Department of Electrica Engineering
Universidade Federal do Maranhão
São Luís-MA, Brazil

Abstract: Knowledge of the human brain's ability to process information is used in methods to extract a mixture of sounds. They simulate that ability, mainly, in the attempt of increasing the recognition rate and intelligibility. On the other hand, speech extraction can be achieved from one or from various channels. In this work, we propose to extract speech from a single channel, by handling speech features by means of the concept of efficient coding, which mimics the way that the auditory cortex code information. For that aim, we use independent component analysis (ICA). In order to show the efficiency of the method used in this research, simulations and results are presented.

Keywords: Cocktail Party Effort, Analysis of Independent Components and Single Speech Channel.

1. INTRODUCTION

In real environments, human speech usually happens in a context of multiple sonorous interferences; nevertheless, humans have the striking ability to filter the speech of a speaker from other mixed sounds. The term auditory scene analysis (ASA) was introduced to describe that process (Bregman, 1990). Among problems within ASA, one of the best known is the so called c*ocktail party* problem, which can be understood as a party, because in a party the frequency components of several signal sources (speech, music, air, noise, etc) reach a person's ear totally mixed, and yet with reverberation. Our task is to select one or more of those original signals and improve their intelligibility. Such a situation can be assessed by means of two approaches (Arons, 1992): those which use multi-channels measurements or using a single channel, being the latter the most difficult once the mutual interrelationship between channels can not be explored.

While monaural segregation remains a difficult challenge for computational system, humans show an impressive capacity for monaural segregation. The design of computational systems for monaural auditory scene analysis (CASA) is growing (Wang *et al.*, 1999), and a number of solutions to improve speech intelligibility and quality, with noise reduction, can be cited: Hu and Wang (Hu *et al.*, 2002) used the failure of most CASA systems to deal with high frequencies, and proposed a system for separating human speech grasping both, low and high frequencies. The signal was tested and assessed with various kinds of mixtures, being soon after compared with Wang-Brown's system (Wang *et al*., 1999), because of its similar characteristics. Besides that, there were also developed algorithms to improve a single channel using the discrete Fourier transform (DFT), as well as algorithms that use ICA (Barros *et al*., 2002) or algorithms of the subtractive type (Virag, 1999), based on the family of subtractive algorithms that explore the human hearing system masking properties.

The single channel subtractive-type speech enhancement has two major drawbacks: (1) the introduction of a musical residual noise with an unnatural structure in the enhanced speech; (2) a speech recognizing system cannot correctly segment noisy speech signal at very low SNR. Virag found the best trade-off for noise reduction, but she did not overcome the second problem. (Martin, 2001) proposed an algorithm to find out the variation of SNR for several environmental noises, but the algorithm also introduced a very annoying musical noise in the silent speech interstices. Combining (Virag, 1999) and (Martin, 2001), advantages, (Xiaoping, *et. al*., 2002) proposed an algorithm based in the minimum statistics to delimit the SNR variation, and explored the human hearing system masking properties to overcome the limitations of the improvement of subtractive type channel facing background noise added to a very low SNR.

In this paper, we propose the use of the available knowledge in the literature related with the manner in which humans process information to extract the

speech signal in a single channel by means of noise reduction and consequent speech recognition. *Masking* and *efficient coding* are the two characteristics of how the human brain uses to codify information. While the first occur at the cochlea level, it is assumed that the second happens at the auditory cortex level. In our proposal we use the latter approach.

Efficient coding helped to explain how the receptive fields in the primary visual cortex code information. Applying this theory to the auditory system, (Lewicki, 2002), used ICA for deriving efficient coding for different classes of natural sounds, including sounds originated from animals, the environment and human speech, proving that animal vocalizations are very alike to a Fourier transform, environmental, non biological sounds are alike the Wavelet transform, and the human speech, in its turn, seemed to be encoded as a mixture of those two classes.

Based upon Lewicki studies, we propose a neural network architecture that learns the basis functions of the input signals by ICA. This work is divided as follows: system outline, brief introduction on the principal component analysis (PCA), independent component analysis (ICA) and automatic speech recognition (ASR). After that, we show the simulations, the results and finally the conclusions.

