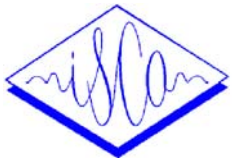

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

RECOGNIZING SIMULTANEOUS SPEECH: A GENETIC ALGORITHM APPROACH

Athanasios Koutras, Evangelos Dermatas and George Kokkinakis

WCL, Electrical and Computer Engineering Dept.
University of Patras, 26100 Patras, HELLAS.
e-mail:koutras@giapi.wcl2.ee.upatras.gr

10.21437/Eurospeech.1999-643

ABSTRACT

In this paper it is shown experimentally that a new blind signal separation method in the frequency domain improves significantly the speaker signal to interference ratio (SIR) and the phoneme recognition score of a continuous speech, speaker-independent acoustic decoder in a two-simultaneous-speaker environment. The implemented two-sensor separation method is based on evolutionary minimization of the cross-correlation of the separated speech signals. Extensive experiments have been conducted in three types of artificially created mixture scenarios: instantaneous, time delayed and convolutive, using real room impulse responses. The experiments showed that in the worst case (convolutive mixture scenario) a mean improvement of 11dB SIR is achieved by the proposed GaBSS method for both output channels. Furthermore, the phoneme recognition rate of the separated signals was found to approach the rate measured with the clean signals in all experiments. The recognition rate improvement is maximised in the case of convoluted mixing of equal energy speech signals.

1. INTRODUCTION

The state of the art speech recognition technology is still vulnerable in the presence of acoustic interference. One of the most difficult problems encountered is the interfering speech from competing stationary speakers, or even worse, from moving speakers. In such environment, robust speech recognition still remains a challenging task.

Humans have the ability to focus their listening attention on a single talker among a din of conversations and background noise, and recognize a specific voice (*"cocktail party effect"*). In order to solve the Blind Signal Separation (BSS) problem, several general purpose methods have already been proposed and tested in various applications such as array signal processing [1], communication problems [2], medical signal processing [3], etc, based on the hypothesis that the mixed signals are stochastically independent.

Recently [4], a hands-free speech recognition system was evaluated in a simulated, two-simultaneous-speaker environment using real-life recordings. In particular, the phoneme recognition accuracy improvement was measured by processing the mixed speech signals of two omni-directional microphones and the separated signals obtained by the Output Decorrelation Filtering method in the frequency domain. The recognition, as well as the SIR results, although promising for the field of speech recognition, showed a certain incompetence of the method to perform efficiently due to the nature of the stochastic gradient descent optimization technique that was used. The presence of multiple local optimum attractors traps the solution obtained by the gradient-based methods away from the global optimum one. Furthermore, the convergence behavior of the above optimization methods depends on the choice of the step size as well as on the initial values of the separation filters coefficients.

The Genetic Algorithm (GA) is a global optimization technique, which is able to find the global optimum solution without being trapped in local minima. As a result it has been successfully employed in a variety of multi-modal and multi-objective optimization problems. In signal processing, the GA was applied to the delay estimation of a sampled signal [5], the weight training of the feedforward neural networks [6] and the parameter estimation of linear and non-linear adaptive filters [7].

In this optimization framework we propose a novel Blind Signal Separation method, based on the minimization of the cross-correlation function of the separated signals in the frequency domain (GaBSS). The proposed method is evaluated in an artificially created multi-simultaneous-speaker environment using a sub-set of the TIMIT speech corpus. Specifically, the case of two simultaneous speakers is studied in instantaneous, time-delayed and convolutive, artificially mixed environments.

The structure of this paper is as follows: In the following section 2 the fundamentals of the Blind Signal Separation for instantaneous and convolutive mixtures both in time and frequency domain are described. In section 3, we present the Genetic Algorithm module and its integration into the speech separation system. In section 4 the speech recognition system and the experimental conditions for solving the specific *"cocktail party"* problem are presented. In section 5 the evaluation results of the proposed method are given.

