Noise Reduction in Reverberant Environments with a Blind Source Separation Algorithm

Felipe Augusto Pereira de Figueiredo and Carlos Alberto Ynoguti

***Abstract*— In this paper we propose the use of a blind source separation (BSS) algorithm as a means to enhance convolutively mixed audio signals by separating the interference signals, such as cars, footsteps, trains, etc., from the desired speech source. Computer simulations confirm that the algorithm adopted in this work can efficiently extract the desired speech signal from a convolutive mixture. These results encourage its possible use as a speech enhancement front-end for automatic speech recognition (ASR) systems for example.**

***Keywords*— BSS, speech enhancement, convolutive mixtures.**

I. Introduction

In order to deploy automatic speech recognition (ASR) effectively in real world scenarios it is necessary to handle hostile environments with multiple speech and noise sources. One classical example is the so-called *cocktail party problem* [1], where a number of people are talking simultaneously in a room and the ASR task is to recognize the speech content of one or more target speakers amidst other interfering sources. Although the human brain and auditory system can handle this everyday problem with ease it is very hard to solve it with computational algorithms.

The speech recognition technology is still vulnerable when dealing with signals in the presence of acoustic interference [20]. Robust speech recognition in real environments still remains a challenging task.

On the other hand, the objective of Blind Source Separation (BSS) is to extract one or several source signals from the observed multichannel mixture signal. The signals of interest depend on the application: for instance, in the context of speech enhancement for mobile phones, the only source signal of interest is the user's speech. Undesired sources may then include speech signals from surrounding people and environmental noises produced by cars, wind or rain. Noise can also originate from clinking glasses or footsteps.

In this paper we propose the use of an offline blind signal separation method in the time domain to separate the speech of a target speaker from all the undesired sound sources that might have been convolutively mixed with the desired source signal. The algorithm adopted here was firstly presented in [2]. The mixture is simulated through the convolution of the sound signals (target speaker and undesired sound sources) and the room impulse response generated by the image method [3, 4].

The structure of this paper is as follows: in the next section the system model used to blindly separate speech signals in reverberant environments is shown and then, in Section III, an algorithm for blind separation of convolutive mixtures is presented. In Section IV, the experiments carried out for the evaluation of the aforementioned algorithm's performance are presented. Finally, in the last section some conclusions and ideas for further research are given.

II. Model of Blind Source Separation in Reverberant Environment

The problem of blind source separation is illustrated in Figure 1:

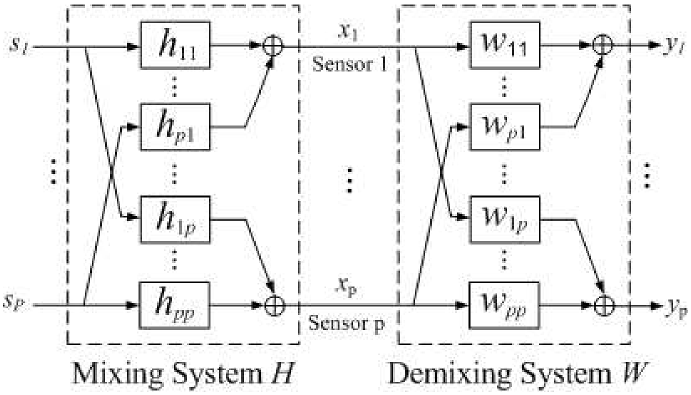

Figure 1. Linear MIMO model for BSS.

In this work it is assumed a MIMO (Multiple Inputs Multiple Outputs) model, in which the signals are convolutively mixed. Also, the number of source signals $s_q(n), q = 1, ..., Q$ is assumed to be equal to the number of sensor signals $x_p(n), p = 1, ..., P$.

Each of the outputs of the mixing system H is described by

$$x_p(n) = \sum_{q=1}^{P} \sum_{k=0}^{M-1} h_{qp}(k) s_q(n-k) \qquad (1)$$

where $h_{qp}(k), k = 0, ..., M-1$ denote the coefficients of the filter from the q-th source to the p-th sensor. The constant M gives the order of the filters employed to model de mixing system.

The problem of BSS consists in finding a corresponding demixing system according to Fig. 1, where the output signals $y_q(n), q = 1, ..., P(P = Q)$ are described by

$$y_q(n) = \sum_{p=1}^{P} \sum_{k=0}^{L-1} w_{pq}(k) x_p(n-k) \qquad (2)$$

Felipe Augusto Pereira de Figueiredo and Carlos Alberto Ynoguti¸ Instituto Nacional de Telecomunicações - INATEL, Minas Gerais, Brasil, E-mails: felipeaugusto@mtel.inatel.br, ynoguti@inatel.br.

where L is the length of the filters of the demixing system.

