AUDIO SOURCE SEPARATION USING SPARSITY

[1]*A. Aïssa-El-Bey,* [1] [2]*H. Bousbia-Salah,* [1]*K. Abed-Meraim and* [1]*Y. Grenier*

{elbey, bousbia, abed, grenier}@tsi.enst.fr
[1]ENST-Paris, TSI department, 46 rue Barrault 75634, Paris Cedex 13, France
[2]ENP, electronics department, 10 avenue Hassan Badi, Algiers, Algeria

ABSTRACT

In this paper, we are interested in blind source separation from instantaneous mixtures of audio signals. Using the sparsity property of audio signals, we propose an iterative method that relies on a relative gradient technique which minimizes a contrast function based on the ℓ_p norm. This norm is considered as a good sparsity measure. The simulations show that the proposed method outperforms other methods based on source independency.

1. INTRODUCTION

This paper deals with blind source separation (BSS). The blind context means that neither the sources nor the mixing matrix are known. The goal of BSS is to recover the sources up to scaling and permutation by, only, using the mixtures. Blind source separation (BSS) has applications in several areas, such as communication, speech / audio processing, and biomedical engineering [1]. A fundamental and necessary assumption of BSS is that the sources are statistically independent and thus are often separated using higher-order statistical information [2]. If some information about the sources is available at hand, such as temporal coherency [3], source nonstationarity [4], or source cyclostationarity [5], then one can remain in the second-order statistical scenario.

In the case of non-stationary signals (including audio signals), certain solutions using time-frequency analysis of the observations exist [6]. Other solutions use the statistical independence of the sources assuming a local stationarity to solve the BSS problem [7]. This is a strong assumption that is not always verified [8]. To avoid this problem, we propose a new approach that handles the general linear instantaneous model (possibly noisy) by using the *sparsity* assumption of the sources in the time domain. The use of sparsity to handle this model, has arisen in several papers in the area of source separation [1, 9]. We first present a sparsity contrast function for BSS. Then, in order to achieve BSS, we optimize the considered contrast function using an iterative algorithm based on the gradient technique.

In the following section, we discuss the data model that formulates our problem. Next, we detail the different steps of the proposed algorithm. In Section 4, some simulations are undertaken to validate our algorithm and to show the usefulness of the proposed method.

2. DATA MODEL

Assume that N audio signals impinge on an array of $M \geq N$ sensors. The measured array output is a weighted superposition of the signals, corrupted by additive noise, i.e.

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{w}(t) \qquad t = 0, \ldots, T-1 \qquad (1)$$

where $\mathbf{s}(t) = [s_1(t), \cdots, s_N(t)]^T$ is the $N \times 1$ *sparse source vector*, $\mathbf{w}(t) = [w_1(t), \cdots, w_M(t)]^T$ is the $M \times 1$ gaussian complex *noise vector*, $\mathbf{A}$ is the $M \times N$ full column rank *mixing matrix* (i.e., $M \geq N$), and the superscript T denotes the transpose operator. The purpose of blind source separation is to find a separating matrix, i.e. a $N \times M$ matrix such that $\widehat{\mathbf{s}}(t) = \mathbf{B}\mathbf{x}(t)$ is an estimate of the source signals.

Before proceeding, note that complete blind identification of separating matrix $\mathbf{B}$ (or the equivalently mixing matrix $\mathbf{A}$) is impossible in this context, because the exchange of a fixed scalar between the source signal and the corresponding column of $\mathbf{A}$ leaves the observations unaffected. Also note that the *numbering* of the signals is immaterial. It follows that the best that can be done is to determine $\mathbf{B}$ up to a permutation and scalar shifts of its columns, i.e., $\mathbf{B}$ is a separating matrix iff:

$$\mathbf{B}\mathbf{x}(t) = \mathbf{P}\mathbf{\Lambda}\mathbf{s}(t) \qquad (2)$$

where $\mathbf{P}$ is a permutation matrix and $\mathbf{\Lambda}$ a non-singular diagonal matrix.

