Blind Speech Separation of Moving Speakers Using Hybrid Neural Networks

Athanasios Koutras, Evangelos Dermatas and George Kokkinakis

WCL, Department of Electrical and Computer Engineering
University of Patras, Greece
koutras@giapi.wcl2.ee.upatras.gr

Abstract

In this paper we present a novel method for Blind Speech Separation of convolutive speech signals of moving speakers in highly reverberant rooms. The separation network used is a hybrid neural network, which performs separation of convolutive speech mixtures in the time domain, without any prior knowledge of the propagation media, based on the Maximum Likelihood Estimation (MLE) principle. The proposed method improves significantly (more than 13% in all adverse mixing situations) the performance of a phoneme-based continuous speech recognition system and therefore can be used as a front-end to separate simultaneous speech of speakers who are moving in reverberant rooms.

1. Introduction

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, which is known as the "*cocktail party effect*". The problem of Blind source separation (BSS) consists of recovering unknown signals or "sources" from their several observed mixtures. Typically, these mixtures are acquired by a number of sensors, where each sensor receives signals from all sources. The term "blind" is justified by the fact that the only *a-priori* knowledge that we have for the signals is their statistical independence. No other information about the signal distortion on the transfer paths from the sources to the sensors is available beforehand.

There are many potential applications of blind signal separation, some of them referring but not restricted to: communication systems [1], biomedical signal analysis such as MEG, ECG, EEG [2], speech enhancement and noise reduction [3-6], and speech recognition [7-10].

The state of the art speech recognition technology is still vulnerable in the presence of acoustic interference. Specifically, one of the most difficult problems encountered is the interfering speech from competing stationary speakers, or even worse, from moving speakers in highly reverberant rooms. In the latter case, robust speech recognition in real room environments still remains a challenging task.

Generally, a great number of algorithms for BSS of speech signals have already been proposed [11,12]. However, most of them deal with the instantaneous mixture of sources [11-12] and only a few methods examine the case of convolutive mixtures of speech signals [3-10]. BSS for improving the speech recognition rate in real reverberant environments has been mostly tested for the case of stationary speakers positioned near to a microphone [6-9]. On the other hand, the presentation of algorithms capable of dealing with the case of moving speakers has been very limited [10,13]. Particularly, in [10], a recurrent neural network topology was presented that could separate competing moving speakers in a real room environment.

In this paper we propose an on-line blind signal separation method in the time domain for separating competing moving speakers in a reverberant real room environment based on a hybrid structure neural network topology [8]. Blind speech separation is performed by decomposing the speech signals in short time intervals and applying the MLE criterion. Extensive phoneme recognition experiments on a speaker independent, automatic speech recognition system, have shown that the proposed network performs better than the one in [10] and is capable of improving the recognition rate of the separated speech signals significantly in comparison to the rates achieved with the sensors signals. Specifically, the experimental results show an improvement of the phoneme recognition accuracy that reaches 13% for both separated speakers in high interference environments. In addition, our experiments show that the proposed BSS method outperforms the standard BSS method for stationary speakers by a mean value of 10% even in adverse mixing environments.

The structure of this paper is as follows: In the next section we present the basic method for separating convolutively mixed stationary speech signals, based on the MLE criterion and a hybrid neural network. Furthermore, the extension of the algorithm that deals with the case of non-stationary, moving speakers in the time domain is given. In section 3 the speech recognition experiments that were carried out for the evaluation of the aforementioned algorithm's performance are presented. In section 4 the speech recognition results are given in detail. Finally, in the last section some conclusions and remarks are given.

2. The Hybrid Neural Network

2.1. Blind Separation of Stationary speakers

Let us assume that we have M speech sources, denoted by $s_j(t)$ for $j \in [1,2,\ldots,M]$, which are considered to be zero meaned, mutually stochastic independent, in a real room environment. In addition let N be the number of the sensors located at fixed positions in the room. The microphones acquire the *convolutive* mixture of the speech signals denoted by $x_i(t)$, for $i \in [1,2,\ldots N]$.

$$x_i(t) = \sum_{j=1}^{N} \sum_{k=0}^{L} a_{ij}(k) s_j(t-k) \qquad (1)$$

Due to room acoustics, the sensors acquire besides the speakers' speech signals, delayed versions produced by multiple echoes that are propagated in the room. To solve the BSS problem, we make the assumption that the number of the speech signals that must be separated is known beforehand.

