2007 IEEE International Conference on Signal Processing and Communications (ICSPC 2007), 24-27 November 2007, Dubai, United Arab Emirates

ADAPTIVE SPARSE SOURCE SEPARATION WITH APPLICATION TO SPEECH SIGNALS

Elham Azizi, G. Hosein Mohimani, Massoud Babaie-Zadeh

Electrical Engineering Department, Sharif University of Technology, Tehran, Iran

ABSTRACT

In this paper, a Sparse Component Analysis algorithm is presented for the case in which the number of sources is less than or equal to the number of sensors, but the channel (mixing matrix) is time-varying. The method is based on a smoothed ℓ^0 norm for the sparsity criteria, and takes advantage of the idea that sparsity of the sources is decreased when they are mixed. The method is able to separate synthetic and speech data, which require very weak sparsity restrictions. It can separate up to 50 mixed signals while being adaptive to channel variation and robust against noise.

Index Terms— Blind Source Separation; Sparse Component Analysis; Adaptive Source Separation; Smoothed ℓ^0 Norm.

1. INTRODUCTION

The problem of Blind Source Separation (BSS) has been extensively studied in the last two decades, because of its many potential applications in science and technology [1, 2, 3]. This problem consists of separating a set of mixed signals from their mixtures taking advantage of a very weak *a priori* information about the source signals [4]. In early literature, the priors 'non-stationarity' of the source signals [5] and their 'temporal correlation' [6] have already been considered.

Independent Component Analysis (ICA) has been successfully used to solve blind source separation problems in several application areas which consider statistical independence of sources as a priori. However, in some cases, e.g. sound and music signals, independency is usually not a valid assumption. For example, the fundamental frequencies of the sources in a music signal are often in a harmonic relationship, and the sources have dependencies because of rhythmic concordance in time domain [7].

Another prior information is the *sparsity* of source signals [8, 9, 10]. A signal is sparse when it is zero or nearly zero (inactive) in most of it's samples. Such a signal has a probability density function with a sharp peak at zero and fat tails [11, 12]. This prior also permits source separation for the case in which the number of sources exceeds the number of sensors [8, 10, 13, 14]. Moreover, it is a practical assumption for many sources. Non-sparse sources in time domain may be sparse in another linear transformed domain; and since the mixing system is identical in both domains, these approaches may be used in the transformed domain. For instance, speech sources may not be sparse enough in time domain, whereas they possess sparsity in time-frequency (using Short-Time Frequency Transform) or time-scale (utilizing wavelet packet) domains [9]. These methods which mainly rely on the sparsity of sources are usually called Sparse Component Analysis (SCA) [8].

On the other hand, a natural environment is confounded by signal reverberations in sensors or channel variations might occur in many cases. In some applications, it is required to obtain an estimation of the channel immediately after applying the mixtures, i.e. an on-line or instantaneous estimation. In some others, the channel varies with time and a separating system with the ability to track channel variations is required.

In the present paper, we address the issue of BSS for instantaneous but time-dependent channels. The number of sensors is greater than or equal to the number of sources. Considering sparsity as a priori, a reasonable prediction of the mixture matrix based on previous observations is obtained through an adaptive procedure which is able to track variations in channels (mixing matrices). The basic idea is the fact that the sparsity is decreased in mixtures of sparse sources. The proposed algorithm shows significant abilities in robustness against noise and satisfactory separation for medium dimension of mixing matrices.

