Blind audio source separation using sparsity based criterion for convolutive mixture case

Abdeldjalil Aissa El Bey, Karim Abed-Meraim, Yves Grenier

Blind audio source separation using sparsity based criterion for convolutive mixture case

A. Aïssa-El-Bey, K. Abed-Meraim and Y. Grenier

ENST-Paris, TSI Department, 46 rue Barrault 75634, Paris Cedex 13, France
{elbey, abed, grenier}@tsi.enst.fr

Abstract. In this paper, we are interested in the separation of audio sources from their instantaneous or convolutive mixtures. We propose a new separation method that exploits the sparsity of the audio signals via an ℓ_p-norm based contrast function. A simple and efficient natural gradient technique is used for the optimization of the contrast function in an instantaneous mixture case. We extend this method to the convolutive mixture case, by exploiting the property of the Fourier transform. The resulting algorithm is shown to outperform existing techniques in terms of separation quality and computational cost.

1 Introduction

Blind Source Separation (BSS) is an approach to estimate and recover independent source signals using only the information within the mixtures observed at each channel. Many algorithms have been proposed to solve the standard blind source separation problem in which the mixtures are assumed to be instantaneous. A fundamental and necessary assumption of BSS is that the sources are statistically independent and thus are often separated using higher-order statistical information [1]. If extra information about the sources is available at hand, such as temporal coherency [2], source nonstationarity [3], or source cyclostationarity [4], then one can remain in the second-order statistical scenario, to achieve the BSS.
In the case of non-stationary signals (including audio signals), certain solutions using time-frequency analysis of the observations exist [5]. Other solutions use the statistical independence of the sources assuming a local stationarity to solve the BSS problem [6]. This is a strong assumption that is not always verified [7]. 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. Then, we extend this algorithm to the convolutive mixture case, by transforming the convolutive problem into instantaneous problem in the frequency domain, and separating the instantaneous mixtures in every frequency bin. The use of sparsity to handle this model, has arisen in several papers in the area of source separation [8, 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 relative 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 compare its performance to other existing BSS techniques.

2 Instantaneous mixture case

2.1 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.

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

where $\boldsymbol{s}(t) = [s_1(t), \cdots, s_N(t)]^T$ is the $N \times 1$ sparse source vector, $\boldsymbol{w}(t) = [w_1(t), \cdots, w_M(t)]^T$ is the $M \times 1$ complex noise vector, $\boldsymbol{A}$ 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 $\boldsymbol{B}$ such that $\widehat{\boldsymbol{s}}(t) = \boldsymbol{B}\boldsymbol{x}(t)$ is an estimate of the source signals.
Before proceeding, note that complete blind identification of separating matrix $\boldsymbol{B}$ (or equivalently, the mixing matrix $\boldsymbol{A}$) is impossible in this context, because the exchange of a fixed scalar between the source signal and the corresponding column of $\boldsymbol{A}$ 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 $\boldsymbol{B}$ up to a permutation and scalar shifts of its columns, i.e., $\boldsymbol{B}$ is a separating matrix iff :

$$\boldsymbol{B}\boldsymbol{x}(t) = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{s}(t) \tag{2}$$

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

2.2 Sparsity-based BSS algorithm

Before starting, we propose to use 'an optional' whitening step which set the mixtures to the same energy level and reduces the number of parameters to be estimated. More precisely, the whitening step is applied to the signal mixtures before using our separation algorithm. The whitening is achieved by applying a $N \times M$ matrix $\boldsymbol{W}$ to the signal mixtures in such a way $\mathrm{Cov}(\boldsymbol{W}\boldsymbol{x}) = \boldsymbol{I}$ in the noiseless case, where $\mathrm{Cov}(\cdot)$ stands for the covariance operator. As shown in [2], $\boldsymbol{W}$ can be computed as the inverse square root of the noiseless covariance matrix of the signal mixtures (see [2] for more details). In the following, we apply our separation algorithm on the whitened data :

$$\boldsymbol{x}_w(t) = \boldsymbol{W}\boldsymbol{x}(t).$$

We propose an iterative algorithm for the separation of sparse audio signals, namely the ISBS for Iterative Sparse Blind Separation. It is well known that

audio signals are characterized by their sparsity property in the time domain [8, 9] which is measured by their ℓ_p norm where $0 \leq p < 2$. More specifically, one can define the following sparsity based contrast function