2. THE METHOD

We assume here the observed signal $x(t)$ is encoded in a set of M responses $a(t) = [a_1(t)...a_M(t)]^T$. The goal efficient coding is to derive a vector of filters $\phi(t) = [\phi_1(t)...\phi_M(t)]^T$, or basis functions that minimizes the mutual statistical dependence between the responses. The observed signal can be written as follow:

$$x = a^T \phi \qquad (1)$$

An estimations of ϕ and a can be found by either principal component analysis (PCA) or ICA, once it assumes that any elements of ϕ are mutually uncorrelated or independent, respectively. Theoretically, an ideal coding transforms the signals input to allow the outputs be statistically independent, removing all redundancies.

2.1 Principal Component Analysis (PCA)

PCA has been widely used in efficient coding of human speech signals. PCA and ICA are forms of rotating a given vector. PCA, in particular, forces the elements of ϕ to be orthonormal functions and the a coefficients to be mutually orthogonal.

PCA maps one space into another, based on the correlation between elements, so that the structure of the new data corresponds to the "energy" involved in the process that generated the original data. The disadvantage of this strategy is the fact that it really rotates the axis correctly in the direction of the independent signals only in the case in which those signals have a Gaussian distribution. ICA appeared as a solution to this problem.

The PCA and ICA define their objective functions in complete different way. PCA uses only second order statistics, while it is impossible to perform the ICA by using only second order statistics. PCA emphasizes size reduction, while ICA can reduce the size, augment it or leave it as it is.

2.2 Independent Component Analysis (ICA)

ICA is an analysis method proposed by Jutten and Herault from 1985 on, and it is also an extension of the PCA. It has being widely studied in the context of blind source separation (BBS) from their linear mixtures, using statistical independency as criterion for the sources separation without knowing the mixing coefficients, nor the probable distribution of signal sources. (Karhunen, *et al.,* 1999*;* Lee, 1998; Freisleben, *et al*., 1997).

Much attention was paid to this method due to its potential in applications of signal processing, such as telecommunications, physiological signal processing and human speech recognition.

The estimation of the data model of independent component analysis is usually performed by formulating an objective function and an optimization algorithm.

ICA may be understood in the framework of probability distribution estimation. Let us first remember that the joint distribution of a set of independent observations $x(t) = [x_1, x_2...x_N]^T$ taken from the same distribution $p(x, \theta)$ is given by the product of the marginal distributions when the elements are mutually independents, i.e.

$$p(x, \theta) = \coprod_i^N p(x, \theta) \qquad (2)$$

Given the modifiable parameters $\hat{\theta}$, ICA algorithms should find an estimator of the true density $\hat{p}(x, \hat{\theta})$ where θ is the true parameter.

One way of estimating ICA is to find a transformation that minimizes the mutual information between components, where this mutual information is a natural measure of the dependency among random variables. The use of mutual information can be motivated by handling the Kullback-Lieber

divergence or by means of high order cumulants, such as the kurtosis.

After selecting an estimation principle for ICA, we need to find out a practical method to run it. And there are several algorithms for ICA with different characteristics, such as JADE (Cardoso, 1993), FastICA (Hyvarinen, *et al*., 1997) e RICA (Barros, 2001).

2.3 Automatic Speech Recognition (ASR)

Voice is a particular type of sounds produced by the larynx, while the speech demands, besides the production of speech, a clear sound articulation.

As technology advances so do the machines in almost all scenarios. Thus, nothing better than providing those equipments with the capacity of perception and understanding of the human speech, which is the simplest, natural and efficient way humans use to explain their thoughts and, in this way, humanize even more the relationship man-machine.

Automatic Speech Recognition (ASR) is the process by which linguistic information of speech signal is automatically extracted. Linguistic information contained in the speech signal is encoded as to make that the high degree of signal variability, caused by the environment and the speaker, practically do not interfere in the human perception of the information.

Speech recognition, however, is constrained by certain problems that make difficult its processing. The principal difficulties related with speech recognition can be summarized as follows:

1. The same word pronounced several times can produce different waveforms due to the articulation of organs of the speech apparatus;

2. Difficulties in the segmentation of speech: it does not exist exactly one way for limiting phonemes (the shorter speech unit), embarrassing the recognition of continued speech. This lack of accuracy of the limit arises from the large variation of speech signals and from the mutual interaction between them;

3. Variation in the speech characteristics: we can find acoustical, non linear differences in time rhythm, in timbre and in intensity;

4. With frequent misuse of linguistic knowledge: speech may not grasp all linguistic information (as for example: mistakes in Portuguese language and accent).

ASR systems can be subdivided, according to the words treatment, in: isolated words recognition, chained or continuous words, in which a silent interval must be considered corresponding to the interval between one and another pronounced word, to establish the difference in this classification. The ASR also considers the pronunciations which depend upon a particular speaker as well the ones that do not depend on any particular speaker, considering only what was uttered and not who uttered a word (Foster, *et al*., 1993).