The problem can be stated as follows. Consider the linear model:

$$\mathbf{X} = \mathbf{AS}+\mathbf{N} \tag{1}$$

where $\mathbf{A} = [\mathbf{a}_1 \dots \mathbf{a}_n] \in \mathbb{R}^{m\times n}$ is the mixing matrix, $\mathbf{S} = [\mathbf{s}_1 \dots \mathbf{s}_T] \in \mathbb{R}^{n\times T}$ and $\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_T] \in \mathbb{R}^{m\times T}$ are the matrices of n sources and m observed signals and $\mathbf{N}$ is the matrix of Additive White Gaussian Noise. Each column of $\mathbf{S}$, $\mathbf{X}$ and $\mathbf{N}$ corresponds to an instant of 'time' and T is the number of time samples. Sparsity of source signals implies that in each column of $\mathbf{S}$, there are few significant values (active sources). The intention of SCA is then to estimate $\mathbf{A}$ and $\mathbf{S}$, only from $\mathbf{X}$ and the sparsity assumption.

2. THEORY AND MATERIAL

In this section, we discuss a new SCA method for detecting the mixing matrix $\mathbf{A}$ and the sources $\mathbf{S}$ in the com-

This work has been partially funded by Iran National Science Foundation (INSF).

1-4244-1236-6/07/$25.00 ©2007 IEEE

plete case (number of mixtures are greater than or equal to sources) using the model $\mathbf{AS} = \mathbf{X}$. The over-complete problem, is not considered. The basic idea is the fact that sparsity is decreased in mixtures of sparse sources. In fact, sparsity is decreased by mixing sparse sources, only when uncorrelated sparse signals are mixed. For example, the addition of a sparse source $\mathbf{s}$ and $-\mathbf{s}$ would be sparser than the both sources (zero), due to correlation between $\mathbf{s}$ and $-\mathbf{s}$. In the continuation, we assume the sources to be uncorrelated. The idea is then to multiply the mixture matrix $\mathbf{X}$ by a separating matrix $\mathbf{B}$ and try to maximize the sparsity. To accomplish this goal, a smooth measure of sparsity is defined as in [15]. A sparse signal is inactive in a large percent of its samples, and hence it's ℓ^0 norm is small, where the ℓ^0 norm is defined as:

$$\|\mathbf{s}\|_0 = \sum_{i=1}^{n} \nu(s_i) \tag{2}$$

where:

$$\nu(s) = \begin{cases} 1 & s \neq 0 \\ 0 & s = 0 \end{cases} \tag{3}$$

In order to find the sparsest solution, it is required to minimize the ℓ^0 norm. However, optimization of the ℓ^0 norm is not straightforward. Considering the discontinuities of the function ν, gradient-based methods are impossible to be applied directly. Thus, a smooth estimation of ν is employed in (2). This may also provide more robustness against noise. Various functions can be utilized for this aim. Here, we use a zero-mean Gaussian family of functions because of their differentiability. By defining:

$$f_\sigma(s) = \exp(-s^2/2\sigma^2), \tag{4}$$

we have:

$$\lim_{\sigma \to 0} f_\sigma(s) = \begin{cases} 1 & s = 0 \\ 0 & s \neq 0 \end{cases} \tag{5}$$

Consequently, $\lim_{\sigma \to 0} f_\sigma(s) = 1 - \nu(s)$, and therefore if we define:

$$F_\sigma(\mathbf{s}) = \sum_{i=1}^{n} f_\sigma(s_i), \tag{6}$$

we have:

$$\lim_{\sigma \to 0} F_\sigma(\mathbf{s}) = \sum_{i=1}^{n} (1 - \nu(s_i)) = n - \|\mathbf{s}\|_0. \tag{7}$$

We then take $n - F_\sigma(\mathbf{s})$ as an approximation to $\|\mathbf{s}\|_0$. The value of σ indicates a trade-off between the accuracy and the smoothness of the approximation: the smaller values of σ result in better approximations, while the larger values lead to smoother estimations.