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

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 $\boldsymbol{B}$ such as,

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

where

$$\mathcal{G}_p(\boldsymbol{B}) \triangleq G_p(\boldsymbol{z}) \ , \tag{6}$$

and $\boldsymbol{z}(t) \triangleq \boldsymbol{B}\boldsymbol{x}_w(t)$ 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 iteratively the minimum of (5) in the form :

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

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

where $\boldsymbol{I}$ denotes the identity matrix. At iteration k, a matrix $\boldsymbol{\epsilon}^{(k)}$ is determined from a local linearization of $G_p(\boldsymbol{B}^{(k+1)}\boldsymbol{x}_w)$. 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 $\boldsymbol{B}^{(k)}$ is close to a separating matrix. This procedure is illustrated in the following :

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 $\boldsymbol{B}^{(k)}$ 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 :

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

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 (9), equation (3) can be rewritten in more compact form as :

$$\mathcal{G}_p\left(\boldsymbol{B}^{(k+1)}\right) = \mathcal{G}_p\left(\boldsymbol{B}^{(k)}\right) + \Re e\left\{Tr\left(\overline{\boldsymbol{\epsilon}}^{(k)}\boldsymbol{\mathcal{R}}^{(k)H}\boldsymbol{D}^{(k)H}\right)\right\} \tag{10}$$

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

$$\mathcal{R}_{ij}^{(k)} = \frac{1}{T}\sum_{t=0}^{T-1}|z_i^{(k)}(t)|^{p-1}e^{-\jmath\phi_i^{(k)}(t)}z_j^{(k)}(t) \tag{11}$$

$$\boldsymbol{D}^{(k)} = \left[\mathrm{diag}\left(\mathcal{R}_{11}^{(k)},\ldots,\mathcal{R}_{NN}^{(k)}\right)\right]^{\frac{1}{p}-1} . \tag{12}$$

Using a gradient technique, $\boldsymbol{\epsilon}^{(k)}$ can be chosen as :

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

where $\mu > 0$ is the gradient step. Replacing (13) into (10) leads to,

$$\mathcal{G}_p\left(\boldsymbol{B}^{(k+1)}\right) = \mathcal{G}_p\left(\boldsymbol{B}^{(k)}\right) - \mu\|\boldsymbol{D}^{(k)}\boldsymbol{\mathcal{R}}^{(k)}\|^2 . \tag{14}$$

So μ controls the decrement of the criterion. Now, to avoid the algorithm's convergence to the trivial solution $\boldsymbol{B} = \boldsymbol{0}$, one normalizes the outputs of the separating matrix to unit-power, i.e. $\rho_{z_i}^{(k+1)} \triangleq \frac{1}{T}\sum_{t=0}^{T-1}|z_i^{(k+1)}(t)|^2 = 1, \quad \forall\ i$. Using first order approximation, this normalization leads to :

$$\epsilon_{ii}^{(k)} = \frac{1-\rho_{z_i}^{(k)}}{2\rho_{z_i}^{(k)}} . \tag{15}$$

After convergence of the algorithm, the separation matrix $\boldsymbol{B} = \boldsymbol{B}^{(\mathcal{K})}$ is applied to the whitened signal mixtures $\boldsymbol{x}_w$ to obtain an estimation of the original source signals. $\mathcal{K}$ denotes here the number of iterations that can be either chosen a priori or given by a stopping criterion of the form $\|\boldsymbol{B}^{(k+1)} - \boldsymbol{B}^{(k)}\| < \delta$ where δ is a small threshold value.

3 Convolutive mixture case

Unfortunately, instantaneous mixing is very rarely encountered in real-world situations, where multipath propagation with large channel delay spread occurs, in which case convolutive mixtures are considered. In this case, the signal can be modeled by the following equation :

$$\boldsymbol{x}(t) = \sum_{l=0}^{L} \boldsymbol{H}(l)\boldsymbol{s}(t-l) + \boldsymbol{w}(t) \tag{16}$$

where $\boldsymbol{H}(l)$ are $M \times N$ matrices for $l \in [0, L]$ representing the impulse response coefficients of the channel and the polynomial matrix $\boldsymbol{H}(z) = \sum_{l=0}^{L} \boldsymbol{H}(l)z^{-l}$ is assumed to be irreducible (i.e. $\boldsymbol{H}(z)$ is of full column rank for all z).
If we apply a short time Fourier transform (STFT) to the observed data $\boldsymbol{x}(t)$, the model in (16) (in the noiseless case) becomes approximately