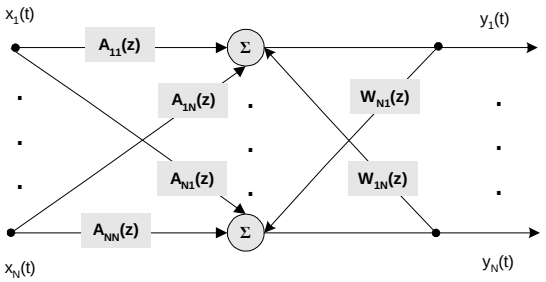


Figure 1. The hybrid network used for blind separation of speech signals.

The implemented network that we used to solve the BSS problem is a hybrid network structure, as shown in Figure 1. Hybrid neural networks have been already applied to the Blind Separation problem of stationary speakers with successful results [8]. In this work we use the hybrid neural networks to separate moving speakers as well. The separating filters $\boldsymbol{\alpha}_{ij}$, $\mathbf{w}_{ij}$ that need to be estimated are modeled by FIR linear filters. If the number of sensors is equal to the number of sources (N=M), the signals at the output of the network are:

$$y_i(t) = \sum_{j=1}^{N}\sum_{k=0}^{L} a_{ij}(k)x_j(t-k) + \sum_{\substack{j=1\\ j\neq i}}^{N}\sum_{k=0}^{L} w_{ij}(k)y_j(t-k) \quad (2)$$

where L is the order of the FIR filters.

In order to achieve signal separation of the output signals, we have to assure statistical independence; that is the joint probability density function (pdf) of the output signals $y_i(t)$ must be factorized. To this end the Maximum Likelihood Estimation Principle (MLE) was employed. The normalized log-likelihood of the parametric density estimator that needs to be maximized for the above network is defined by:

$$L(\mathbf{A},\mathbf{W}) = \log(p(\mathbf{x};\mathbf{A},\mathbf{W})) = \log\left(\left|\mathbf{A}_0\right|\right) + \log\left(p(\mathbf{y};\mathbf{A},\mathbf{W})\right) \quad (3)$$

where $\mathbf{A}_0$ is a NxN matrix with the leading weights of each feedforward FIR separating filter $\alpha_{ij}(0)$ and $\mathbf{A}$, $\mathbf{W}$ the matrices with the separating feedforward and feedback FIR filters α_{ij}, w_{ij} in their elements respectively.

The separating filter coefficients can be estimated using the stochastic gradient of $L(\mathbf{A},\mathbf{W})$ with respect to each filter coefficient $\alpha_{ij}(k)$, $w_{ij}(k)$ as follows:

$$\frac{\partial L}{\partial \mathbf{A}_0} = \left[\mathbf{A}_0^T\right]^{-1} - f(\mathbf{y})\cdot\mathbf{x}^T \quad (4)$$

and

$$\frac{\partial L}{\partial \mathbf{A}_k} = -f(\mathbf{y})\cdot\mathbf{x}(t-k)^T \quad (5)$$

$$\frac{\partial L}{\partial w_{ij}(k)} = -f(y_i(t))\cdot y_j(t-k)^T \quad (6)$$

where $\mathbf{A}_k$, is the NxN matrix with the k^{th} order separating filter coefficients $\alpha_{ij}(k)$.

However, the adaptation rule in (4) presents the drawback that it requires computationally expensive matrix inversion operations. To overcome this problem we use the natural gradient approach [14], which results in:

$$\frac{\partial L}{\partial \mathbf{A}_0} = \left(\left[\mathbf{A}_0^T\right]^{-1} - f(\mathbf{y})\cdot\mathbf{x}^T\right)\cdot\mathbf{A}_0{}^T\mathbf{A}_0 = \left(\mathbf{I} - f(\mathbf{y})\mathbf{x}^T\cdot\mathbf{A}_o^T\right)\cdot\mathbf{A}_0 \quad (7)$$

The non-linear function f presented in the above equations is equal to the cumulative density function of the source signals and its choice plays an important role to the efficacy of the BSS neural network. For the case of speech signals it has been shown that their pdf can be approximated by a gamma distribution variant or the Laplacian density. For the latter case, as Charkani and Deville have reported [15], the choice of $f(y_i(t)) = sign(y_i(t))$ is the best for speech signals. In this case, the weight update equations for the feedforward and the recurrent part of the neural network are given by:

$$\mathbf{A}_k^{(n+1)} = \mathbf{A}_k^{(n)} - \mu\cdot\left(sign(\mathbf{y}(t))\mathbf{x}^{\mathrm{T}}(t-k)\right) \quad (8)$$

$$\mathbf{A}_o^{(n+1)} = \mathbf{A}_o^{(n)} - \mu\cdot\left(\mathbf{I} - sign(\mathbf{y}(t))\mathbf{x}^{\mathrm{T}}(t)\right)\cdot\mathbf{A}_o^{(n)} \quad (9)$$

$$w_{ij}^{(n+1)}(k) = w_{ij}^{(n)}(k) - \mu\cdot\left(sign(y_i(t))y_j{}^{\mathrm{T}}(t-k)\right) \quad (10)$$