It can be shown (see, e.g., [1]) that the MIMO demixing system coefficients $w_{pq}(k)$ can in fact reconstruct [5] the sources up to an unknown permutation and an unknown filtering of the individual signals, where L should be chosen at least equal to M.

With the intention of estimating the P^2L coefficients of $w_{pq}(k)$ the MIMO demixing filter W, we consider in this work an approach using second-order statistics [6], which exploits the nonwhiteness and nonstationarity properties of the signals. The nonwhiteness property is exploited by simultaneous diagonalization of output correlation matrices over multiple time-lags, e.g., [7], and the nonstationarity property is exploited by simultaneous diagonalization of short-time output correlation matrices at different time intervals, e.g., [8, 9, 10]. In the sequence, we present an algorithm for convolutive mixtures by first introducing a general matrix formulation for convolutive mixtures following [11] that includes all time lags.

III. Time-domain Algorithm for BSS

In this section we introduce the matrix formulation that will allow us to derive a time-domain algorithm [12] from a cost function which inherently takes into account the nonstationarity and nonwhiteness properties.

A. Matrix notation for Convolutive Mixtures

From Figure 1, it can be seen that the output signals $y_q(n), q = 1, ..., P$ of the demixing system at time n are given by

$$y_q(n) = \sum_{p=1}^{P} x_p^T(n) w_{pq}, \tag{3}$$

where

$$\mathbf{x}_p(n) = [x_p(n), x_p(n-1), ..., x_p(n-L+1)]^T \tag{4}$$

is a vector containing the latest L samples of the sensor signal of the p-th channel and

$$\mathbf{w}_{pq}(n) = [w_{pq,0}, w_{pq,1}, ..., w_{pq,L-1}]^T \tag{5}$$

contains the current weights of the MIMO filter taps from the p-th sensor channel to the q-th output channel.

An algorithm for BSS of convolutive signals which exploits those two signal properties can be obtained from the definition of the following matrix

$$\mathbf{Y}_q(m) = \begin{bmatrix} y_q(mL) & \cdots & y_q(mL-D+1) \\ y_q(mL+1) & \cdots & y_q(mL-D+2) \\ \vdots & \ddots & \vdots \\ y_q(mL+N-1) & \cdots & y_q(mL-D+N) \end{bmatrix}, \tag{6}$$

where m denotes the time index of the block being processed and N is the length of the system output blocks taken into account for the estimates of short-time correlations used below. This matrix captures L subsequent output signal vectors

$$\mathbf{y}_q(m) = [y_q(mL), ..., y_q(mL+N-1)]^T \tag{7}$$

in order to incorporate L time-lags in the cost function and thus the algorithm will be able to exploit the nonwhiteness property.

With the definitions above, (2) can be rewritten as

$$\mathbf{Y}_q(m) = \sum_{p=1}^{P} \mathbf{X}_p(m) \mathbf{W}_{pq} \tag{8}$$

The approach followed here is carried out with overlapping data blocks to increase the convergence rate and reduce signal delay. Overlapping is introduced by simply replacing the time index mL in the equations $m(L/\alpha)$ by with the overlapping factor $1 \leq \alpha \leq L$. The matrices $\mathbf{X}_p(m), p = 1, ..., P$ used in Equation (8) are defined as

$$\mathbf{X}_p(m) = \begin{bmatrix} x_p(mL) & \cdots & x_p(mL-L+1) \\ x_p(mL+1) & \cdots & x_p(mL-L+2) \\ \vdots & \ddots & \vdots \\ x_p(mL+N-1) & \cdots & x_p(mL-L+N) \end{bmatrix} \tag{9}$$

Those matrices are Toeplitz with dimension ($N \times 2L$), so the first row contains $2L$ input samples and each subsequent row is shifted to the right by one sample and thus, contains one new input sample. $\mathbf{W}_{pq}$ are $2L \times L$ Sylvester matrices, which are defined as

$$\mathbf{W}_{pq}(m) = \begin{bmatrix} w_{pq,0} & 0 & \cdots & 0 \\ w_{pq,1} & w_{pq,0} & \ddots & \vdots \\ \vdots & w_{pq,1} & \ddots & 0 \\ w_{pq,L-1} & \vdots & \ddots & w_{pq,0} \\ 0 & w_{pq,L-1} & \ddots & w_{pq,1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & w_{pq,L-1} \\ 0 & \cdots & 0 & 0 \\ \vdots & \cdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \end{bmatrix} \tag{10}$$

