Hindawi Publishing Corporation
EURASIP Journal on Audio, Speech, and Music Processing
Volume 2007, Article ID 85438, 15 pages
doi:10.1155/2007/85438

Research Article

Underdetermined Blind Audio Source Separation Using Modal Decomposition

Abdeldjalil Aïssa-El-Bey, Karim Abed-Meraim, and Yves Grenier

Départment TSI, École Nationale Supérieure des Télécommunications (ENST), 46 Rue Barrault, 75634 Paris Cedex 13, France

Received 1 July 2006; Revised 20 November 2006; Accepted 14 December 2006

Recommended by Patrick A. Naylor

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-step approach consisting of a signal analysis (extraction of the modal components) followed by a signal synthesis (grouping of the components belonging to the same source) using vector clustering. For the signal analysis, two existing 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 validity for both instantaneous and convolutive mixtures and its ability to separate more sources than sensors. Simulation results are given to compare and assess the performance of the proposed algorithms.

1. INTRODUCTION

The problem of blind source separation (BSS) consists of finding "independent" source signals from their observed mixtures without a priori knowledge on the actual mixing channels.

The source separation problem is of interest in various applications [1, 2] 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 component analysis of biomedical signals (EEG or ECG), multispectral astronomical imaging, geophysical data processing, and so forth [2].

This problem has been intensively studied in the literature and many effective solutions have been proposed so far [1–3]. Nevertheless, the literature intended for the underdetermined case where the number of sources is larger than the number of sensors (observations) is relatively limited, and achieving the BSS in that context is one of the challenging problems in this field. Existing methods for the underdetermined BSS (UBSS) include the matching pursuit methods in [4, 5], the separation methods for finite alphabet sources in [6, 7], the probabilistic-based (using maximum a posteriori criterion) methods in [8–10], and the sparsity-based techniques in [11, 12]. In the case of nonstationary signals (including the audio signals), certain solutions using time-frequency analysis of the observations exist for the underdetermined case [13–15]. In this paper, we propose an alternative approach named MD-UBSS (for modal decomposition UBSS) using modal decomposition of the received signals [16, 17]. 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 [18, 19], 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. Although we consider here an audio application, the proposed method can be used for any other application where the source signals can be represented by a sum of sinusoidal components. This includes in particular the separation of NMR (nuclear magnetic resonance) signals in [20, 21] and the rotating machine signals in [22]. To start, we consider first the case of instantaneous mixtures, then we treat the more challenging problem of convolutive mixtures in the underdetermined case.

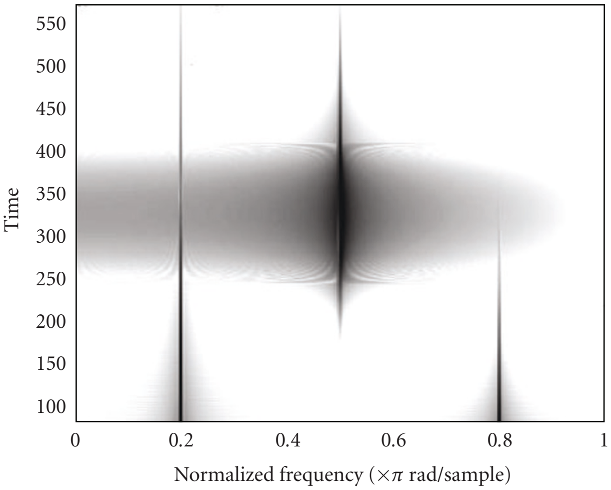

FIGURE 1: Time-frequency representation of a three-modal-component signal (using short-time Fourier transform).

Note that this modal representation of the sources is a particular case of signal sparsity often used to separate the sources in the underdetermined case [23]. Indeed, a signal given by a sum of sinusoids (or damped sinusoids) occupies only a small region in the time-frequency (TF) domain, that is, its TF representation is sparse. This is illustrated by Figure 1 where we represent the time-frequency distribution of a three-modal-component signal.

The paper is organized as follows. Section 2 formulates the UBSS problem and introduces the assumptions necessary for the separation of audio sources using modal decomposition. Section 3 proposes two MD-UBSS algorithms for instantaneous mixture case while Section 4 introduces a modified version of MD-UBSS that relaxes the quasiorthogonality assumption of the source modal components. In Section 5, we extend our MD-UBSS algorithm to the convolutive mixture case. Some discussions on the proposed methods are given in Section 6. The performance of the above methods is numerically evaluated in Section 7. The last section is for the conclusion and final remarks.