The above learning rules are valid only for the case that the sources are considered stationary, that is the speakers are positioned in the room and they are not moving. The above algorithms perform on-line and the adaptation of the separating filter coefficients is carried out each time a new sample is presented to the hybrid network.

2.2. Blind Separation of moving speakers

Let us now consider the case of moving speech sources. In this case the mixing matrix of filters $\mathbf{H}$ becomes time-dependent $\mathbf{H}(t)$. Assuming that $\mathbf{H}(t)$ is smoothly and slowly varying compared to the typical time-scale of the speech sources, and choosing an appropriate time scale T, so that for any time window of size T the sources can be considered stationary and at the same time we have sufficient statistics of them, the mixture matrix of filters can be considered to be nearly constant. In this case, we propose a solution for solving the blind signal separation problem of moving speakers: For each time window $B_k[t\text{-}T, t]$, under the assumption that the sources are stationary in this interval, we compute a set of separating filters using the standard BSS equations (8)-(10).

The proposed algorithm performs on-line as well, and calculates the set of separating filters $\boldsymbol{\alpha}_{ij}$, $\mathbf{w}_{ij}$ and the separated speech signal segments in each time window B_k. After the separation procedure, the speech segments from every time window B_k are used to reconstruct the separated speech signals. To this end, in each time window B_k we apply the following weight update equation:

$$\mathbf{A}_k^{(n+1)}(T) = \mathbf{A}_k^{(n)}(T) - \mu\cdot\left(sign(\mathbf{y}(t))\mathbf{x}^{\mathrm{T}}(t-k)\right) \quad (11)$$

$$\mathbf{A}_o^{(n+1)}(T) = \mathbf{A}_o^{(n)}(T) - \mu\cdot\left(\mathbf{I} - sign(\mathbf{y}(t))\mathbf{x}^{\mathrm{T}}(t)\right)\cdot\mathbf{A}_o^{(n)}(T) \quad (12)$$

$$w_{ij}^{(n+1)}(T,k) = w_{ij}^{(n)}(T,k) - \mu\cdot\left(sign(y_i(t))y_j{}^{\mathrm{T}}(t-k)\right) \quad (13)$$

where the index T denotes the particular time interval for which the separation is performed.

The choice of the time interval length B_k is very crucial for the algorithm's performance. If B_k is set very large, then the assumption for smooth and slow varying mixing matrix doesn't hold which results to a significant deterioration of the separation performance. On the other hand the choice of a small length time interval B_k may not lead to accurate separation as well, due to lack of sufficient statistics of the sources that prevent the algorithm's training, even though the stationarity of the source signals is valid. So it can be deduced

that the performance of BSS for moving sources depends highly on the length of the time interval B_k.

3. Experiments

The experiments presented in this paper are focused primarily on the test of the proposed BSS method in the case of two simultaneous non-stationary speakers in reverberant real room environment. This environment imposes several additional problems such as background noise, reflections and absorptions of the sound signals due to the walls and furniture in the room. In our experiments, the speakers were considered to walk in a room, while two microphones received the convolutive mixtures of their speeches.

3.1. Experimental Setup

To test our method, two sets of experiments were carried out. In all experiments speech recordings from the TIMIT speech database were used to simulate the speakers in a real room environment. In the first set two speech sources were activated simultaneously; the first source was fixed in a position, while the second one was moved around the room. In the second set both sources were moved in the same room.

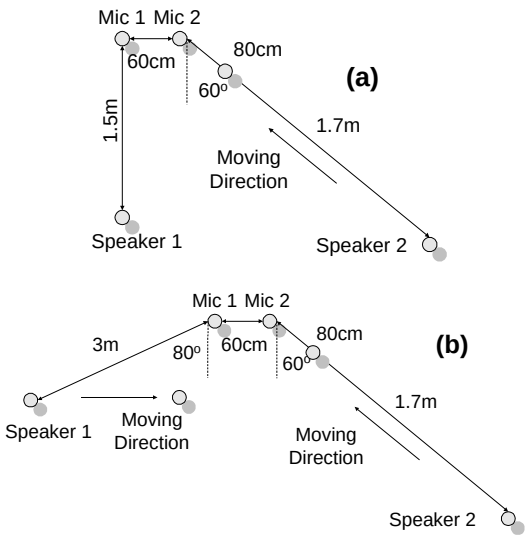


Figure 2. The topologies used in the experiments

The room where the experiments took place, was a typical office room, with possible background noise, of dimensions 6.5m x 4.5m x 2.5m. Two omni-directional microphones were placed 60cm apart from each other.