3. THE ALGORITHM

Similar to ICA procedures, whitening is performed in the first stage of the algorithm. The energies of the sources are assumed to be equal to one because their energy is insignificant (and impossible to calculate). As a result of uncorrelated sources we have $E(\mathbf{S}\mathbf{S}^T) = \mathbf{I}$. Whitening $\mathbf{X}$ would achieve $\widetilde{\mathbf{X}} = \mathbf{W}\mathbf{X}$ where $\mathbf{W}$ is the whitening matrix. In the continuation, the separating matrix $\mathbf{B}$ represents the inverse of the mixing matrix $\mathbf{A}$. By defining $\widetilde{\mathbf{B}} = \mathbf{B}\mathbf{W}^{-1}$, $\widetilde{\mathbf{B}}$ is orthonormal [2, 16]. By using the sparsity measure introduced in the previous section we attempt to find

$$\underset{\widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^T = \mathbf{I}_m}{\operatorname{argmax}} H_\sigma(\widetilde{\mathbf{B}}) = E\{F_\sigma(\widetilde{\mathbf{B}}\widetilde{\mathbf{X}})\}$$

for some small value of σ, where the expectation is calculated by averaging in time. The relative gradient [17] of the function H_σ can be calculated as:

$$\nabla_{\mathbf{B}} H_\sigma \triangleq \frac{\partial H_\sigma}{\partial \mathbf{B}} \mathbf{B}^T = E\{(\mathbf{Y} .* f_\sigma(\mathbf{Y}))\mathbf{Y}^T\} \tag{8}$$

where $\mathbf{Y} = \widetilde{\mathbf{B}}\widetilde{\mathbf{X}}$ and .* denotes element-wise multiplication.

For escaping local maxima, the idea is to apply a gradual decline in the value of σ: for each value of σ a steepest ascent algorithm is implied to maximize H_σ, and the initial value of this steepest ascent algorithm is the maximizer of H_σ obtained for the previous (larger) value of σ. So, the steepest ascent algorithm is initialized not far from the actual maximum . The natural (or relative) gradient [17] is applied to update $\widetilde{\mathbf{B}}$.

The algorithm shown in Fig. 1 is very similar to fast ICA methods. The main difference is in the optimization criteria, which is maximizing the sparsity criteria defined by (6), compared to optimizing an independence criteria in ICA methods. Note that the whitening step is an essential part of the presented algorithm. Emitting the whitening step (and orthonormality assumption of $\widetilde{\mathbf{B}}$), results in repeated sources. Discussion on the other steps can be found in [2].

4. ADAPTIVE CASE

The mentioned algorithm has been modified to achieve an adaptive SCA method which is able to track channel variations occurred with time and consequently estimate the channel mixing matrix immediately after applying the mixtures. In other words an on-line or instantaneous estimation is intended.

All the steps of Fig. 1 may be done on-line, except the whitening part which needs a modification. An on-line whitening algorithm which has been presented in [2] is applied in the first stage. The adaptive version of the algorithm is shown in Fig.2 [1].

5. EXPERIMENTAL RESULTS

In order to justify the performance of the presented method in the adaptive case, a number of experiments were con-

[1] Initializations of σ, μ and $\widetilde{\mathbf{B}}$ from the previous algorithm should be repeated here.

1. Apply whitening to the mixture matrix.
   (a) Let $\mathbf{E}$ and $\mathbf{V}$ be the eigenvalue and eigenvector matrices of the covariance matrix of the mixtures.
   (b) Set the whitening matrix $\mathbf{W} = \mathbf{E}^{(-1/2)}\mathbf{V}^T$.
   (c) Set $\widetilde{\mathbf{X}} = \mathbf{W}\mathbf{X}$
2. Start with $\widetilde{\mathbf{B}} = \mathbf{I}_m$ (for the case where $m = n$).
3. Choose a suitable decreasing sequence of $[\sigma_1 \dots \sigma_l]$, a value for L, number of repetitions (usually about 100), and a value for the factor μ (about 0.001).
4. For $i = 1 \dots l$ set $\sigma = \sigma_i$ and maximize H_σ by starting from the current $\widetilde{\mathbf{B}}$ by repeating the following loop L times:
   (a) Set $\mathbf{Y} = \widetilde{\mathbf{B}}\widetilde{\mathbf{X}}$ (estimation of the sources).
   (b) Set $\mathbf{D} = E\{(\mathbf{Y}. * f_\sigma(\mathbf{Y}))\mathbf{Y}^T\}$ where .* denotes element-wise multiplication.
   (c) Set $\widetilde{\mathbf{B}} = (\mathbf{I}_m - \mu\mathbf{D})\widetilde{\mathbf{B}}$.
   (d) Orthonormalize $\widetilde{\mathbf{B}}$; Set $\widetilde{\mathbf{B}} = \frac{3}{2}\widetilde{\mathbf{B}} - \frac{1}{2}\widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^T\widetilde{\mathbf{B}}$. (refer to [2], sec. 6.5).
5. Set $\mathbf{B} = \widetilde{\mathbf{B}}\mathbf{W}$