2. BSS IN THE FREQUENCY DOMAIN

Observing instantaneous mixtures of sources we approache the elementary Blind Signal Separation (BSS) problem. Let us assume that N signals s_i are ordered in a vector $\mathbf{s}^T(t) = [s_1(t) \ldots s_N(t)]$ where t is the time index. These N signals are acquired from a set of N sensors, so we obtain: $\mathbf{x}^T(t) = [x_1(t) \ldots x_N(t)]$. By assuming linear superposition, the vector $\mathbf{x}(t)$ can be expressed as:

$$\mathbf{x}(t) = \mathbf{A} \cdot \mathbf{s}(t) \qquad (1)$$

where $\mathbf{A}$ is the mixing matrix. The BSS objective is to recover the original signals $\mathbf{s}(t)$, without any prior knowledge of the mixing coefficients given the mixed signals $\mathbf{x}(t)$.

This problem is faced using different separation criteria like the Information Maximization, the Entropy maximization or the Output Decorrelation Criterion and different optimization techniques (the natural or stochastic gradient descent method [8,9]). However the calculated unmixing matrix is found to recover the original sources arbitrarily scaled [8]. In addition the rows of the unmixing matrix may have a different ordering than the true inverse of the mixing matrix. These are referred to as the scaling and permutation problems.

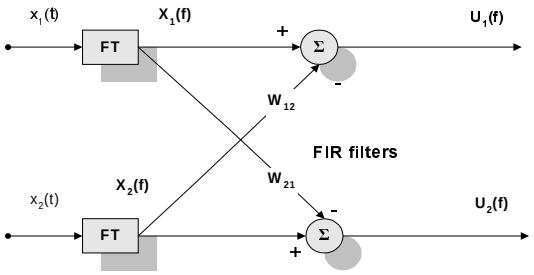

Figure 1 Basic Two Input Two Output (TITO) Blind Signal Separation Network

However simple it may be, the previous case is very rarely encountered in real-world situations, due to the filtering imposed on the sources by the room environment, propagation delays and wall reflections. Under these circumstances we have to use a more general mixing scenario known as convolutive mixture. To adequately express the mixing phase we make use of the FIR Matrix Algebra proposed by Lambert [10]. Using its notation, we can express the convolved mixing case in the form:

$$\mathbf{x}(t) = \underline{\mathbf{A}} \cdot \mathbf{s}(t) \tag{2}$$

where $\underline{\mathbf{A}}$ is the mixing matrix with each element being a FIR filter. For example, for $\underline{\mathbf{A}}_{2x2}$, equation (2) gives:

$$x_1(t) = a_{11}(t) \otimes s_1(t) + a_{12}(t) \otimes s_2(t),$$
$$x_2(t) = a_{21}(t) \otimes s_1(t) + a_{22}(t) \otimes s_2(t)$$

where ⊗ denotes the convolution operation.

Much work has been done in this field by various researchers worldwide. However, one of the major drawbacks of the proposed approaches is that they require significant computational power, while some of them have the side effect of whitening the spectrum of the output data [11]. This effect acts as a local minimum that hinders the convergence of algorithms that rely on gradient descent methods.

The basic two-input two-output (TITO) separation network in the frequency domain is shown in Figure 1. Lower case letters denote time-domain signals, while the corresponding capital letters denote their frequency domain equivalent. Let $x_1(t)$ and $x_2(t)$ be the mixed speech signal of both speakers obtained by two omni-directional microphones. The network output $U_1(f)$ and $U_2(f)$ denotes the Fourier transform of the speakers' separated signals. In the frequency domain the FIR filter matrix is transformed to a FIR polynomial complex-matrix by performing frequency transformation on each element. So the new derived form of the unmixing system becomes:

$$\underline{\underline{\mathbf{X}}} = \underline{\underline{\mathbf{A}}} \cdot \underline{\underline{\mathbf{S}}} \tag{3}$$

where now the convolution operation has been replaced by element-wise multiplications of polynomials. The matrix $\underline{\mathbf{X}}$ and $\underline{\mathbf{S}}$ contain the frequency transforms of $\mathbf{x}(t)$, $\mathbf{s}(t)$ respectively and $\underline{\mathbf{A}}$ is the frequency transformation of the mixing matrix. By closer inspection of the above equation, we can rewrite it in the form of:

$$\mathbf{X}(f) = \mathbf{A}(f) \cdot \mathbf{S}(f) \tag{4}$$

where $\mathbf{S}(f)$ and $\mathbf{X}(f)$ are vectors with one element for each source which is the frequency transform element of $\mathbf{s}_i(t)$ and $\mathbf{x}_i(t)$ at the frequency bin f, and $\mathbf{A}(f)$ is the matrix with the elements of the mixing filters in the frequency domain. By closer examination we can also see that the above equation has the same form with that of equation 1. This is actually the case as the frequency elements acquired from our sensors are in fact instantaneous mixtures of the original frequency elements of the sources.

3. THE GENETIC ALGORITHM

The GA is based on the mechanics of natural selection and genetics to emulate the evolutionary behavior of biological systems. In general, it consists of three processes: selection, reproduction and mutation, which make the transition from one population generation to the next.

In the proposed signal separation method for a TITO network, multiple GAs are applied to minimize the output cross-correlation of the signals *at every frequency bin f*:

$$C(f) = \sum_{k=1}^{N} (U_1^{(k)}(f) \cdot [U_2^{(k)}(f)]^H)^2 \tag{5}$$

where N is the total number of frames.

Each GA operates on a population of "chromosomes" coding four real numbers, i.e. the two complex numbers of the separation filter coefficients. The magnitude and the phase parameters are limited in the speech separation problem in reverberant rooms. Specifically, the binary representation of the magnitude is limited in the range [0,1] and the phase parameters are binary coded in the range $[-\pi,\pi]$. The binary alphabet offers the maximum number of schemata than any other coding method and the followed binary representation method maximizes the searching capabilities of the GA for the specific problem.

4. EXPERIMENTS

In this section we present the experiments that were conducted to demonstrate the efficiency of the implemented GaBSS method in the frequency domain. In our experiments, speech recordings from a subset of the TIMIT database were used to simulate the *"cocktail party effect"*. In particular, recordings from the test set of the TIMIT database were used to formulate a set of 360 sentences in a scenario where two speakers are talking simultaneously under various Relative Energy Level (REL) values. The speech signals were preprocessed so that they had zero-mean value and REL that ranged from as low as –20dB to 20dB for each speech channel respectively. All signals were sampled at 16 kHz. The mixing scenario consisted of three parts:

a) Instantaneous mixtures.
b) Mixtures with time-delayed cross-filters at 7.5 and 5 ms.
c) Mixtures with two linear FIR cross-filters of 256 coefficients.

In all the above cases the channel distortions from talker 1 to microphone 1 and talker 2 to microphone 2 were assumed negligible.

4.1 The GA Setup

The length of each "chromosome" was chosen to be 120 bits, coding the magnitude and the phase real values for each one of the two cross-filters at each frequency bin. Each real free parameter of the optimization problem was binary coded using a string of 30 bits equally distributed in the search space. The maximum precision of the GA found solution is $9.31*10^{-10}$ for the magnitude and $9.31*10^{-10}$ for the phase parameters. A population of 100 "chromosomes" is generated repeatedly 300 times. The mutation probability was selected to be 0.4 for each "chromosome", while the crossover probability was equal to 0.9.

The separation filters were chosen to have 256 taps length in all mixing scenarios and a 512 point FFT algorithm was used to transform short time domain signal frames into their frequency domain equivalent representation.

4.2 Speech recognition

The speech signal frames were decomposed in critical rectangular bands (the first 20 bands from [12], page 142), by using 32 ms FFT computed every 5 ms. The feature vector consisted of the normalized log-energy of each critical band with respect to the total frame log-energy.