In the first case, the stationary source was placed in a distance of 1.5m from the first microphone. The second source was moved from a distance of 2.5m from the second microphone towards the microphone, until it reached a distance of 80cm (Figure 2a).

In the second set of the experiments we kept the same topology of the omni-directional microphones (Figure 2b). The first source was moved from a distance of 3m from the first microphone in a straight line parallel to the microphones until it reached a position at about 20 degrees left of the first microphone. The second source was moved in the same way as in the first experiment. The time window B_k denoted in section 2 was experimentally chosen to be 125ms.

3.2. Speech Corpus

Recordings from the test set of the TIMIT database were used to formulate a set of 100 pairs of sentences under various Relative Energy Level (REL) values. The speech signals were preprocessed so that they had zero mean value and RELs that ranged from –20dB to 20dB in each speech channel.

3.3. Speech Recognition System

The feature extraction module extracted the 12 Mel-cepstrum coefficients plus the energy parameter from the speech signals. The mean value of the Mel-cepstrum coefficients was subtracted from each coefficient and the first- and second-order differences were formed to capture the dynamic evolution of speech signals, resulting to a total number of 39 parameters.

The phoneme recognition experiments were carried out on a continuous speech, speaker-independent acoustic decoder based on five states left to right Continuous Density Hidden Markov Models (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. The complete training set from the TIMIT database was used for the training of the recognition system, while for the testing we used the 100 pairs of sentences taken from the testing set. For the recognition experiments, a set of 39 different phoneme categories was employed.

4. Results

To evaluate the accuracy of the proposed BSS method we used the recognition system on the original (anechoic) recordings (CLEAN), the convolutive mixtures from both microphones (MIXED), the output signals from algorithms (8)-(10) (SBSS) and the output signals from algorithms (11)-(13) (NSBSS).

4.1. One moving speaker

In this section we present the percent phoneme recognition results for the first case of our experiments where the first speech source is stationary and the second source is moving. In tables 1 and 2 the recognition results for both output channels are presented.

REL (dB)	-20	-10	0	10	20
CLEAN	69.78	69.78	69.78	69.78	69.78
MIXED	26.58	35.12	42.45	53.7	62.33
SBSS	31.99	37.52	44.88	56.99	62.51
NSBSS	41.12	45.90	51.6	59.14	62.42

Table 1. Percent phoneme recognition rate for channel 1.

REL (dB)	-20	-10	0	10	20
CLEAN	53.86	53.86	53.86	53.86	53.86
MIXED	40.15	38.04	32.39	29.18	25.12
SBSS	42.4	41.22	35.34	31.8	31.15
NSBSS	45.82	44.9	43.12	41.17	40.83

Table 2. Percent phoneme recognition rate channel 2.

From the above tables we can see that the proposed method succeeds in separating the moving and the stationary

speakers. The algorithm achieved separation even in the most difficult mixing situations of REL –20 and 20 dB for speaker1 and 2 respectively. In this case a percent phoneme recognition improvement of about 15% was achieved in comparison to the mixed sources results. The overall mean accuracy of the phoneme recognizer was found to be 52% for channel 1 and 43% for the second channel. In addition, the proposed method worked better than the standard BSS method for stationary speakers especially in the adverse environments of low RELs, showing an improvement of about 10%.

4.2. Two moving speakers

For the second case, the percent recognition rates for both moving speech sources are presented in Tables 3 and 4.

REL (dB)	-20	-10	0	10	20
CLEAN	69.78	69.78	69.78	69.78	69.78
MIXED	28.12	32.15	39.98	44.18	49.01
SBSS	29.7	33.87	42.68	48.16	54.98
NSBSS	39.44	41.89	52.11	53.98	57.02

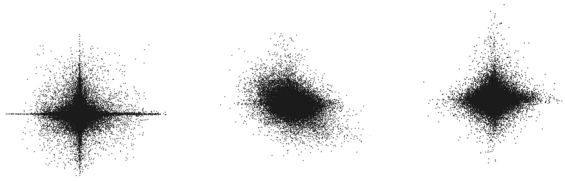
Table 3. Percent phoneme recognition rate for channel 1.