3. ITERATIVE SPARSE ALGORITHM

In this section, we propose an iterative algorithm for the separation of sparse audio signals ISBS for Iterative Sparse Blind Separation. As well known, audio signals are characterized by their sparsity property in the time domain [1,

9] which is measured by their ℓ_p norm where $0 \leq p \leq 1$. This norm represents how the "energy" is concentrated on a small number of coefficients. Based on this, one can define the following sparsity contrast function,

$$G_p(\mathbf{s}) = \frac{1}{N}\sum_{i=1}^{N}[\mathcal{J}_p(s_i)]^{\frac{1}{p}} \tag{3}$$

where

$$\mathcal{J}_p(s_i) = \frac{1}{T}\sum_{t=0}^{T-1}|s_i(t)|^p \tag{4}$$

The algorithm finds a separating matrix $\mathbf{B}$ such as,

$$\mathbf{B} = \arg\min_{\mathbf{B}}\{\mathcal{G}_p(\mathbf{B})\} \tag{5}$$

where

$$\mathcal{G}_p(\mathbf{B}) \stackrel{\text{def}}{=} G_p(\mathbf{z}) \tag{6}$$

and $\mathbf{z}(t) = \mathbf{B}\mathbf{x}(t)$ represents the estimated sources. The approach we choose to solve (5) is inspired from [10]. It is a block technique based on the processing of T received samples and consists in searching the minimum of the sample version of (5). Solutions are obtained iteratively in the form:

$$\mathbf{B}^{(k+1)} = (\mathbf{I} + \boldsymbol{\epsilon}^{(k)})\mathbf{B}^{(k)} \tag{7}$$

$$\mathbf{z}^{(k+1)}(t) = (\mathbf{I} + \boldsymbol{\epsilon}^{(k)})\mathbf{z}^{(k)}(t) \tag{8}$$

where $\mathbf{I}$ denotes the identity matrix. At iteration k, a matrix $\boldsymbol{\epsilon}^{(k)}$ is determined from a local linearization of $G_p(\mathbf{B}\mathbf{x}(t))$. It is an approximate Newton technique with the benefit that $\boldsymbol{\epsilon}^{(k)}$ can be very simply computed (no Hessian inversion) under the additional assumption that $\mathbf{B}^{(k)}$ is close to a separating matrix. This procedure is illustrated in the following steps:
At the $(k+1)^{th}$ iteration, the proposed criterion (4) can be developed as follows:

$$\begin{aligned}\mathcal{J}_p(z_i^{(k+1)}) &= \frac{1}{T}\sum_{t=0}^{T-1}\left|z_i^{(k)}(t) + \sum_{j=1}^{N}\epsilon_{ij}^{(k)}z_j^{(k)}(t)\right|^p \\ &= \frac{1}{T}\sum_{t=0}^{T-1}|z_i^{(k)}(t)|^p\left|1 + \sum_{j=1}^{N}\epsilon_{ij}^{(k)}\frac{z_j^{(k)}(t)}{z_i^{(k)}(t)}\right|^p\end{aligned}$$

Under the assumption that $\mathbf{B}^{(k)}$ is close to a separating matrix, we have

$$|\epsilon_{ij}^{(k)}| \ll 1$$

and thus, a first order approximation of $\mathcal{J}_p(z_i^{(k+1)})$ is given by:

$$\begin{aligned}\mathcal{J}_p(z_i^{(k+1)}) \approx \tfrac{1}{T}\sum_{t=0}^{T-1}|z_i^{(k)}(t)|^p\Bigg(1 + p\sum_{j=1}^{N}\Big\{\Re e(\epsilon_{ij}^{(k)}) \\ \Re e\left(\frac{z_j^{(k)}(t)}{z_i^{(k)}(t)}\right) - \Im m(\epsilon_{ij}^{(k)})\ \Im m\left(\frac{z_j^{(k)}(t)}{z_i^{(k)}(t)}\right)\Big\}\Bigg)\end{aligned} \tag{9}$$