The phoneme recognition experiments were carried out on a speaker-independent acoustic decoder based on Continuous Density Hidden Markov Models (CDHMM). Each phoneme-unit HMM was considered to be a four states left to right CDHMM with no state skip. The output distribution probabilities were modeled by means of a Gaussian component with diagonal covariance matrix. The classification was achieved by reaching the maximum forward probability of the observation sequence for each phoneme model. In the training process the segmental K-Means algorithm was used to estimate each CDHMM's parameter from multiple observations. A total number of 25000 manually labeled phonemes were used for training while for the testing we used approximately 5000 phonemes. A set of 39 different phonemes was employed that was created by a unification of the 49 original phonemes of the TIMIT database, on the basis of their acoustic similarity and the work of Lee and Hon [13].

5. RESULTS

In this section we present two types of experimental results of our GaBSS method. In particular the mean SIR improvement at both output channels of the separation network and the phoneme recognition rate for all the different REL mixing cases is presented

5.1 SIR improvement

INSTANT	Channel 1		Channel 2	
REL	**Before**	**After**	**Before**	**After**
-20	-16.31	5.49	29.87	22.00
-10	-6.90	19.35	20.45	14.80
0	3.09	21.98	10.45	17.57
10	13.10	25.10	0.44	21.07
20	23.14	32.77	-9.58	19.79
DELAYED	**Channel 1**		**Channel 2**	
REL	**Before**	**After**	**Before**	**After**
-20	-16.31	1.25	27.37	25.60
-10	-6.9	6.65	17.95	18.99
0	3.09	15.17	7.95	17.11
10	13.10	19.23	-2.05	13.78
20	23.14	24.05	-12.08	6.19
CONVOLUTIVE	**Channel 1**		**Channel 2**	
REL	**Before**	**After**	**Before**	**After**
-20	-9.85	2.19	21.55	22.63
-10	0.69	15.07	12.13	19.94
0	10.73	17.63	2.13	17.55
10	20.74	19.80	-7.87	11.12
20	30.78	26.42	-17.90	4.10

Table 1 SIRs for the three types of mixing scenarios (in dB).

In Table 1, we show the Signal to Interference Ratios for the case of instantaneous, time-delayed and convolutive mixing

	INSTANT		DELAYED		CONVOLUTIVE	
REL	**Ch 1**	**Ch 2**	**Ch 1**	**Ch 2**	**Ch 1**	**Ch 2**
-20	21.80	-7.87	17.56	-1.7	12.05	1.08
-10	26.50	-5.65	13.55	1.03	14.37	7.81
0	18.88	7.11	12.08	9.15	6.89	15.41
10	11.99	20.63	6.12	15.83	-0.94	19.00
20	9.62	29.38	0.90	18.28	-4.35	22.01

Table 2: SIR improvement for the instantaneous, time-delayed and convolutive mixing scenarios (in dBs).

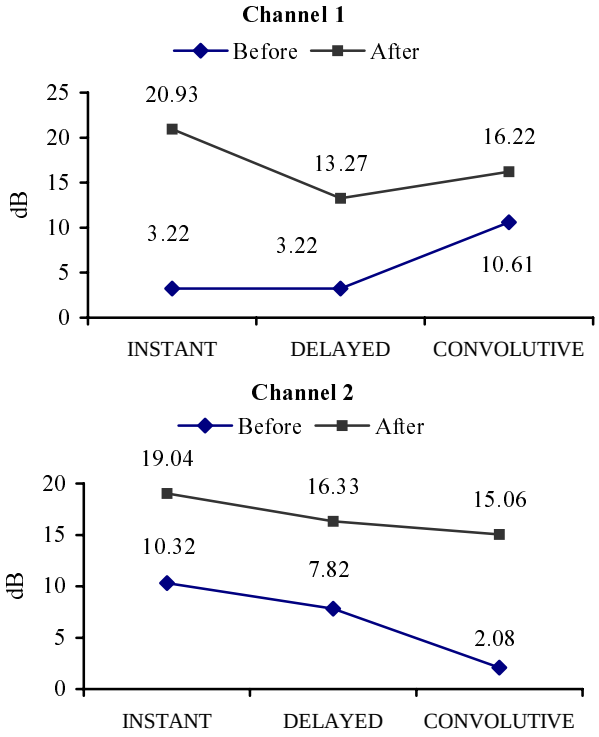

Figure 2 Mean SIR before and after signal separation for the three mixing scenarios

environments for both output channels. The proposed speech separation method improves the SIR ratio significantly, reaching 17.5 dB. This result is measured in the worst case, where the speech signals are filtered using real room transfer functions at 0dB REL.