REL (dB)	-20	-10	0	10	20
CLEAN	53.86	53.86	53.86	53.86	53.86
MIXED	29.54	27.36	26.34	24.12	22.01
SBSS	40.5	35.78	30.33	28.4	23.87
NSBSS	43.11	40.42	40.03	37.96	36.65

Table 4. Percent phoneme recognition rate for the channel 2.

The proposed algorithm succeeded in separating the two speakers' signals in the most difficult and frequently encountered case where both speakers are moving in a real room environment. The mean phoneme recognition rate that was achieved was 49% for the first speaker and 40% for the second speaker respectively. Again, the efficacy of the proposed BSS method for moving speakers is evident compared to the standard BSS method's performance. A 10% improvement was measured as well in the adverse mixing environments (REL -20dB and 20dB for the first and the second speaker).

In Figure 3 we present the scatter plots of the joint probability density function of the clean speech signals, the convolutively mixed signals and the separated speech signals using the proposed algorithm. It is clearly evident that factorization of the joint pdf of the separated signals was achieved. Audible results can be accessed at: http://www.wcl2.ee.upatras.gr/koutras/on-line.htm.



(a) **(b)** **(c)**

Figure 3 Scatter plots of the joint pdf for the (a) clean, (b) convolutively mixed and (c) separated speech signals for two moving speakers.

5. Conclusions

In this paper we presented an on-line method for separating simultaneous moving speakers in a reverberant real room environment using a hybrid neural network topology. The separation criterion that was used was derived from the information theoretic approach based on the Maximum Likelihood Estimator and the decomposition of the signals in short time intervals where the mixture matrix of filters is considered to be constant. Extensive phoneme recognition tests used to evaluate the proposed method showed that the algorithm can perform satisfactory in a variety of mixing situations especially under the presence of high Interference.

6. References

1. Feng M. and Kammeyer K. “Application of source separation algorithms for mobile communication environment”, *1st Int. Workshop on ICA & Signal Separation*, pp. 431-436, 1999.
2. Makeig S., Bell A., Jung T., Sejnowski T. “ICA in electroencephalographic data”, *Advances in Neural Information Processing Systems*, (8):145-151, 1996.
3. Lee T., Bell A., Lambert R. “Blind separation of delayed and convolved sources”, *Advances in Neural Information Processing Systems*, (9):758-764, 1997.
4. Girolami M. “Noise reduction and speech enhancement via temporal anti-Hebbian learning”, *ICASSP*, Vol 2, pp. 1233-1236, 1998.
5. Choi S., Cichocki A. “Adaptive blind separation of speech signals: Cocktail Party Problem”, *International Conference of Speech Processing*, Seoul, Korea, pp. 617-622, 1997.
6. Yen K., Huang J., Zhao Y. “Co-channel speech separation in the presence of correlated and uncorrelated noises”, *EuroSpeech*, Vol. 6, pp. 2587-2590, 1999.
7. Koutras A., Dermatas E., Kokkinakis G. “Recognizing simultaneous speakers: A genetic algorithm approach”, *EuroSpeech*, Vol. 6, pp. 2551-2554, 1999.
8. Koutras A., Dermatas E., Kokkinakis G., “Blind separation of speakers in noisy reverberant environments: A neural network approach”, *Neural Network World Journal*, 10(4):619-630, 2000.
9. Yen K., Zhao Y. “Co-channel speech separation for robust Automatic Speech recognition”, *ICASSP*, Vol 2, pp.859-862, 1997.
10. Koutras A., Dermatas E., Kokkinakis G., “Blind speech separation of moving speakers in real reverberant environments”, *ICASSP*, Vol 2, pp. 1133-1136, 2000.
11. Amari S., Cichocki A., Adaptive Blind Signal Processing-Neural Network Approaches. *IEEE Proceedings*, 86(10):2026-2048, 1998.
12. Cardoso J. “Blind Signal Separation: Statistical Principles”, *IEEE Proceedings*, 86(10):2009-2025, 1998.
13. Anemuller J., Gramms T. “On-line blind separation of moving sound sources”. *1st Int. Workshop on ICA & Signal Separation*, pp. 331-334, 1999.
14. Amari S., “Natural Gradient works efficiently in learning”, *Neural Computation*, 10:251-276, 1998.
15. Charkani N., Deville Y. “Optimization of the asymptotic performance of time-domain convolutive source separation algorithms”. *ESANN*, pp. 273-278, 1997.