Finally, to allow a convenient notation of the algorithm combining all channels, (8) can be compactly rewritten as

$$\mathbf{Y}(m) = \mathbf{X}(m)\mathbf{W} \tag{11}$$

where

$$\mathbf{Y}(m) = [\mathbf{Y}_1(m), \mathbf{Y}_2(m), ..., \mathbf{Y}_p(m)] \tag{12}$$

$$\mathbf{X}(m) = [\mathbf{X}_1(m), \mathbf{X}_2(m), ..., \mathbf{X}_p(m)] \tag{13}$$

$$\mathbf{W} = \begin{bmatrix} \mathbf{W_{11}} & \cdots & \mathbf{W_{1P}} \\ \vdots & \ddots & \vdots \\ \mathbf{W_{P1}} & \cdots & \mathbf{W_{PP}} \end{bmatrix} \tag{14}$$

B. Cost Function and Algorithm Derivation

Having defined the compact matrix formulation (11) for the block-MIMO filtering, a following cost function that explicitly contains correlation matrices including several time-lags under the assumption of short-time stationarity is defined. This cost function is based on a generalization of Shannon's mutual information [11, 13] and simultaneously accounts for those two properties of the signals used here:

$$\mathbf{J}(m) = \frac{1}{M}\sum_{i=0}^{M-1}\{\log|\operatorname{bdiag}\mathbf{Y}^T(i)\mathbf{Y}(i)| - \log|\mathbf{Y}^T(i)\mathbf{Y}(i)|\} \tag{15}$$

This cost function was firstly derived in [14] as a generalization of [15]. Since the matrix formulation (11) is used for calculating the short-time correlation matrices $\mathbf{Y}^T(m)\mathbf{Y}(m)$, the cost function inherently includes all L time-lags of all auto-correlations and cross-correlations of the BSS output signals. By Oppenheim's inequality $\sum_q \log|\mathbf{Y}_q^T\mathbf{Y}_q| \geq \log|\mathbf{Y}^T\mathbf{Y}|$ [16], it is ensured that the first term in the braces of (14) is always greater than or equal to the second term, where the equality holds if all block-diagonal elements of $\mathbf{Y}^T(m)\mathbf{Y}(m)$, i.e., the output cross-correlation over all time-lags, vanish. The algorithm is based on the first-order gradient and in order to express the update equations of the filter coefficients exclusively by Sylvester matrices W, we take the gradient with respect to W and ensure the Sylvester structure of the result by selecting the non redundant values using a constraint.

$$\triangledown_{\mathbf{W}}\mathbf{J}(m) = \frac{\partial\mathbf{J}(m)}{\partial\mathbf{W}} \tag{16}$$

And as result,

$$\begin{aligned}\triangledown_{\mathbf{W}}\mathbf{J}(m) = \tfrac{2}{M}\textstyle\sum_{i=0}^{M-1}\mathbf{R}_{xy}(i)\mathbf{R}_{yy}^{-1}(i)\{\mathbf{R}_{yy}(i)\\ -\operatorname{bdiag}\mathbf{R}_{yy}(i)\}\operatorname{bdiag}^{-1}\mathbf{R}_{yy}(i)\end{aligned} \tag{17}$$

With an iterative optimization procedure, the current demixing matrix is obtained by the recursive update equation

$$\mathbf{W}(m) = \mathbf{W}(m-1) - \mu(m)\Delta\mathbf{W}(m) \tag{18}$$

The $\mu(m)$ parameter gives the length of the step in the negative gradient direction and it is often called the step size or learning rate. As is known, non quadratic cost functions may have many local maxima and minima and therefore, good choices for initial values are important.

C. Natural Gradient

The gradient of a function $\mathbf{J}(m)$ points in the steepest direction in the Euclidean orthogonal coordinate system. However, the parameter space is not always Euclidean; in fact it has a Riemannian metric structure, as pointed out by Amari [17]. In such a case, the steepest direction is given by the so-called natural gradient instead. Therefore, in order to use the natural gradient as the update term $\Delta\mathbf{W}(m)$ the following modification should be applied to the descent gradient:

$$\triangledown_{\mathbf{W}}^{NG}\mathbf{J}(m) = \mathbf{W}\mathbf{W}^T\triangledown_{\mathbf{W}}\mathbf{J}(m) = \mathbf{W}\mathbf{W}^T\frac{\partial\mathbf{J}(m)}{\partial\mathbf{W}} \tag{19}$$

And then we have

$$\begin{aligned}\triangledown_{\mathbf{W}}^{NG}\mathbf{J}(m) = \tfrac{2}{M}\textstyle\sum_{i=0}^{M-1}\mathbf{W}(m)\{\mathbf{R}_{yy}(i) - \operatorname{bdiag}\mathbf{R}_{yy}(i)\}\\ \operatorname{bdiag}^{-1}\mathbf{R}_{yy}(i)\end{aligned} \tag{20}$$

IV. Computer Simulations

Speech enhancement in reverberant and noisy environments is a very challenging task which has a handful of applications and it's very useful in the areas of robust speech recognition and telecommunications.