On the other hand, the SIR of the separated signals is slightly decreased in case of high REL mixture signals, e.g the SIR in case of separating the convoluted speech signal of channel 1at 20 dB REL is reduced from 30.78 dB to 26.42 dB. However, this is not really a problem as the SIR of the signal still remains at a high level, so that the speech recognition rate is not affected.

Finally, in Table 2 and in Figure 2 we present the mean SIR improvement for all three mixing scenarios for both output channels. The overall mean SIR improvement taking all cases into account reaches 11.11dB and 10.09dB respectively. In all cases extensive acoustic tests proved the efficiency of our speech separation method.

5.2 Phoneme Recognition

In Table 3, the recognition rates achieved by a 4 hidden state CDHMM based phoneme classifier for both channels are

given. Clean speech signals from the TIMIT database were used in the training process, while the evaluation of the implemented phoneme recognition system was performed by using the clean, the mixed and the separated signals of the two channels in all mixing scenarios. The system achieved a recognition rate of 23.42% for the first channel (recognition score of the clean speech signals 26.25% and that of the mixed signals 17.03%) and 23.25% for the second channel (clean signals' score: 26.97% – mixed signals' score: 17.12%). In Figure 3 the mean phoneme recognition rate is shown for channels 1 and 2 and for the three mixing scenarios.

(a)	CLEAN		MIXED		SEPARATED	
REL	**Ch1**	**Ch2**	**Ch1**	**Ch2**	**Ch1**	**Ch2**
-20	28.51	28.41	8.18	23.8	22.26	25.49
-10	28.51	28.41	11.22	22.44	24.18	23.11
0	28.51	28.41	16.23	19.52	25.54	23.94
10	28.51	28.41	21.35	15.14	26.49	24.22
20	28.51	28.41	25.57	10.43	27.69	22.70
(b)	CLEAN		MIXED		SEPARATED	
REL	**Ch1**	**Ch2**	**Ch1**	**Ch2**	**Ch1**	**Ch2**
-20	23.51	27.90	8.11	23.36	20.96	26.35
-10	25.23	26.77	11.11	21.27	21.29	25.04
0	25.51	26.21	16.30	18.40	21.61	23.17
10	26.10	24.40	21.37	14.00	23.95	21.30
20	28.31	23.84	25.45	9.47	25.99	21.06
(c)	CLEAN		MIXED		SEPARATED	
REL	**Ch1**	**Ch2**	**Ch1**	**Ch2**	**Ch1**	**Ch2**
-20	20.95	28.10	8.70	22.38	19.15	26.15
-10	22.22	27.65	12.70	19.79	20.91	22.86
0	24.83	27.42	18.53	16.51	23.30	22.84
10	26.45	26.60	23.65	12.09	23.28	20.89
20	28.15	23.84	27.13	8.24	24.79	19.67

Table 3 Phoneme recognition rates (percent) for (a) instant, (b) delayed, and (c) convolutive mixing, before and after the implementation of the separating algorithm.