The major problem found in the advancement of the technology for speaker recognition is the same found in the technology for speech recognition. There is a large variability intra–speaker within time due to health factors (respiratory illnesses, laryngitis, etc), stress, emotional factors, effort and speed of the speech. All these factors contribute to the starting up of significant changes in the training and testing parameters, turning null the recognition.

Many of the techniques used in the field of speech processing are also used in other areas of signal processing, as for example in transforms and filters. Nevertheless, the speech signal has special characteristics that lead to the usage of knowledge about the way in which humans produce and perceive it.

2.4 The Model

The model proposed here is not very different from in artificial neural networks (ANN). It is divided in a training stage (learning) and a working stage. The difference is that we do not use the non linearity of the neural model. The learning stage, shown in Fig.1, at the left-hand is where the system learns the basis functions through the ICA algorithm, and the second stage (running phase) is where the desired signal is computed by the mean square error (MSE) simple estimation.

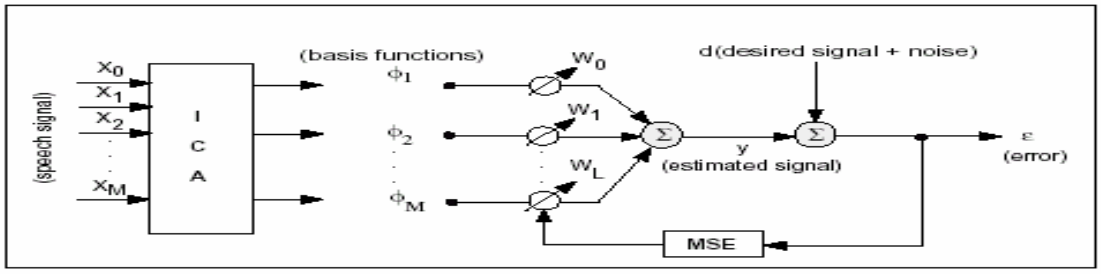

Fig. 1. System block diagram.

3. SIMULATIONS

Simulations consisted in using a set of speech signals, uttered numbers, from zero to nine, in Brazilian Portuguese language (zero, um, dois, três, quatro, cinco, seis, sete, oito, nove). A male speaker repeated each number ten times. Initially, we removed manually parts of each uttered word, preserving only the one containing the actual speech.

We tested the theory validity with different amounts of basis functions, by randomly extracting, from any part of the speech signals, a training vector of 200 length. We trained the ICA algorithm which is capable of introducing those vectors and took the basis functions with the highest Kurtosis.

We extracted 40, 60, 80, 100, 120 and 140 basis functions and computed the error estimation using the mean square error algorithm (MSE), to create another input randomly chosen, as described in the system block diagram. After 100 repetitions of the task, we calculated the mean error as shown in Fig. 3, according to the number of basis functions for both, the PCA and ICA, without adding noise to the system.

Different types of noise were used, Gaussian and non-Gaussian, to test the method's robustness in front of undesirable signals, and also performed the recovering of a signal through the base functions of another uttered signal.

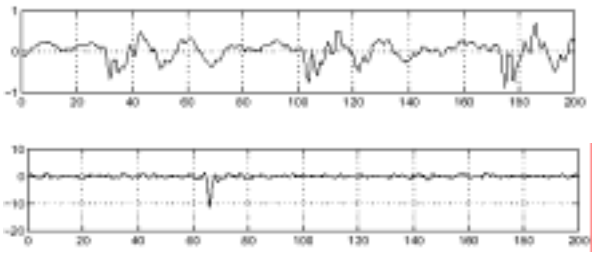

Fig. 2. Example of original signal and the obtained basis functions.

4. CONCLUSIONS

It is worth to notice some interesting points in this work: it has a strong biological basis, which is that efficient coding, besides a statistic basement. Those advantages are reflected in our simulation which showed that the PCA adds almost no benefit to speech recovering and is less affected by the increase in the number of basis functions. For some numbers the error is really higher with the increase of the basis. On the other hand, the ICA seems to be the more appropriate technique. We can easily see that there is an exponential decay as the number of basis functions is increased.

Our simulations confirm that this is a robust technique in front of interferences (different kinds of noise), and it was also possible to get the recovering of a speech signal, randomly selected, through speech signal basis functions the speaker used to utter another number.