The degree of difficulty in enhancing speech signals strongly depends on the environment conditions in which the speaker is located. In the case where the speaker is located near to a microphone, reverberating effects are minimum and standard methods are able to deal with moderate noise levels. However, when the speaker is far from the sensors (microphones), there will be severe distortions which include lots of noise and noticeable reverberation. Denoising and de-reverberation of speech signals in such conditions have proven to be a very challenging task.

In this section we present the results of experiments which show the usefulness of the algorithm adopted in this work in denoising and de-reverberating speech signals convolutively mixed with noise signals provided by the ETSI/Aurora database [18]. It includes recordings of five different kinds of noise which were captured in the following places: airport, restaurant, meeting room, street and train station. These noise signals are convolutively mixed with the speech signal of a speaker specifically recorded for this work.

The experiments carried out for our research were focused on assessing the quality of the signal separation achieved by the BSS algorithm adopted here in the case where a speaker's signal is corrupted by background noise in a reverberant simulated environment. Filters with 447 taps (M = 447) were generated by the image method with the purpose of simulating the acoustical behavior of a real room.

Both recordings, speaker's and background noise signals, have 4 second length. They were convolved with the synthetic impulse response of a room generated by the image method [3, 4].

The recordings were taken in a low noise environment with 8000 Hz sampling frequency and 16 bits of resolution. The technique used to adapt the step size of the BSS method is known as fixed step size [19], and it is made equal to 0.04. The length L of the demixing filters is made equal to the length M of the mixing filters, so we have L = M = 447. The demixing filters $\mathbf{W}_{pp}$ were initialized with an unit impulse at the first tap and all the taps of the filters $\mathbf{W}_{pq}, p \neq q$ are made equal to zero.

The Signal-to-Interference Ratio (SIR), which is defined as the ratio of the signal power of the target signal to the signal power from the jammer signal, was used to evaluate the performance of the algorithm. For this work the SIR measure can be interpreted as being the Signal-to-Noise Ratio (SNR) measure, once the interfering signal is noise. Therefore, from now on we adopt the SNR term. Figure 2 depicts the SNR achieved by the BSS algorithm for each one of the mixtures between the speaker's signal and the 5 noise signals from the Aurora data base. The SNR is averaged over both output channels. The desired separated signal, i.e., the speaker's signal is chosen listening to both output signals.

Analyzing Figure 2 and the table above, it can be noticed that the BSS method adopted here performs pretty well the

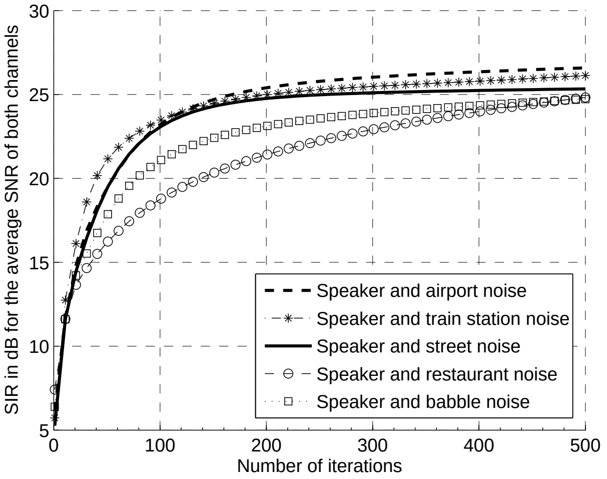

Figure 2. Results of the separation between speech and noise signals.

TABLE I

INITIAL AND FINAL SNR FOR THE EXPERIMENTS WITH NOISE.

Noise	Initial SNR (dB)	Final SNR (dB)	SNR Gain(dB)
Airport	6,13	26,59	20,46
Babble	6,38	24,73	18,34
Restaurant	7,42	24,83	17,41
Street	5,27	25,34	20,07
Train station	5,73	26,13	20,41

task of separating speech signal from background noise. The SNR gain is greater than 17 dB for the worst case and greater than 20 dB for the best one.