Figure 1. The final algorithm.

ducted. The first experiment analyzes the method on synthetic data and the second experiment, simulates the method with speech signals. To evaluate the performance of the algorithm, let $\mathbf{C} = \mathbf{BA}$ be the global mixing-separating matrix. SNR (Signal to Noise Ratio in dB) is defined as $10 \log \frac{\sum_{i=j}^{n} c_{ij}^2}{\sum_{i \neq j} c_{ij}^2}$.

Synthetic Data: Sparse sources are artificially created using the Mixture of Gaussians (MoG) model

$$s_i \sim p \cdot \mathcal{N}(0, \sigma_{\text{on}}) + (1-p) \cdot \mathcal{N}(0, \sigma_{\text{off}}), \qquad (9)$$

where p denotes the probability of activity of the sources. σ_{on} and σ_{off} are the standard deviations of the sources in active and inactive mode, respectively. The parameters required to satisfy the sparsity conditions are $\sigma_{\text{off}} \ll \sigma_{\text{on}}$ and $p \ll 1$. In the simulation σ_{on} is set to 1 and σ_{off} is set to zero (a spiky model). Each column of the mixing matrix is randomly generated using the normal distribution which is then normalized to unity. The model (1) generates the mixtures, where $\mathbf{N}$ represents an Additive White Gaussian Noise, where $\sigma_n \mathbf{I}_m$ is the covariance matrix of $\mathbf{N}$. In this experiment, the mixing matrix changes linearly with time according to $\mathbf{A} = \mathbf{A}_1 + \mathbf{A}_2 t/T$ where $\mathbf{A}_1$ and $\mathbf{A}_2$ are chosen randomly and T is set to 100000. The parameter μ was set to 0.001 for synthetic signals. Fig. 3 shows that the algorithm will converge after some time[2]. It can completely separate up to 50 sources while being robust against noise and working for weak restrictions on sparsity (even for $p = 0.5$, see Fig. 3).

Speech Data: In the second experiment, 8 typical speech signals (with 16KHz sampling rate and without any pre-processing) are mixed through an arbitrary mixing matrix, with diagonal elements 1, and all other elements 0.5. The experiment was performed 50 times for 50 random shifts of speech signals (to create different signals). The average instantaneous SNR (Fig. 4) grows rapidly and remains stable. Therefore, except for the first samples of signals, the sources are recovered perfectly (see Fig. 4). By increasing the amount of μ the algorithm converges more rapidly, but it converges to a smaller value of SNR[3].