2. PROBLEM FORMULATION IN THE INSTANTANEOUS MIXTURE CASE

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

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

This can be represented compactly by the mixing equation

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

where $\mathbf{s}(t) \overset{\text{def}}{=} [s_1(t), \dots, s_N(t)]^T$ is an $N \times 1$ column vector collecting the real-valued source signals, vector $\mathbf{x}(t)$, similarly, collects the M observed signals, and the $M \times N$ mixing matrix $\mathbf{A} \overset{\text{def}}{=} [\mathbf{a}_1, \dots, \mathbf{a}_N]$ with $\mathbf{a}_i = [a_{1i}, \dots, a_{Mi}]^T$ contains the mixture coefficients.

Now, if $N > M$, that is, there are more sources than sensors, we are in the underdetermined case, and BSS becomes UBSS (U stands for underdetermined). By underdeterminacy, we cannot, from the set of equations in (2), algebraically obtain a unique solution, because this system contains more variables (sources) than equations (sensors). In this case, $\mathbf{A}$ is no longer left invertible, because it has more columns than rows. Consequently, due to the underdetermined representation, the above system of (2) cannot be solved completely even with the full knowledge of $\mathbf{A}$, unless we have some specific knowledge about the underlying sources.

Next, we will make some assumptions about the data model in (2), necessary for our method to achieve the UBSS.

Assumption 1. The column vectors of $\mathbf{A}$ are pairwise linearly independent.

That is, for any index pair $i \neq j \in \mathcal{N}$, where $\mathcal{N} = \{1, \dots, N\}$, vectors $\mathbf{a}_i$ and $\mathbf{a}_j$ are linearly independent. This assumption is necessary because if otherwise, we have $\mathbf{a}_2 = \alpha \mathbf{a}_1$ for example, then the input/output relation (2) can be reduced to

$$\mathbf{x}(t) = [\mathbf{a}_1, \mathbf{a}_3, \dots, \mathbf{a}_N][s_1(t) + \alpha s_2(t), s_3(t), \dots, s_N(t)]^T, \tag{3}$$

and hence the separation of $s_1(t)$ and $s_2(t)$ is inherently impossible. This assumption is used later (in the clustering step) to separate the source modal components using their spatial directions given by the column vectors of $\mathbf{A}$.

It is known that BSS is only possible up to some scaling and permutation [3]. We take the advantage of these indeterminacies to further make the following assumption without loss of generality.

Assumption 2. The column vectors of $\mathbf{A}$ are of unit norm.

That is, $\|\mathbf{a}_i\| = 1$ for all $i \in \mathcal{N}$, where the norm hereafter is given in the Frobenius sense.

As mentioned previously, solving the UBSS problem requires strong a priori assumptions on the source signals. In our case, signal sparsity is considered in terms of modal representation of the input signals as stated by the fundamental assumption below.

Assumption 3. The source signals are sum of modal components.

Indeed, we assume here that each source signal $s_i(t)$ is a sum of l_i modal components $c_i^j(t)$, $j = 1, \dots, l_i$, that is,

$$s_i(t) = \sum_{j=1}^{l_i} c_i^j(t), \quad t = 0, \dots, T-1, \tag{4}$$

where $c_i^j(t)$ are damped sinusoids or (quasi)harmonic signals, and T is the sample size.

Standard BSS techniques are based on the source independence assumption. In the UBSS case, the source independence is often replaced by the disjointness of the sources. This means that there exists a transform domain where the source representation has disjoint or quasidisjoint supports. The quasidisjointness assumption of the sources translates in our case into the quasiorthogonality of the modal components.

Assumption 4. The sources are quasiorthogonal, in the sense that

$$\frac{\langle c_i^j \mid c_{i'}^{j'} \rangle}{\|c_i^j\|\,\|c_{i'}^{j'}\|} \approx 0, \quad \text{for } (i,j) \neq (i',j'), \tag{5}$$

where

$$\begin{gathered}\langle c_i^j \mid c_{i'}^{j'} \rangle \overset{\text{def}}{=} \sum_{t=0}^{T-1} c_i^j(t) c_{i'}^{j'}(t), \\ \|c_i^j\|^2 = \langle c_i^j \mid c_i^j \rangle.\end{gathered} \tag{6}$$

In the case of sinusoidal signals, the quasiorthogonality of the modal components is nothing else than the Fourier quasiorthogonality of two sinusoidal components with distinct frequencies. This can be observed in the frequency domain through the disjointness of their supports. This property is also preserved by filtering, which does not affect the frequency support, and hence the quasiorthogonality assumption of the signals (this is used later when considering the convolutive case).

