Blind separation of audio sources using modal decomposition

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

BLIND SEPARATION OF AUDIO SOURCES USING MODAL DECOMPOSITION

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

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

ABSTRACT

This paper introduces new algorithms for the blind separation of audio sources using modal decomposition. Indeed, audio signals and, in particular, musical signals can be well approximated by a sum of damped sinusoidal (modal) components. Based on this representation, we propose a two steps approach consisting of a signal analysis (extraction of the modal components) followed by a signal synthesis (pairing of the components belonging to the same source) using vector clustering. For the signal analysis, two algorithms are considered and compared: namely the EMD (Empirical Mode Decomposition) algorithm and a parametric estimation algorithm using ESPRIT technique. A major advantage of the proposed method resides in its ability to separate more sources than sensors. Simulation results are given to compare and assess the performances of the proposed algorithms.

1. INTRODUCTION

The problem of blind source separation consists of finding independent source signals from their observed mixtures without a priori knowledge on the actual mixing matrix.
The source separation problem is of interest in various applications [1] such as the localization and tracking of targets using radars and sonars, separation of speakers (problem known as "cocktail party"), detection and separation in multiple access communication systems, independent components analysis of biomedical signals (EEG or ECG), multispectral astronomical images etc.
This problem has been intensively studied in the literature and many effective solutions have been proposed so far [1]. Nevertheless, the underdetermined case where the number of sources is greater than the number of sensors (observations) remains relatively poorly treated, and its resolution is one of the open problems of blind source separation. In the case of non-stationary signals (including the audio signals), certain solutions using time-frequency analysis of the observations exist for the underdetermined case [6, 7]. In this paper, we propose an alternative approach using modal decomposition of the received signals [2, 3]. More precisely we propose to decompose a supposed *locally periodic* signal which is not necessarily harmonic in the Fourier sense into its various modes. The audio signals and more particularly the musical signals can be modeled by a sum of damped sinusoids [8] and hence are well suited for our separation approach. We propose here to exploit this last property for the separation of audio sources by means of modal decomposition.

2. DATA MODEL

The blind source separation model assumes the existence of N independent signals $s_1(t), \ldots, s_N(t)$ and M observations $x_1(t), \ldots, x_M(t)$ that represent the mixtures. These mixtures are supposed linear and instantaneous, i.e.

$$x_i(t) = \sum_{j=1}^{N} a_{ij} s_j(t) \qquad i = 1, \ldots, M \tag{1}$$

This can be represented compactly by the mixing equation

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) \tag{2}$$

where $\mathbf{s}(t) \stackrel{\text{def}}{=} [s_1(t), \ldots, s_N(t)]^T$ is a $N \times 1$ column vector collecting the source signals, vector $\mathbf{x}(t)$ similarly collects the M observed signals, and the $M \times N$ mixing matrix $\mathbf{A} \stackrel{\text{def}}{=} [\mathbf{a}_1, \ldots, \mathbf{a}_N]$ with $\mathbf{a}_i = [a_{1i}, \ldots, a_{Mi}]^T$ contains the mixture coefficients. We will suppose that for any pair (i, j) with $i \neq j$, the vectors $\mathbf{a}_i$ and $\mathbf{a}_j$ are linearly independent.
The source signals are supposed to be decomposable in a sum of modal components $c_i^j(t)$, i.e:

$$s_i(t) = \sum_{j=1}^{l_i} c_i^j(t) \qquad t = 0, \ldots, T-1 \tag{3}$$

The usual source independence assumption is replaced here by a quasi-orthogonality assumption of the modal components, i.e.

$$\frac{\langle c_i^j | c_{i'}^{j'} \rangle}{\|c_i^j\| \|c_{i'}^{j'}\|} \approx 0 \quad \text{for} \quad (i,j) \neq (i',j') \tag{4}$$

where

$$\langle c_i^j | c_{i'}^{j'} \rangle \stackrel{\text{def}}{=} \sum_{t=0}^{T-1} c_i^j(t) c_{i'}^{j'}(t)^* \tag{5}$$

and

$$\|c_i^j\|^2 = \langle c_i^j | c_i^j \rangle \tag{6}$$

Remark: Assumption (4) may be restrictive in certain applications. However, it can be relaxed in such a way to allow common modal components to different sources as shown in [11].