[2]To have a perception of SNR values, it should be noted that for a normalized mixing-separating matrix with 0.9902 on its main axis and 0.02 on all other elements, SNR is approximately 17dB. Whereas a main axis of 0.9367 and other elements of 0.05, result in approximately 8.5dB

1. Initialize the whitening matrix $\mathbf{W} = \mathbf{I}_m$.
2. Repeat the following steps for $t = 1 \dots T$ and in each time sample, for $[\sigma_1 \dots \sigma_l]$:
   (a) Set $\widetilde{\mathbf{X}}(t) = \mathbf{W}\mathbf{X}(t)$.
   (b) Set $\mathbf{W} = \mathbf{W} + \mu(\mathbf{I}_m - \widetilde{\mathbf{X}}^T(t)\widetilde{\mathbf{X}}(t))$. (refer to [2], sec. 6.4)
   (c) Set $\mathbf{Y}(t) = \widetilde{\mathbf{B}}\widetilde{\mathbf{X}}(t)$ (estimation of the sources).
   (d) Set $\mathbf{D} = [\mathbf{Y}(t). * f_\sigma(\mathbf{Y}(t))]\mathbf{Y}(t)^T$ where .* denotes element-wise multiplication.
   (e) Set $\widetilde{\mathbf{B}} = (\mathbf{I}_m - \mu\mathbf{D})\widetilde{\mathbf{B}}$.
   (f) Orthonormalize $\widetilde{\mathbf{B}}$; Set $\widetilde{\mathbf{B}} = \frac{3}{2}\widetilde{\mathbf{B}} - \frac{1}{2}\widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^T\widetilde{\mathbf{B}}$.
3. Set $\mathbf{B} = \widetilde{\mathbf{B}}\mathbf{W}$

Figure 2. The adaptive algorithm.

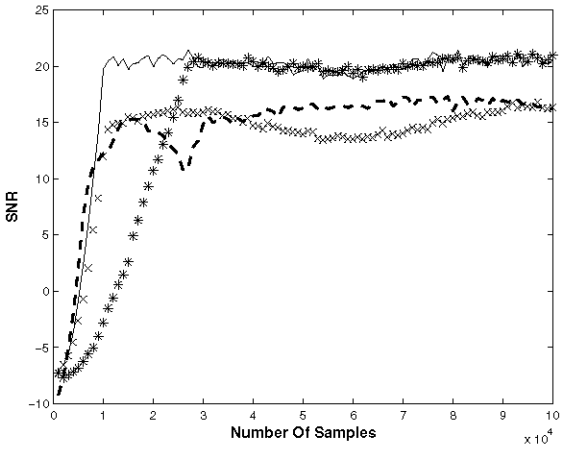

Figure 3. Synthetic source separating results: SNR shown for $n = 30$, $p = .1$, $\sigma_n = .01$ (solid line); scale comparison: $n = 50$, $p = .1$, $\sigma_n = .01$ (dashed line); Noise robustness: $n = 30$, $p = .1$, $\sigma_n = .05$ (crosses); and sparsity decreased with $n = 30$, $p = .5$, $\sigma_n = .01$ (asterisks).

6. DISCUSSION AND CONCLUSIONS

In this paper, an adaptive sparse source separation method is proposed. It has been shown that the method can track

[3]The reason for smaller μ in speech experiment (0.0002 for speech, compared to 0.001 for synthetic) lies in temporal correlation of speech signals in short-length time windows.