3. MD-UBSS ALGORITHM

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

(i) 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 [16, 17] and a parametric algorithm which estimates the parameters of the modal components modeled as damped sinusoids.

(ii) 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 spatial direction evaluated by correlation of the extracted (component) signal with the observed antenna signal.

(1) Extraction of all harmonic components from each sensor by applying modal decomposition.
(2) Spatial direction estimation by (14) and vector clustering by k-means algorithm [24].
(3) Source estimation by grouping together the modal components corresponding to the same spatial direction.
(4) Source grouping and source selection by (18).

Algorithm 1: MD-UBSS algorithm in instantaneous mixture case using modal decomposition.

Note that, by this method, each sensor output leads to an estimate of the source signals. Therefore, we end up with M estimates for each source signal. As the quality of source signal extraction depends strongly on the mixture coefficients, we propose a blind source selection procedure to choose the "best" of the M estimates. This algorithm is summarized in Algorithm 1.

3.1. Modal component estimation

3.1.1. Signal analysis using EMD

A new nonlinear technique, referred to as *empirical mode decomposition* (EMD), has recently been introduced by Huang et al. for representing nonstationary signals as sum of zero-mean AM-FM components [16]. 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 [17]:

(1) identify all extrema of $z(t)$. This is done by the algorithm in [25];
(2) interpolate between minima (resp., maxima), ending up with some envelope $e_{\min}(t)$ (resp., $e_{\max}(t)$). Several interpolation techniques can be used. In our simulation, we have used a spline interpolation as in [25];
(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[1] $m(t)$ until $m(t) = 0$ (in practice, we stop the algorithm when $\|m(t)\| \leq \epsilon$, where ϵ is a given threshold value).

By applying the EMD algorithm to the ith 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 $\hat{c}_j^k(t)$ of components $c_j^k(t)$ (up to the scalar constant a_{ij}).

3.1.2. Parametric signal analysis

In this section, we present an alternative solution for signal analysis. For that, we represent the source signal as sum of

[1] Indeed, the mean signal $m(t)$ is also the residual signal after extracting the detail component $d(t)$, that is, $m(t) = z(t) - d(t)$.

damped sinusoids:

$$s_i(t) = \Re e\left\{ \sum_{j=1}^{l_i} \alpha_i^j (z_i^j)^t \right\}, \tag{7}$$

corresponding to

$$c_i^j(t) = \Re e\left\{ \alpha_i^j (z_i^j)^t \right\}, \tag{8}$$

where $\alpha_i^j = \beta_i^j e^{\theta_i^j}$ represents the complex amplitude and $z_i^j = e^{d_i^j + \jmath \omega_i^j}$ is the jth pole of the source s_i, where d_i^j is the negative damping factor and ω_i^j is the angular frequency. $\Re e(\cdot)$ represents the real part of a complex entity. We denote by L_{tot} the total number of modal components, that is, $L_{\text{tot}} = \sum_{i=1}^{N} l_i$.

For the extraction of the modal components, we propose to use the ESPRIT (estimation of signal parameters via rotational invariance technique) algorithm that estimates the poles of the signals by exploiting the row-shifting invariance property of the $D \times (T - D)$ data Hankel matrix $[\mathcal{H}(x_k)]_{n_1 n_2} \overset{\text{def}}{=} x_k(n_1+n_2)$, D being a window parameter chosen in the range $T/3 \leq D \leq 2T/3$.

More precisely, we use Kung's algorithm given in [26] that can be summarized in the following steps:

(1) form the data Hankel matrix $\mathcal{H}(x_k)$;

(2) estimate the $2L_{\text{tot}}$-dimensional signal subspace $\mathbf{U}^{(L_{\text{tot}})} = [\mathbf{u}_1, \ldots, \mathbf{u}_{2L_{\text{tot}}}]$ of $\mathcal{H}(x_k)$ by means of the SVD of $\mathcal{H}(x_k)$ ($\mathbf{u}_1, \ldots, \mathbf{u}_{2L_{\text{tot}}}$ are the principal left singular eigenvectors of $\mathcal{H}(x_k)$);