Table 1: Iterative Sparse Blind Separation (ISBS) algorithm

1. Initialize $\mathbf{B}^{(1)}$ randomly ($\mathbf{z}^{(1)}(t) = \mathbf{B}^{(1)}\mathbf{x}(t)$).
2. For $k = 1, \cdots, K$, compute $\boldsymbol{\mathcal{R}}^{(k)}$ by (12).
3. Update the separation matrix $\mathbf{B}^{(k+1)}$ by (15).
4. Update the source estimate (16).

thus,

$$\begin{aligned}\mathcal{J}_p(z_i^{(k+1)}) \approx \tfrac{1}{T}\sum_{t=0}^{T-1}\Big(|z_i^{(k)}(t)|^p + \\ p\sum_{j=1}^{N}\Big\{\Re e(\epsilon_{ij}^{(k)})\Re e\left(|z_i^{(k)}(t)|^{p-1}e^{-\jmath\phi_i^{(k)}(t)}z_j^{(k)}(t)\right) \\ -\Im m(\epsilon_{ij}^{(k)})\Im m\left(|z_i^{(k)}(t)|^{p-1}e^{-\jmath\phi_i^{(k)}(t)}z_j^{(k)}(t)\right)\Big\}\Big)\end{aligned} \tag{10}$$

where $\Re e(x)$ and $\Im m(x)$ denote the real and imaginary parts of x and $\phi_i^{(k)}(t)$ is the argument of the complex number $z_i^{(k)}(t)$.
Using equation (3), minimization of the above criterion (10) is similar to minimization of $G_p(\mathbf{z}^{(k+1)})$. Equation (3) can be rewritten in more compact form as:

$$\mathcal{G}_p\left(\mathbf{I} + \boldsymbol{\epsilon}^{(k)}\right) = \mathcal{G}_p\left(\mathbf{I}\right) + \Re e\left\{Tr\left(\overline{\boldsymbol{\epsilon}}^{(k)}\boldsymbol{\mathcal{R}}^{(k)H}\right)\right\} \tag{11}$$

where $\overline{(\cdot)}$ denotes the conjugate of $(\cdot)$ and the ij^{th} entry of matrix $\boldsymbol{\mathcal{R}}^{(k)}$ is given by:

$$\mathcal{R}_{ij}^{(k)} = \frac{1}{T}\sum_{t=0}^{T-1}|z_i^{(k)}(t)|^{p-1}e^{-\jmath\phi_i^{(k)}(t)}z_j^{(k)}(t) \tag{12}$$

and Tr is the matrix trace operator. Using a gradient technique, $\boldsymbol{\epsilon}^{(k)}$ can be written as:

$$\boldsymbol{\epsilon}^{(k)} = -\mu\overline{\boldsymbol{\mathcal{R}}}^{(k)} \tag{13}$$

where $\mu > 0$ is the gradient step. Replacing (13) into (11) leads to,

$$\mathcal{G}_p\left(\mathbf{I} + \boldsymbol{\epsilon}^{(k)}\right) = \mathcal{G}_p\left(\mathbf{I}\right) - \mu\|\boldsymbol{\mathcal{R}}^{(k)}\|^2 \tag{14}$$

So μ controls the decrement of the criterion. Hence, at the $(k+1)^{th}$ iteration, we have

$$\mathbf{B}^{(k+1)} = (\mathbf{I} - \mu\overline{\boldsymbol{\mathcal{R}}}^{(k)})\mathbf{B}^{(k)} \tag{15}$$

$$\mathbf{z}^{(k+1)} = (\mathbf{I} - \mu\overline{\boldsymbol{\mathcal{R}}}^{(k)})\mathbf{z}^{(k)} \tag{16}$$

This algorithm is summarized in Table 1.