$$\mathcal{S}_{\boldsymbol{x}}(t, f) \approx \boldsymbol{H}(f)\mathcal{S}_{\boldsymbol{s}}(t, f) \tag{17}$$

where $\mathcal{S}_{\boldsymbol{x}}(t, f)$ is the mixture STFT vector, $\mathcal{S}_{\boldsymbol{s}}(t, f)$ is the source STFT vector and $\boldsymbol{H}(f)$ is the channel Fourier Transform matrix. It shows that, for each frequency bin, the convolutive mixtures reduce to simple instantaneous mixtures. Therefore we can apply our ISBS algorithm for each frequency and separate the signals. As a result, in each frequency bin, we obtain the STFT source estimate

$$\mathcal{S}_{\widehat{\boldsymbol{s}}}(t, f) = \boldsymbol{B}(f)\mathcal{S}_{\boldsymbol{x}}(t, f) \ . \tag{18}$$

It seems natural to reconstruct the separated signals by aligning these $\mathcal{S}_{\widehat{\boldsymbol{s}}}(t, f)$ obtained for each frequency bin and applying the inverse short time Fourier transform. For that we need first to solve the permutation and scaling ambiguities as shown next.

3.1 Removing the scaling end permutation ambiguities

In this stage, the output of the separation filter is processed with the permutation matrix $\boldsymbol{\Pi}(f)$ and the scaling matrix $\boldsymbol{\mathcal{C}}(f)$.

$$\boldsymbol{G}(f) = \boldsymbol{\Pi}(f)\boldsymbol{\mathcal{C}}(f)\boldsymbol{B}(f) \ . \tag{19}$$

The scaling matrix $\boldsymbol{\mathcal{C}}(f)$ is a $N \times N$ diagonal matrix found as in [11] by $\boldsymbol{\mathcal{C}}(f) = \mathrm{diag}[\boldsymbol{B}(f)^{\#}]$. For the permutation matrix $\boldsymbol{\Pi}(f)$, we exploit the continuity property of the acoustic filter in the frequency domain [12]. To align the estimated sources at two successive frequency bins, we test of the closeness of $\boldsymbol{G}(f_n)\boldsymbol{G}(f_{n-1})^{\#}$ to a diagonal matrix. Indeed, by using the representation (19), one can find the permutation matrix by minimizing :

$$\boldsymbol{\Pi}(f_n) = \arg\min_{\widetilde{\boldsymbol{\Pi}}} \left\{ \sum_{i \neq j} \left(\widetilde{\boldsymbol{\Pi}}\boldsymbol{\mathcal{C}}(f_n)\boldsymbol{B}(f_n)\boldsymbol{G}(f_{n-1})^{\#} \right)_{ij}^{2} \right\} \ . \tag{20}$$

In our simulations, we have used an exhaustive search to solve (20). However, when the number of sources is large, the exhaustive search becomes prohibitive. In that case, one can estimate $\boldsymbol{\Pi}(f_n)$ as the matrix with ones at the ij^{th} entry satisfying $|\boldsymbol{\mathcal{M}}(f_n)|_{ij} = \max_k |\boldsymbol{\mathcal{M}}(f_n)|_{ik}$ and zeros elsewhere with $\boldsymbol{\mathcal{M}}(f_n) = \boldsymbol{\mathcal{C}}(f_n)\boldsymbol{B}(f_n)\boldsymbol{G}(f_{n-1})^{\#}$. This solution has the advantage of simplicity but may lead to erroneous solution in difficult context. An alternative solution would be to decompose $\boldsymbol{\Pi}(f_n)$ as product of elementary permutations[1] $\boldsymbol{\Pi}_{(pq)}$. The latter is considered at a given iteration, only if it decrease criterion (20), if

$$|\boldsymbol{\mathcal{M}}(f_n)|^2_{pq} + |\boldsymbol{\mathcal{M}}(f_n)|^2_{qp} > |\boldsymbol{\mathcal{M}}(f_n)|^2_{pp} + |\boldsymbol{\mathcal{M}}(f_n)|^2_{qq}$$

Finally, we obtain :

$$\boldsymbol{\Pi}(f_n) = \prod_{\text{nb of iterations}} \prod_{1 \leq p < q \leq N} \widetilde{\boldsymbol{\Pi}}_{(pq)} \ , \tag{21}$$