(3) solve (in the least-squares sense) the shift invariance equation

$$\mathbf{U}_{\downarrow}^{(L_{\text{tot}})} \Psi = \mathbf{U}_{\uparrow}^{(L_{\text{tot}})} \Longleftrightarrow \Psi = \mathbf{U}_{\downarrow}^{(L_{\text{tot}})\#} \mathbf{U}_{\uparrow}^{(L_{\text{tot}})}, \tag{9}$$

where $\Psi = \Phi \Delta \Phi^{-1}$, Φ being a nonsingular $2L_{\text{tot}} \times 2L_{\text{tot}}$ matrix, and $\Delta = \operatorname{diag}(z_1^1, z_1^{1*}, \ldots, z_1^{l_1}, z_1^{l_1*}, \ldots, z_N^{l_N}, z_N^{l_N*})$. $(\cdot)^*$ represents the complex conjugation, $(\cdot)^\#$ denotes the pseudoinversion operation, and arrows ↓ and ↑ 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

$$\min_{\boldsymbol{\alpha}_k} \|\mathbf{x}_k - \mathbf{Z}\boldsymbol{\alpha}_k\|^2 \Longleftrightarrow \boldsymbol{\alpha}_k = \mathbf{Z}^\# \mathbf{x}_k, \tag{10}$$

where $\mathbf{x}_k = [x_k(0), \ldots, x_k(T-1)]^T$ is the observation vector, $\mathbf{Z}$ is a Vandermonde matrix constructed from the estimated poles, that is,

$$\mathbf{Z} = [\mathbf{z}_1^1, \mathbf{z}_1^{1*}, \ldots, \mathbf{z}_1^{l_1}, \mathbf{z}_1^{l_1*}, \ldots, \mathbf{z}_N^{l_N}, \mathbf{z}_N^{l_N*}], \tag{11}$$

with $\mathbf{z}_i^j = [1, z_i^j, (z_i^j)^2, \ldots, (z_i^j)^{T-1}]^T$, and $\boldsymbol{\alpha}_k$ is the vector of complex amplitudes, that is,

$$\boldsymbol{\alpha}_k = \frac{1}{2}[a_{k1}\alpha_1^1, a_{k1}\alpha_1^{1*}, \ldots, a_{k1}\alpha_1^{l_1*}, \ldots, a_{kN}\alpha_N^{l_N*}]^T. \tag{12}$$

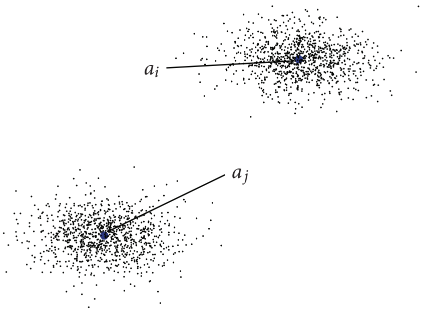

Figure 2: Data clustering illustration, where we represent the different estimates $\hat{\mathbf{a}}_i^j$ and their centroids.

3.2. Clustering and source estimation

3.2.1. Signal synthesis using vector clustering

For the synthesis of the source signals, one observes that thanks to the quasiorthogonality assumption, one has

$$\frac{\langle \mathbf{x} \mid c_i^j \rangle}{\|c_i^j\|^2} \overset{\text{def}}{=} \frac{1}{\|c_i^j\|^2} \begin{bmatrix} \langle x_1 \mid c_i^j \rangle \\ \vdots \\ \langle x_M \mid c_i^j \rangle \end{bmatrix} \approx \mathbf{a}_i, \tag{13}$$

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

$$\hat{\mathbf{a}}_j^k = \frac{\langle \mathbf{x} \mid \hat{c}_j^k \rangle}{\|\hat{c}_j^k\|^2}. \tag{14}$$

Vector $\hat{\mathbf{a}}_j^k$ would be equal approximately to $\mathbf{a}_i$ (up to a scalar constant) if $\hat{c}_j^k$ is an estimate of a modal component of source i. Hence, two components of a same source signal are associated to colinear spatial direction of to the same column vector of $\mathbf{A}$. Therefore, we propose to gather these components by clustering their directional vectors into N classes (see Figure 2). For that, we compute first the normalized vectors

$$\overline{\mathbf{a}}_j^k = \frac{\hat{\mathbf{a}}_j^k e^{-\jmath \psi_j^k}}{\|\hat{\mathbf{a}}_j^k\|}, \tag{15}$$

where ψ_j^k is the phase argument of the first entry of $\hat{\mathbf{a}}_j^k$ (this is to force the first entry to be real positive). Then, these vectors are clustered by k-means algorithm [24] that can be summarized in the following steps.

(1) Place N points into the space represented by the vectors that are being clustered. These points represent initial group centroids. One popular way to start is to randomly choose N vectors among the set of vectors to be clustered.