4. SIMULATION RESULTS

We present here some numerical simulations to evaluate the performance of our algorithm. We consider an array of $M = 5$ sensors with half wavelength spacing receiving two audio signals in the presence of stationary complex temporally white noise of covariance $\sigma^2\mathbf{I}$ (σ^2 being the noise power). 10000 samples are used with a sampling frequency of 8Khz. The sources arrive from the directions $\theta_1 = 30$ and $\theta_2 = 45$ degree.
In order to evaluate the performance, the separation quality is measured using two different criteria, the first one is the mean rejection level criterion [3] defined as:

$$\mathcal{I}perf \stackrel{\text{def}}{=} \sum_{p\neq q} \frac{E\left(|(\mathbf{BA})_{pq}|^2\right)\rho_q}{E\left(|(\mathbf{BA})_{pp}|^2\right)\rho_p} \tag{17}$$

where $\rho_i = E(|s_i(t)|^2)$ is the i^{th} source power evaluated here as $\frac{1}{T}\sum_{t=0}^{T-1}|s_i(t)|^2$. The second is the normalized mean square error (NMSE) of the sources defined as:

$$NMSE_i \stackrel{\text{def}}{=} \frac{1}{N_r}\sum_{r=1}^{N_r}\min_{\alpha}\left(\frac{\|\alpha\widehat{\mathbf{s}}_{i,r}-\mathbf{s}_i\|^2}{\|\mathbf{s}_i\|^2}\right) \tag{18}$$

$$NMSE_i = \frac{1}{N_r}\sum_{r=1}^{N_r} 1-\left(\frac{\widehat{\mathbf{s}}_{i,r}\mathbf{s}_i^H}{\|\widehat{\mathbf{s}}_{i,r}\|\|\mathbf{s}_i\|}\right)^2 \tag{19}$$

$$NMSE = \frac{1}{N}\sum_{i=1}^{N} NMSE_i\ . \tag{20}$$

where $\mathbf{s}_i \stackrel{\text{def}}{=} [s_i(0),\ldots,s_i(T-1)]$ and $\widehat{\mathbf{s}}_{i,r}$ is defined similarly and represents the r^{th} estimate of source $\mathbf{s}_i$, α is a scalar factor that compensate for the scale indeterminacy of the BSS problem and N_r is the number of Monte-Carlo runs. Both criteria are estimated over $N_r = 200$ runs. Figure 1 represents the two original sources $(s_1(t),\ s_2(t))$ and the recovered ones $(z_1(t),\ z_2(t))$ by the proposed algorithm in a noiseless case. In Figure 2, the mean rejection level is plotted versus the SNR for the proposed algorithm and the algorithm SOBI [3] which is considered as one of the most performing in separating audio sources. We used SOBI with 6 correlation matrices of respective delays $\tau = 1,\ldots,6$. It is clearly shown that our algorithm (ISBS) performs better in terms of the mean rejection level especially for high SNR. One can observe in Figure 3, that we reach the same conclusion for the $NMSE$. Figure 4 compares the mean rejection level for ISBS and SOBI when the number of sensors increases. For 2 sensors, both algorithms perform equally. When the number of sensors is greater, ISBS has a much lower mean rejection level than SOBI. Figure 5 shows the mean rejection level against the sample size for ISBS and SOBI. When the sample size is small, SOBI outperforms the proposed algorithm ISBS. Whereas, ISBS has a much lower