$\widetilde{\boldsymbol{\Pi}}_{(pq)}$ being either the identity matrix or the above permutation matrix $\boldsymbol{\Pi}_{(pq)}$ depending on the binary decision rule define above. We stop the iterative process, when all matrices $\widetilde{\boldsymbol{\Pi}}_{(pq)}$ are equal to the identity. We have observed that one or, at most, two iterations are sufficient to get the desired permutation. Finally, we apply the updated separation matrix $\boldsymbol{G}(f)$ to the frequency domain mixture :

$$\mathcal{S}_{\widehat{\boldsymbol{s}}}(t,f) = \boldsymbol{G}(f)\mathcal{S}_{\boldsymbol{x}}(t,f) \ . \tag{22}$$

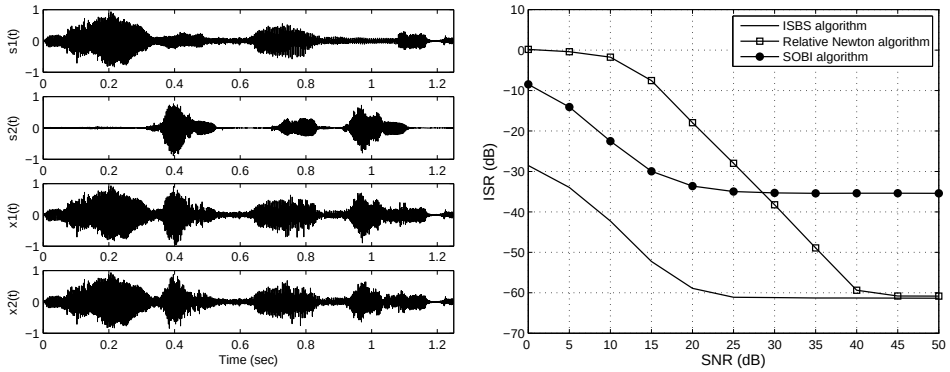
4 Simulation results

We present here some numerical simulations to evaluate the performance of our algorithm. We consider an array of $M = 2$ sensors receiving two audio signals in the presence of stationary 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 (this represents 1.25sec recording). In order to evaluate the performance in the instantaneous mixture case, the separation quality is measured using the *Interference to Signal Ratio* (ISR) criterion [2] defined as :

$$ISR \stackrel{\text{def}}{=} \sum_{p \neq q} \frac{E\left(|(\boldsymbol{BA})_{pq}|^2\right)\rho_q}{E\left(|(\boldsymbol{BA})_{pp}|^2\right)\rho_p} \tag{23}$$

where $\rho_i = E(|s_i(t)|^2)$ is the i^{th} source power evaluated here as $\frac{1}{T}\sum_{t=0}^{T-1}|s_i(t)|^2$. Fig. 1-(a) represents the two original sources and their mixtures in the noiseless case. In Fig. 1-(b), we compare the performance of the proposed algorithm in instantaneous mixture case, to the Relative Newton algorithm developed by Zibulevsky et al. in [9] where the case of sparse sources is considered and to SOBI algorithm developed by Belouchrani et al. in [2]. We plot the residual interference between separated sources (ISR) versus the SNR. It is clearly shown that our algorithm (ISBS) performs better in terms of ISR especially for low SNRs as compared to the two other methods. In Fig. 2-(a), we represent the evolution of

[1] $\boldsymbol{\Pi}_{(pq)}$ is defined such as way that for a given vector $\boldsymbol{y}$, $\widetilde{\boldsymbol{y}} = \boldsymbol{\Pi}_{(pq)}\boldsymbol{y}$ iff $\widetilde{y}(k) = y(k)$, for $k \notin \{p,q\}$, $\widetilde{y}(p) = y(q)$ and $\widetilde{y}(q) = y(p)$.



(a) (b)

Fig. 1. (a) Up the two original source signals and bottom the two signal mixtures. (b) Interference to Signal Ratio (ISR) versus SNR for 2 audio sources and 2 sensors in instantaneous mixture case.