V. Conclusions and Future Work

The use of a method for blind source separation with the purpose of separating a target speech signal from background noise such as foot steps, cars, surrounding people, etc. is studied.

The SNR improvement presented by the use of the offline BSS algorithm firstly proposed by R. Aichner et al. [12] for the task of separating speech signals from background noise indicates that it could be used to improve the performance of speech recognition systems in reverberant and noisy environments in an offline fashion.

As for further research, the implementation of an online version of the algorithm adopted here is to be studied. This online version can be achieved through the addition of a properly chosen weighting function $\beta(i,m)$ to Equation (15), as showed in [21]. Additionally, the use of this online BSS method as a preprocessing tool for ASR systems in reverberating and noisy environments is to be evaluated. The method seems to be able to improve the performance of such systems.

References

[1] E. Cherry, *Some experiments on the recognition of speech, with one and with two ears*. Acoustical Society of America, Vol. 25, No. 5, pp. 975-979, 1953.

[2] H. Buchner, R. Aichner and W. Kellermann, *A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics*. IEEE Trans. Speech Audio Process, pp. 120-134, Jan. 2005.

[3] J. H. Rindel, *The use of computer modeling in room acoustics*. Journal of Vobroengineering, vol. 3, no. 4, pp. 41-72, 2000.

[4] L. Savioja, *Modeling techniques for virtual acoustics*. Ph.D. Dissertation, Helsinki University of Technology, Aug. 2000.

[5] A. Hyvarinen, J. Karhunen and E. Oja, *Independent Component Analysis*. Wiley & Sons, Inc., New York, 2001.

[6] R. Battiti, *First and second-order methods for learning: between steepest descent and Newton's method*. Technical report, University of Trento, 1991.

[7] L. Molgedey and H. G. Schuster, *Separation of a mixture of independent signals using time delayed correlations*. Review Letters, vol. 72, pp. 3634-3636, 1994.

[8] E. Weinstein, M. Feder and A. Oppenheim, *Multi-channel signal separation by decorrelation*. IEEE Trans. on Speech and Audio Processing, vol 1, no. 4, pp. 405-413, Oct. 1993.

[9] S. Van Gerven and D. Van Compernolle, *Signal separation by symmetric adaptive decorrelation: stability, convergence, and uniqueness*. IEEE Trans. Signal Processing, vol. 43, no. 7, pp. 1602-1612, 1995.

[10] C. L. Fancourt and L. Parra, *The coherence function in blind source separation of convolutive mixtures of non-stationary signals*. in Proc. Int. Workshop on Neural Networks for Signal Processing (NNSP), 2001.

[11] C. E. Shannon and W. Weaver, *The mathematical theory of communication*. University of Illinois Press, Urbana, Illinois.

[12] R. Aichner et al., *Time-domain blind source separation of non-stationary convolved signals with utilization of geometric beamforming*. in Proc. Int. Workshop on Neural Networks for Signal Processing, Martigny, Switzerland, 2002.

[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley & Sons, New York, 1991.

[14] H. Buchner, R. Aichner and W. Kellermann, *A generalization of a class of blind source separation algorithms for convolutive mixtures*. Proc. IEEE Int. Symposium on Independent Component Analysis and Blind Signal Separation (ICA), Nara, Japan, pp. 945-950, Apr. 2003.

[15] K. Matsuoka, M. Ohya and M. Kawamoto, *A neural net for blind separation of nonstationary signals*. Neural Networks, vol. 8, no. 3, pp. 411-419, 1995.

[16] A. Oppenheim, *Inequalities connected with definite hermitian forms*. J. London Math. Soc., vol. 5, pp. 114-119, 1930.

[17] S. I. Amari, *Natural Gradient Works Efficiently in Learning*. Neural Computation, Vol. 10, No. 2, Pages 251-276, Feb.1998.

[18] *Aurora project database*. http://aurora.hsnr.de/aurora-5.html.

[19] F. A. P. Figueiredo and C. A. Ynoguti, *On the improvement of the learning rate in Blind Source Separation using techniques from Artificial Neural Networks theory*. Proceedings of the International Workshop on Telecommunications, IWT, 2009.

[20] R. Cole et al., *The challenge of spoken language systems: Research directions for the nineties*. IEEE Trans. Speech Audio Processing, vol. 3, pp. 1-21, Jan. 1995.

[21] R. Aichner, H. Buchner, F. Yan and W. Kellermann, *A real-time blind source separation scheme and its application to reverberant and noisy acoustic environments*. Signal Processing, 86(6):1260-1277, Jun. 2006.