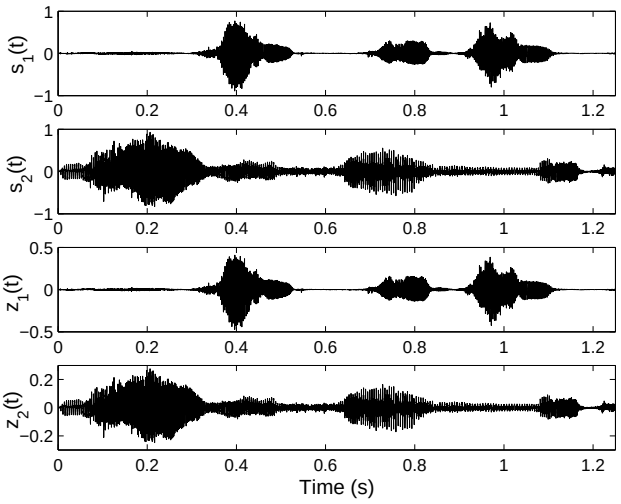

Figure 1: *Blind source separation example for 2 audio sources and 5 sensors: up the two original source signals and bottom the two estimated sources by our algorithm.*

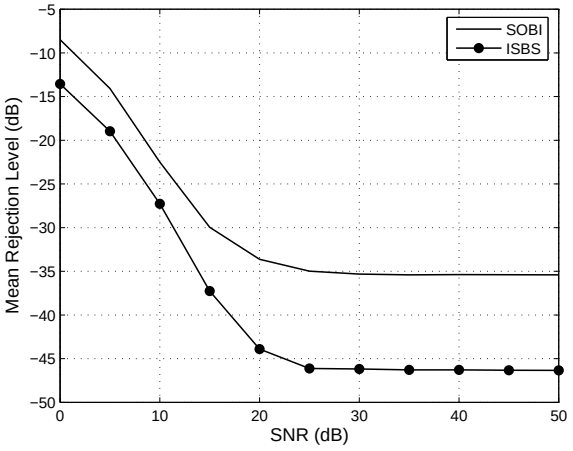

Figure 2: *Mean Rejection Level versus the SNR for 2 audio sources and 5 sensors: comparaison between SOBI and the proposed algorithm.*

mean rejection level when the sample size is larger. It can be explained by the size of the samples: if it increases, the signals present more sparsity, which gives an advantage to ISBS.

5. DISCUSSION

The proposed algorithm outperforms in terms of mean rejection level and $NMSE$ other algorithms that deal with separation from instantaneous mixtures using source independency. It is mostly dedicated to sparse sources in the

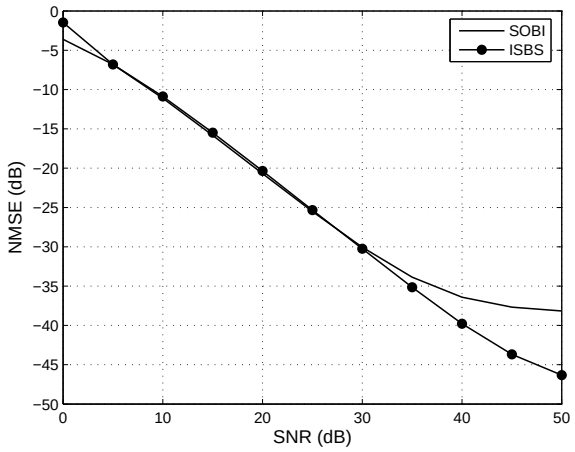

Figure 3: *NMSE versus the SNR for 2 audio sources and 5 sensors: comparaison between SOBI and the proposed algorithm.*

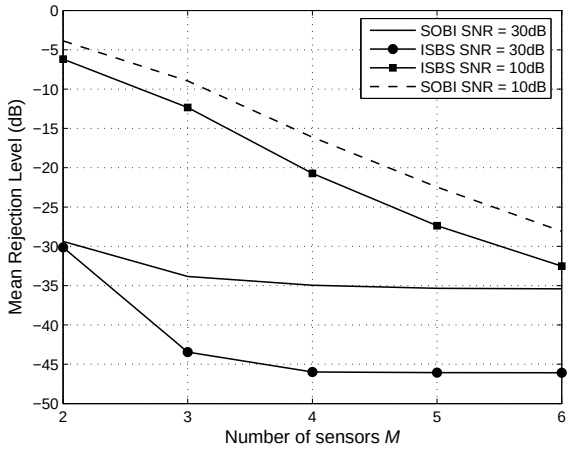

Figure 4: *Mean Rejection Level versus the number of sensors M for 2 audio sources for SNR= 10dB and 30dB.*