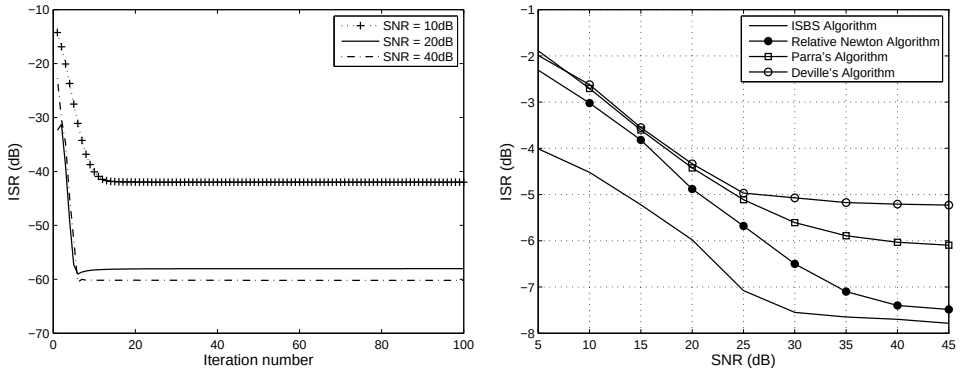
the ISR as a function of the iteration number. A fast convergence rate is observed. In Fig. 2-(b), we compare, in the 2×2 convolutive mixture case the separation performance of our algorithm, Deville's algorithm in [13], Parra's algorithm in [14] and extended version of Zibulevsky's algorithm to the convolutive mixture case. The filter coefficients are chosen randomly and the channel order is $L = 128$. We use in this experiment the ISR criterion defined for the convolutive case in [14] that takes into account the fact the source estimates are obtained up to a scalar filter. We observe a significant performance gain in favor of the proposed method especially at low SNR values. Moreover, the complexity of the proposed algorithm is equal to $2N^2T + \mathcal{O}(N^2)$ flops per iteration whereas the complexity of the Relative Newton algorithm in [9] is $2N^4 + N^3T + N^6/6$.

5 Conclusion

This paper presents a blind source separation method for sparse sources in instantaneous mixture case and its extension to the convolutive mixture case. A sparse contrast function is introduced and an iterative algorithm based on gradient technique is proposed to minimize it and perform the BSS. Numerical simulation results have been given evidence the usefulness of the method. The proposed technique outperforms existing solutions in terms of separation quality and computational cost in both instantaneous and convolutive mixture cases.

References

1. Cardoso, J.F.: Blind signal separation : statistical principles. Proceedings of the IEEE **86**(10) (Oct. 1998) 2009–2025



(a) (b)

Fig. 2. (a) ISR as a function of the iteration number for 2 audio sources and 2 sensors in instantaneous mixture case. (b) ISR versus SNR for 2×2 convolutive mixture case.

2. Belouchrani, A., Abed-Meraim, K., Cardoso, J.F., Moulines, E.: A blind source separation technique using second-order statistics. IEEE T-SP **45**(2) (Feb. 1997)
3. Belouchrani, A., Amin, M.G.: Blind source separation based on time-frequency signal representations. IEEE Transactions on Signal Processing **46**(11) (Nov. 1998)
4. Abed-Meraim, K., Xiang, Y., Manton, J.H., Hua, Y.: Blind source separation using second order cyclostationary statistics. IEEE T-SP **49**(4) (April 2001) 694–701
5. Yilmaz, O., Rickard, S.: Blind separation of speech mixtures via time-frequency masking. IEEE Transactions on Signal Processing **52**(7) (July 2004) 1830–1847
6. Pham, D.T., Cardoso, J.F.: Blind separation of instantaneous mixtures of non stationary sources. IEEE Transactions on Signal Processing **49** (2001) 1837–1848
7. Smith, D., Lukasiak, J., Burnett, I.S.: An analysis of the limitations of blind signal separation application with speech. Signal Processing **86**(2) (Feb. 2006) 353–359
8. Cichocki, A., Amari, S.: Chapter 2. In: Adaptive Blind Signal and Image Processing. Wiley & Sons, Ltd., UK (2003)
9. Zibulevsky, M.: Sparse source separation with relative Newton method. In: Proc. ICA. (April 2003) 897–902
10. Pham, D.T., Garat, P.: Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. IEEE T-SP **45**(7) (July 1997) 1712–1725
11. Murata, N., Ikeda, S., Ziehe, A.: An approach to blind source separation based on temporal structure of speech signals. Neurocomputing **41**(1-4) (Oct. 2001) 1–24
12. Pham, D.T., Servière, C., Boumaraf, H.: Blind separation of convolutive audio mixtures using nonstationarity. In: Proc. ICA, Nara, Japan (April 2003) 981–986
13. Albouy, B., Deville, Y.: Alternative structures and power spectrum criteria for blind segmentation and separation of convolutive speech mixtures. In: Proc. ICA, Nara, Japan (April 2003) 361–366
14. Parra, L., Spence, C.: Convolutive blind separation of non-stationary sources. IEEE Transactions on Speech and Audio Processing **8**(3) (May 2000) 320–327