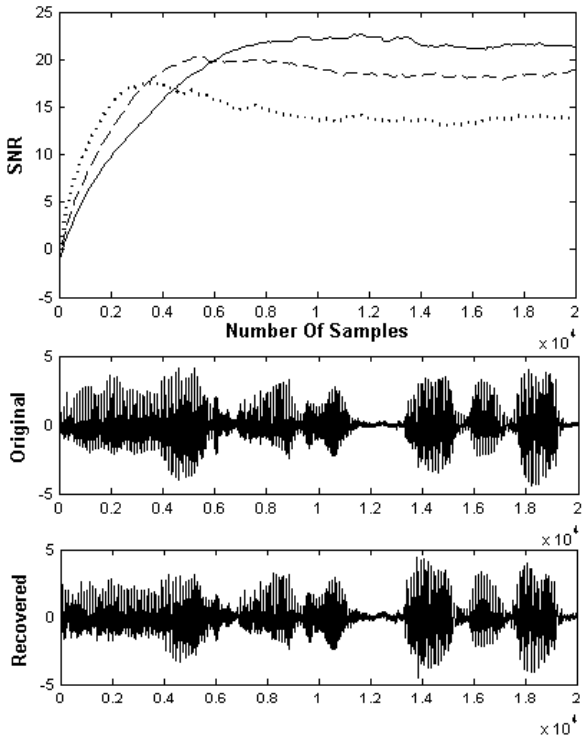

Figure 4. Speech Separation results: Top figure shows the average instantaneous SNR in 50 experiments for $\mu = 0.0002$ (solid line), $\mu = 0.0003$ (dashed line) and $\mu = 0.0005$ (dotted line) . Below, the original and online-recovery of one of the eight signals are depicted.

channel variations and performs separation in medium scale problems, while being robust against noise and requiring a very weak restriction on sparsity. Also, the algorithm can work with un-preprocessed mixtures of speech signals.

7. REFERENCES

[1] A. Cichocki and S. Amari, "Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications," John Wiley and sons, 2002.

[2] A. Hyvarinen, J. Karhunen, and E. Oja, "Independent Component Analysis", John Wiley and Sons, 2001.

[3] C. Jutten, and J. Hérault, "Blind Separation of Sources, Part I: an Adaptive Algorithm Based on a Neuromimetic Architecture", *Signal Processing*, Vol. 24, No. 1, 1991, pp. 1–10.

[4] M. Babaie-Zadeh and C. Jutten, "Semi-Blind Approaches for Source Separation and Independent component Analysis," *Proc. of ESANN'06*, April 2006, pp. 301–312.

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

[6] L. Molgedey and H.G. Schuster, "Separation of a mixture of independent signals using time delayed correlations," *Physical Review Letters*, vol. 72, no. 23, June 1994, pp. 3634–3637.

[7] T. Virtanen, "Separation of Sound Sources by Convolutive Sparse Coding", *Workshop on Statistical and Perceptual Audio Processing SAPA*, Jeju, Korea, 2004.

[8] P. G. Georgiev, F. J. Theis, and A. Cichocki, "Blind source separation and sparse component analysis for over-complete mixtures", in *Proc. of ICASSP'04*, Montreal (Canada), 2004, pp. 493–496.

[9] R. Gribonval and S. Lesage, "A survey of sparse component analysis for blind source separation: principles, perspectives, and new challenges", in *Proc. of ESANN'06*, April 2006, pp. 323–330.

[10] M. Zibulevsky and B. A. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural Computation*, vol. 13, no. 4, pp. 863–882, 2001.

[11] P.D. O'Grady, B.A. Pearlmutter, and S.T. Rickard, "Survey of Sparse and Non-Sparse methods in Source Separation", *Int. Journal of Imaging Systems and Technology*, vol. 15, no. 1, 2005, pp. 18-33.

[12] T. Virtanen, "Monaural Sound Source Separation by Non-Negative Matrix Factorization with Temporal Continuity and Sparseness Criteria", *IEEE Trans. on Audio, Speech, and Language Processing*, accepted for publication.

[13] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal l^1-norm solution is also the sparsest solution", Tech. Rep., 2004.

[14] Y.Q. Li, A. Cichocki, and S. Amari, "Analysis of sparse representation and blind source separation," *Neural Computation*, vol. 16, no. 6, pp. 1193–1234, 2004.

[15] G. H. Mohimani, M. Babaie-Zadeh, and C. Jutten, "Fast Sparse Representation based on Smoothed L0 Norm," *Proc. of ICA 2007*, September 2007, London, UK.

[16] J.-F. Cardoso, "Blind signal separation: statistical principles," *Proceedings IEEE*, vol. 9", pp. 2009–2025, 1998.

[17] J.-F. Cardoso, B.-H. Laheld, "Equivariant Adaptive Source Separation", *IEEE Trans. on Siganl Processing*, vol. 44, no. 12, pp. 3017-3030, 1996.