3. SEPARATION USING MODAL DECOMPOSITION

Based on the previous model, we propose an approach in two steps consisting of:

- *An analysis step*: in this step, one applies an algorithm of modal decomposition to each sensor output in order to extract all the harmonic components from them. We compare, for this modal components extraction two decomposition algorithms that are the EMD (Empirical Mode Decomposition) algorithm introduced in [2, 3] and a parametric algorithm which estimates the parameters of the modal components modeled as damped sinusoids.
- *A synthesis step*: in this step we group together the modal components corresponding to the same source in order to reconstitute the original signal. This is done by observing that all modal components of a given source signal 'live' in the same spatial direction. Therefore, the proposed clustering method is based on the component's direction evaluated by correlation of the extracted (component) signal with the observed antenna signal.

3.1. Signal analysis using EMD

A new nonlinear technique, referred to as *Empirical Mode Decomposition* (EMD), has recently been introduced by N.E. Huang et al. for representing non-stationary signals as sum of zero-mean AM-FM components [2]. The starting point of the EMD is to consider oscillations in signals at a very local level. Given a signal $z(t)$, the EMD algorithm can be summarized as follows [3]:

1. Identify all extrema of $z(t)$.
2. Interpolate between minima (resp. maxima), ending up with some envelope $e_{min}(t)$ (resp. $e_{max}(t)$).
3. Compute the mean $m(t) = (e_{min}(t) + e_{max}(t))/2$.
4. Extract the detail $d(t) = z(t) - m(t)$.
5. Iterate on the residual $m(t)$.

By applying EMD algorithm to the i^{th} mixture signal x_i which is written as $x_i(t) = \sum_{j=1}^{N} a_{ij} s_j(t) = \sum_{j=1}^{N} \sum_{k=1}^{l_j} a_{ij} c_j^k(t)$ one obtains estimates $\widehat{c}_j^k(t)$ of components $c_j^k(t)$.

3.2. Parametric signal analysis

In this section we present an alternative solution for signal analysis. For that, we represent the source signal and hence the observations as sum of damped sinusoids:

$$x_k(t) = \Re \left\{ \sum_{l=1}^{L} \alpha_{l,k} z_l^t \right\} \tag{7}$$

where $\alpha_{l,k}$ represents the complex amplitude and $z_l = e^{d_l + i\omega_l}$ is the l^{th} pole where d_l is the negative damping factor and ω_l is the angular-frequency. $\Re(\cdot)$ represents the real part of a complex entity.
For the extraction of the modal components, we propose to use the ESPRIT-like (Estimation of Signal Parameters via Rotation Invariance Technique) technique that estimates the poles of the signals by exploiting the row-shifting invariance property of the $D \times (T-D)$ data Hankel matrix $[\mathcal{H}(x_k)]_{n_1 n_2} \stackrel{\text{def}}{=} x_k(n_1 + n_2)$, D being a window parameter chosen in the range $T/3 \leq D \leq 2T/3$.
We use of Kung's algorithm given in [5] that can be summarized in the following steps:

1. Form the data Hankel matrix $\mathcal{H}(x_k)$.
2. Estimate the $2L$-dimensional signal subspace $\mathbf{U}^{(L)} = [\mathbf{u}_1 \dots \mathbf{u}_{2L}]$ of $\mathcal{H}(x_k)$ by means of the SVD ($\mathbf{u}_1 \dots \mathbf{u}_{2L}$ are the principal left singular vectors of $\mathcal{H}(x_k)$).
3. Solve (in the least squares sense) the shift invariance equation
$$\mathbf{U}_{\downarrow}^{(L)} \Psi = \mathbf{U}_{\uparrow}^{(L)} \Leftrightarrow \Psi = \mathbf{U}_{\downarrow}^{(L)\#} \mathbf{U}_{\uparrow}^{(L)} \tag{8}$$
where $\Psi = \Phi \Delta \Phi^{-1}$, Φ being a non-singular $2L \times 2L$ matrix and $\Delta = \text{diag}(z_1, z_1^*, \dots, z_L, z_L^*)$. $()^{\#}$ denotes the pseudo-inversion operation and arrows $\downarrow$ and $\uparrow$ denote respectively the last and the first row-deleting operator.
4. Estimate the poles as the eigenvalues of matrix Ψ.
5. Estimate the complex amplitudes by solving the least squares fitting criterion
$$\min_{\boldsymbol{\alpha}} \|\mathbf{x}_k - \mathbf{Z}\boldsymbol{\alpha}\|^2 \Leftrightarrow \boldsymbol{\alpha} = \mathbf{Z}^{\#} \mathbf{x}_k \tag{9}$$
where $\mathbf{x}_k = [x_k(0) \dots x_k(T-1)]^T$ is the observation vector, $\mathbf{Z}$ is a Vandermonde matrix constructed from the estimated poles and $\boldsymbol{\alpha}$ is the vector of complex amplitudes.

3.3. Signal synthesis using vector clustering

For the synthesis of the source signals one observes that thanks to the quasi-orthogonality assumption, one has:

$$\frac{\langle \mathbf{x} | c_i^j \rangle}{\|c_i^j\|^2} \stackrel{\text{def}}{=} \frac{1}{\|c_i^j\|^2} \begin{bmatrix} \langle x_1 | c_i^j \rangle \\ \vdots \\ \langle x_M | c_i^j \rangle \end{bmatrix} \approx \mathbf{a}_i$$

where $\mathbf{a}_i$ represents the i^{th} column vector of $\mathbf{A}$. We can then associate each component $\widehat{c}_j^k$ to a space direction (vector column of $\mathbf{A}$) that is estimated by

$$\widehat{\mathbf{a}}_j^k = \frac{\langle \mathbf{x} | \widehat{c}_j^k \rangle}{\|\widehat{c}_j^k\|^2}$$

Two components of a same source signal are associated to the same column vector of $\mathbf{A}$, Therefore, we propose to gather these components by clustering the vectors $\widehat{\mathbf{a}}_j^k$ into N classes. One will be able to rebuild the initial sources up to a constant by adding the various components within a same class.

3.4. Source pairing and selection

Let us notice, that by applying the approach described previously (analysis plus synthesis) to all the antenna outputs $x_1(t), \cdots, x_M(t)$, we obtain M estimates of each source signal. The estimation quality of a given source signal varies significantly from one sensor to another. Indeed, it depends strongly on the matrix coefficients and, in particular, on the signal to interference ratio (SIR) of the desired source. Consequently, we propose a blind selection method to choose a 'good' estimate among the M we have for each source signal. For that, we need first to pair the source estimates together. This is done by associating each source signal extracted from the first sensor to the $(M-1)$ signals extracted from the $(M-1)$ other sensors that are maximally correlated with it. The correlation factor of two signals s_1 and s_2 is evaluated by $\frac{\langle s_1|s_2\rangle}{\|s_1\|\|s_2\|}$. Once, the source pairing achieved, we propose to select the source estimate of maximal energy, i.e.

$$\widehat{s}_i(t) = \max_j \{E_i^j = \sum_{t=0}^{T-1} |\widehat{s}_i^j(t)|^2, \quad j = 1, \cdots, M\} \quad (10)$$

where E_i^j represents the energy of the i^{th} source extracted from the j^{th} sensor. One can consider other methods of selection based on the dispersion around the centroid of each class, the number of components of each source estimate, etc.

3.5. Discussion

We provide here some comments to get more insight onto the proposed separation method:

- *Over-determined case*: In that case, one is able to separate the sources by left inversion of matrix $\mathbf{A}$. The latter can be estimated from the centroids of the N clustering classes (i.e., the centroid of the i^{th} class represent the estimate of the i^{th} column of $\mathbf{A}$).