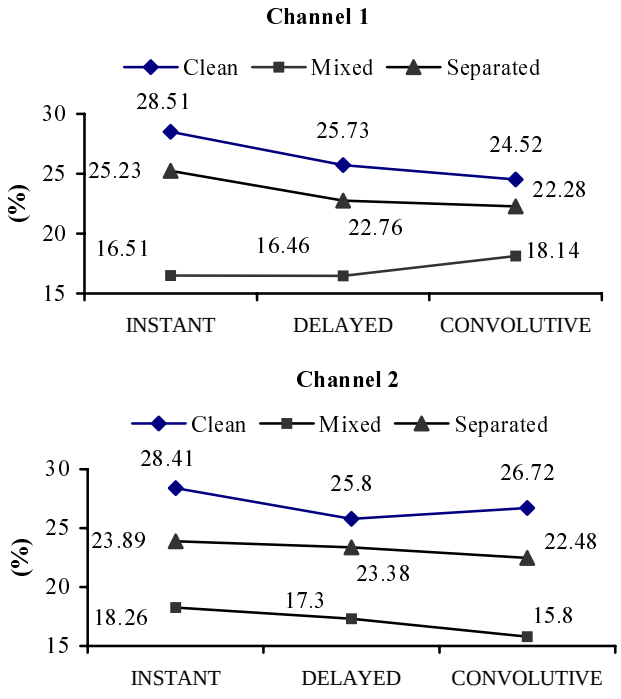

Figure 3 Mean phoneme recognition rate for the three mixing scenarios

6. CONCLUSIONS

In this paper, we presented and evaluated a new BSS technique in the frequency domain, which is based on the minimization of the cross correlation of the output signals using Genetic Algorithms. The evaluation experiments were carried out by measuring the phoneme recognition rate of a HMM acoustic decoder before and after separation, in three mixing scenarios: instantaneous, delayed, and real room transfer function filtered. The experiments showed that the phoneme recognition rate of both outputs increased significantly compared to that in the mixing environment and approached the classification rate achieved in the case of recognizing clean speech. Furthermore, the Signal to Interference Ratio was enormously increased after the application of the separation rules reaching a satisfactory acoustic level even in extremely adverse mixing environments.

REFERENCES

[1] R. W. Liu, "Blind signal processing: an introduction", *in Pro. IEEE Int. Symp. On Circuits and Systems*, Atlanta, GA, May 1996, vol. 2, pp. 81-84.

[2] B. Laheld and J.F. Cardoso, "Adaptive source separation with uniform performance", *in Signal Processing VII: Theories and Applications*, EURASIP 1994, vol. 2, pp. 183-186.

[3] S. Makeig, A. Bell, T.P. Jung, and T. Sejnowski, "Independent Component Analysis of electroencephalographic data", *in Advances in Neural Information Processing Systems 8*, Cambridge, MA: MIT Press, 1996, pp. 145-151.

[4] Koutras A., Dermatas E. and Kokkinakis G.: "Blind signal separation and recognition in the frequency domain".*6th International Conference on Circuits and Systems*, Pafos, Cyprus, (1999), to be published.

[5] D. Etter, M. Masukawa, "A comparison of algorithms for adaptive estimation of the time-delay between sampled signals", *ICASSP*, 1981, pp. 1253-1256.

[6] D. Montana, L. Davis, "Training feedforward neural networks using genetic algorithms", *ICAI, Detroit* 1989, pp. 762-767

[7] D. Etter, M. Hicks and K. Cho: "Recursive adaptive filter design using an adaptive genetic algorithm", *ICASSP*, 1982, pp. 635-638.

[8] A. Bell, T. Sejnowski: "An information maximization approach to blind separation and blind deconvolution", *Neural Computation 7*, 1995, pp. 1129-1159.

[9] S. Amari, A. Cicocki and H. Yang: "A new learning algorithm for blind signal separation", *Advances in Neural Information Processing systems*, Vol 5.

[10] Lambert R.: **"**Multi channel blind deconvolution: FIR matrix algebra and separation of multipath mixtures"**.** *PhD Thesis*, University of Southern California, Dept. Of Electrical Engineering , 1996.

[11] A. Cichocki: "Blind separation and extraction of source signals – Recent results and open Problems", *4th annual Conference of the Institute of Systems, Control and Information Engineers*, Osaka, Japan, 1997, pp. 43-48.

[12] Zwicker E., Fasl H.: "Psychoacoustics: Facts and Models", *Springer Verlang*, Berlin Heidelberg, 1990.

[13] Lee K., Hon H.: **"**Speaker independent phone recognition using Hidden Markov Models". *IEEE Trans. on ASSP* 37(11) 1989, 1641-1648.