time domain. Among its other advantages, the algorithm ISBS shows a low computational complexity and thus can be easily implemented. Furthermore, its flexibility allows us to extend the method to the adaptive case. Nevertheless, the proposed algorithm presents a relative weakness due to the well known disadvantages of the use of gradient techniques such as, the choice of the step gradient μ that the speed convergence depends on and the problem of local minima.

6. CONCLUSION

This paper presents a blind source separation method for sparse sources in the time domain. A sparse contrast function is introduced and an iterative algorithm based on gradient technique is proposed to minimize it and perform BSS. Numerical simulations have been performed to evidence the usefulness of the method. They showed good performance in terms of mean rejection level and $NMSE$ compared to other separation technics (SOBI).

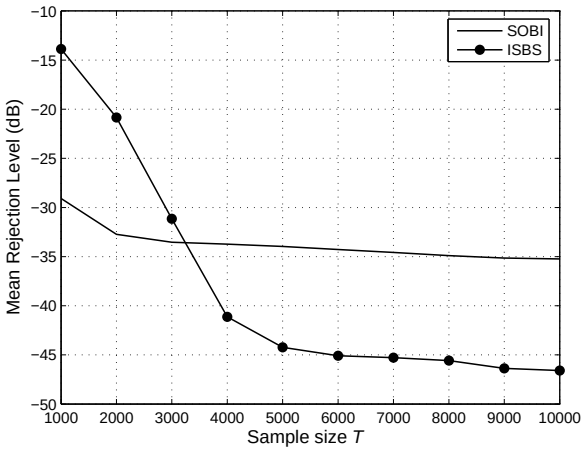

Figure 5: *Mean Rejection Level versus the sample size T for 2 audio sources for SNR=30dB.*

7. REFERENCES

[1] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing*, Wiley & Sons, Ltd., UK, 2003.

[2] J.-F. Cardoso, "Blind signal separation: statistical principles," *Proc. of the IEEE*, vol. 86, pp. 2009–2025, 1998.

[3] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, "A blind source separation technique using second-order statistics," *IEEE Transaction on Signal Processing*, vol. 45, no. 2, pp. 434–444, Feb. 1997.

[4] A. Belouchrani and M. G. Amin, "Blind source separation based on time-frequency signal representations," *IEEE Transaction on Signal Processing*, vol. 46, no. 11, pp. 2888–2897, Nov. 1998.

[5] K. Abed-Meraim, Y. Xiang, J. H. Manton, and Y. Hua, "Blind source separation using second order cyclostationary statistics," *IEEE Transaction on Signal Processing*, vol. 49, no. 4, pp. 694–701, Apr. 2001.

[6] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transaction on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[7] D. T. Pham and J. F. Cardoso, "Blind separation of instantaneous mixtures of non stationary sources," *IEEE Transaction on Signal Processing*, vol. 49, pp. 1837–1848, 2001.

[8] D. Smith, J. Lukasiak, and I. S. Burnett, "An analysis of the limitations of blind signal separation application with speech," *Signal Processing*, vol. 86, pp. 353–359, 2006.

[9] M. Zibulevsky, "Sparse source separation with relative Newton method," in *Proc. ICA*, Apr. 2003, pp. 897–902.

[10] D. T. Pham and P. Garat, "Blind separation of mixture of independent sources through a quasi-maximum likelihood approach," *IEEE Transaction on Signal Processing*, vol. 45, pp. 1712–1725, 1997.