- *Estimation of the number of sources*: This is a difficult and challenging task in the underdetermined case. Few approaches exist based on multi-dimensional tensor decomposition [9] or based on the clustering with joint estimation of the number of classes [4]. However, these methods are very sensitive to noise, to the source amplitude dynamic and to the conditioning of matrix $\mathbf{A}$. In this paper, we assume the number of sources known (or correctly estimated).

- *Number of modal components*: In the parametric approach, we have to choose the number of modal components L needed to well approximate the audio signal. Indeed, small values of L lead to poor signal representation while large value of L increases the computational cost. In fact, L depends on the 'signal complexity' and in general musical signals require less components (for a good modeling) than speech signals. In section 4 we illustrate the effect of the value of L on the separation quality.

- *Hybrid separation approach*: It is most probably that the separation quality can be improved using signal analysis in conjonction with spatial filtering. Indeed, it has been observed that the separation quality depends strongly on the mixture coefficients. Spatial filtering can be used to improve the SIR for a desired source signal and consequently its extraction quality. This will be the focus of a future work.

4. SIMULATION

We present here some simulation results to illustrate the performance of our blind separation algorithms. For that, we consider a uniform linear array with $M = 3$ sensors receiving the signals from $N = 4$ audio sources (except for the third experiment where N varies in the range [2,6]). The angles of arrival of the sources are chosen randomly. The sample size is set to $T = 5000$ samples (the signals are sampled at a rate of 44.1kHz). The observed signals are corrupted by an additive white noise of covariance $\sigma^2\mathbf{I}$ (σ^2 being the noise power). The separation quality is measured by the normalized mean squares estimation errors (NMSE) of the sources evaluated over 100 Monte-Carlo runs. The plots represent the averaged NMSE over the N sources. In figure 1, we compare the separation performance obtained by our algorithm using EMD and the parametric technique with $L = 30$. As a reference, we plot also the NMSE obtained by pseudo-inversion of matrix $\mathbf{A}$ [10] (assumed exactly known). It is observed that both EMD and parametric based separation provide better results than those obtained by pseudo-inversion of the exact mixing matrix. The plots in figure 2 illustrate the effect of the number of components L chosen to model the audio-signal. Too small or too large values of L degrade the performance of the method. In other words, it existe an optimal choice of L that depend on the signal type. In figure 3, we present the separation performance loss that we have when the number of sources increases from 2 to 6 in the noiseless case. For $N = 2$ and $N = 3$ (over-determined case) we estimate the sources by left inversion of the estimate of matrix $\mathbf{A}$. In the underdetermined case,