In this work we showed a system based on efficient coding that encodes and performs noise reduction for the recovering of a desired signal by extracting basis functions by ICA. In addition, simulations and experimental results also showed that this technique appears as a promising technique to speech/speaker identification. As it holds biological evidences and a statistical background, we can imagine, for further studies, a speaker/speech recognition method from a different point of view: firstly design the system as to be robust against noise reverberation effects. When this is done, proceed to create the more complex blocks involving the language to be processed, challenging the rest of the systems available in the market, which, in the first place, are worried with the recognition problem instead of dealing with the interferences problem.

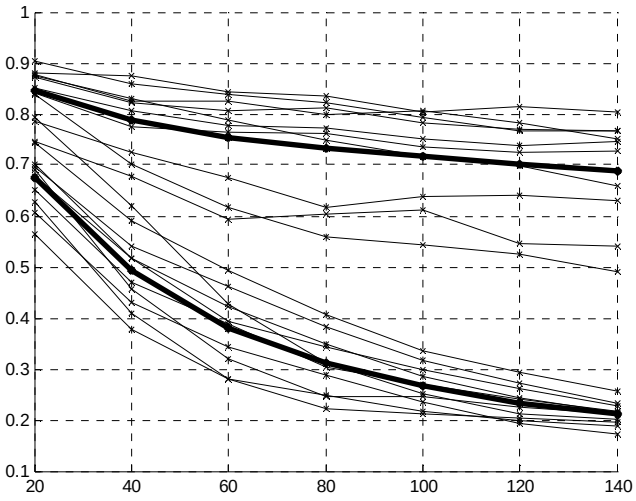

Fig. 3. Error versus number of basis functions for as different words pronounced by a male speaker that uttered numbers, in Brazilian Portuguese language, from zero to nine. The bold line is the average of each, for the PCA algorithm and for the ICA algorithm, without added noise.

REFERENCES

Arons, B., (1992). A Review of the Cocktail Party. Journal of the American Voice I/O Society. pp. 35-50.

Barros A. K. and Cichocki A. (2001). Extraction of specific signals with temporal structure (Neural Computation), Vol. 13, No. 9, pp. 1995-2004.

Barros, A. K., Rutkowski T., Itakura F., Ohnishi N. (2002). Estimation of Speech Embedded in a Reverberant and Noisy Environment by Independent Component Analysis and Wavelets. (IEEE Trans. on Neural Networks. (4)), Vol. 13 pp. 888-893.

Bregman, A.S., (1990). Auditory Scene Analysis, Cambridge, MA: MIT press.

Cardoso, J.-F. and Souloumiac, A. (1993). Blind beamforming for non gaussian signals. (IEEE Procedings) F, 140(6): pp. 362-370.

Foster, P. and Schalk T., (1993) Speech Recognition The Complete Practical Reference Guide, New York, Telecom Library.

Freisleben, B., Hagen C., and Borschbach M. (1997). Blind Separation of Acoustic Signals Using a Neural Network. Departament of Eletrical Engeneering and Computer Science, University of Siegen.

Hu, G. and Wang D. (2002). Monaural Speech Segregation Based on Pitch Tracking and Amplitude Modulation, IEEE

Hyvarinen, A. and Oja E., (1997). A fast fixed-point algorithm for independent component analysis. (Neural Computation (9)), 1483 - 1492.

Hyvarinen, A. Survey on Independent Component Analysis.

Karhunen, J., Vigãrio R., Hurri J. and Oja E. (1997). Applications of Neural Blind Separation to Signal and Image Processing. Helsink (University of Tecnology, Laboratory of Computer and Information Science).

Lee, T. W. (1998). Introduction to Independent Component Analysis. Computational Neurobiology Laboratory. (The Salk Institute).

Lewicki, M.S. (2002). Efficient coding of natural sounds. (Nature Neuroscience 5(4)): 356-363.

Martin, R. (2001). Noise power spectral density estimation based on optimal smoothing and minimum statistics. (IEEE Trans. Speech and Audio Processing), Vol 9, pp. 504-512.

Virag, N. (1999). Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System. (IEEE Trans. on Signal Processing (2)), Vol 7, pp. 126-137.

Wang, D. L. and Brown G. J. (1999). Separation of Speech from Interfering Sounds Based on Oscillatory Correlation. (IEEE Trans. Neural Network), Vol. 10, pp. 684-697.

Xiaoping, J., Hua F. and Tianren Y. (2002). A Single Channel Speech Enhancement Method Based on Masking Properties and Minimum Statistics. (IEEE Trans. on Signal Processing)