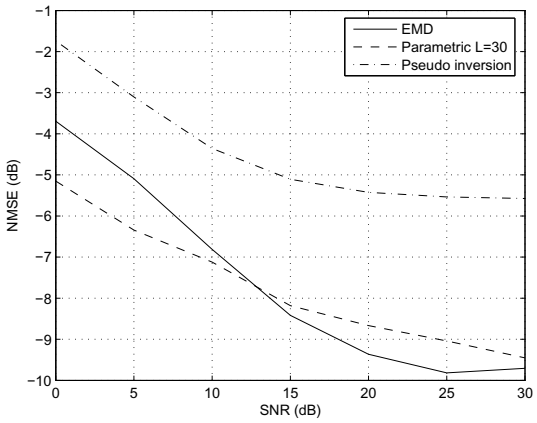

Fig. 1. NMSE versus SNR

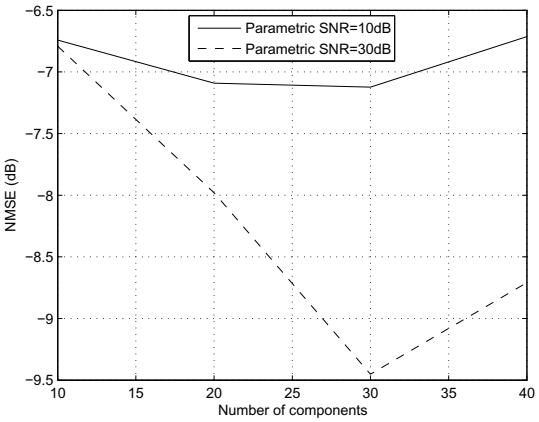

Fig. 2. NMSE versus L: SNR=10 and 30

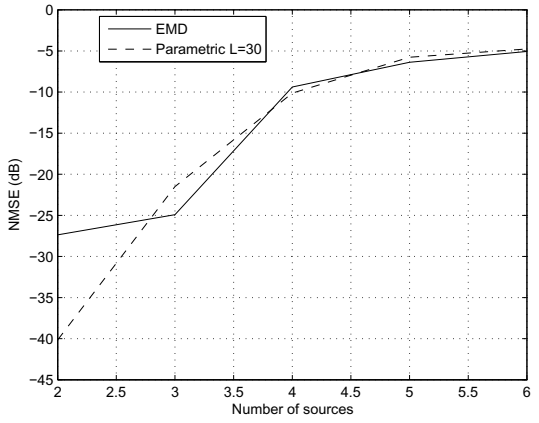

Fig. 3. NMSE versus N (noiseless case)

the EMD and parametric based algorithms present similar performances.

5. CONCLUSION

This paper introduces a new blind separation method for audio-type sources using modal decomposition. The proposed method can separate more sources than sensors and provides, in that case, a better separation quality than the one obtained by pseudo-inversion of the mixture matrix (even if it is known exactly). For the signal analysis step of the proposed method, two algorithms are used and compared using respectively the EMD and the ESPRIT-like technique for the estimation of the poles of the modal components modeled as damped sinusoids.

6. REFERENCES

[1] A.K. Nandi (editor), "Blind estimation using higher-order statistics." *Kluwer Acadimic Publishers*, Boston 1999.

[2] N.E. Huang, Z. Shen, S.R. Long, M.L. Wu, H.H. Shih, Q. Zheng, N.C. Yen, C.C. Trung, H. Liu, "The empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary times series analysis", *Proc. Roy. Soc.* London A, Vol. 454, pp. 903-995, 1998.

[3] P. Flandrin, G. Rilling and P. Goncalvs, "Empirical mode decomposition as a filter bank", *IEEE Sig. Proc. Letters*, pp. 112-114, 2004.

[4] I.E. Frank and R. Todeschini, "The data analysis handbook", *Elsevier, Sci. Pub. Co.*, 1994.

[5] S.Y. Kung, K.S. Arun and D.V. Bhaskan Rao, "Space-time and singular-value decomposition based approximation methods for the harmonic retrieval problem", *J. Opt. Soc.* Vol. 73, no. 12, 1983.

[6] L. Nguyen, A. Belouchrani, K. Abed-Meraim and B. Boashash, "Separating more sources than sensors using time-frequency distributions." in *Proc. ISSPA*, Vol. II, pp. 583-586, 2001.

[7] A. Jourjine, S. Rickard, O. Yilmaz, "Blind separation of disjoint orthogonal signals: demixing n sources from 2 mixtures," in *ICASSP*, pp. 2985-2988, June 2000.

[8] R. Boyer and K. Abed-Meraim, "Audio modeling based on delayed sinusoids." *IEEE-Tr-SAP*, Vol. 12, N2, pp. 110-120 , March 2004.

[9] L. De Lathauwer, B. Moor, J. Vandewalle, "ICA techniques for more sources than sensors", *Higher-oder statistic Proc. of the IEEE Sig. Proc. Workshop*, pp. 116-120, 1999.

[10] M.Z. Ikram,"Blind separation of delayed instantaneous mixtures: A cross-correlation based approach", *ISSPIT*, December 2002.

[11] A. Aissa-El-Bey, K. Abed-Meraim, Y. Grenier, "Séparation aveugle sous-déterminée de sources audio par la méthode EMD (Empirical Mode Decomposition)", *20e Colloque GRETSI sur le traitement du signal et des images*, September 2005.