(2) Assign each vector $\overline{\mathbf{a}}_j^k$ to the group (cluster) that has the closest centroid, that is, if $\mathbf{y}_1, \ldots, \mathbf{y}_N$ are the centroids

of the N clusters, one assigns the vector $\overline{\mathbf{a}}_j^k$ to the cluster i_0 that satisfies

$$i_0 = \arg\min_i ||\overline{\mathbf{a}}_j^k - \mathbf{y}_i||. \tag{16}$$

(3) When all vectors have been assigned, recalculate the positions of the N centroids in the following way: for each cluster, the new centroid's vector is calculated as the mean value of the cluster's vectors.

(4) Repeat steps 2 and 3 until the centroids no longer move. This produces a separation of the vectors into N groups. In practice, in order to increase the convergence rate, one can also use a threshold value and stop the algorithm when the difference between the new and old centroid values is smaller than this threshold for all N clusters.

Finally, one will be able to rebuild the initial sources up to a constant by adding the various components within a same class, that is,

$$\hat{s}_i(t) = \sum_{\mathcal{C}_i} \hat{c}_i^j(t), \tag{17}$$

where $\mathcal{C}_i$ represents the ith cluster.

3.2.2. *Source grouping and selection*

Let us notice that by applying the approach described previously (analysis plus synthesis) to all antenna outputs $x_1(t), \dots, x_M(t)$, 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 $|\langle s_1 \mid s_2\rangle|/\|s_1\|\|s_2\|$.

Once the source grouping is achieved, we propose to select the source estimate of maximal energy, that is,

$$\hat{s}_i(t) = \arg\max_{\hat{s}_i^j(t)} \left\{ E_i^j = \sum_{t=0}^{T-1} \left|\hat{s}_i^j(t)\right|^2,\ j = 1, \dots, M \right\}, \tag{18}$$

where E_i^j represents the energy of the ith source extracted from the jth sensor $\hat{s}_i^j(t)$. One can consider other methods of selection (based, e.g., on the dispersion around the centroid) or instead, a diversity combining technique for the different source estimates. However, the source estimates are very dissimilarly in quality, and hence we have observed in our simulations that the energy-based selection, even though not optimal, provides the best results in terms of source estimation error.

3.3. *Case of common modal components*

We consider here the case where a given component $c_j^k(t)$ associated with the pole z_j^k can be shared by several sources. This is the case, for example, for certain musical signals such as those treated in [27]. To simplify, we suppose that a component belongs to at most two sources. Thus, let us suppose that the sinusoidal component $(z_j^k)^t$ is present in the sources $s_{j_1}(t)$ and $s_{j_2}(t)$ with the amplitudes α_{j_1} and α_{j_2}, respectively (i.e., one modal component of source s_{j_1} (resp., s_{j_2}) is $\Re e(\alpha_{j_1}(z_j^k)^t)$ (resp., $\Re e(\alpha_{j_2}(z_j^k)^t)$)). It follows that the spatial direction associated with this component is a linear combination of the column vectors $\mathbf{a}_{j_1}$ and $\mathbf{a}_{j_2}$. More precisely, we have

$$\hat{\mathbf{a}}_j^k = \frac{1}{||\mathbf{z}_j^k||^2}\begin{bmatrix}\mathbf{x}_1^T\mathbf{z}_j^k \\ \vdots \\ \mathbf{x}_M^T\mathbf{z}_j^k\end{bmatrix} \approx \alpha_{j_1}\mathbf{a}_{j_1} + \alpha_{j_2}\mathbf{a}_{j_2}. \tag{19}$$

It is now a question of finding the indices j_1 and j_2 of the two sources associated with this component, as well as the amplitudes α_{j_1} and α_{j_2}. With this intention, one proposes an approach based on subspace projection. Let us assume that $M > 2$ and that matrix $\mathbf{A}$ is known and satisfies the condition that any triplet of its column vectors is linearly independent. Consequently, we have

$$\mathbf{P}_{\widetilde{\mathbf{A}}}^{\perp}\hat{\mathbf{a}}_j^k = 0, \tag{20}$$

if and only if $\widetilde{\mathbf{A}} = [\mathbf{a}_{j_1}\ \ \mathbf{a}_{j_2}]$, $\widetilde{\mathbf{A}}$ being a matrix formed by a pair of column vectors of $\mathbf{A}$ and $\mathbf{P}_{\widetilde{\mathbf{A}}}^{\perp}$ represents the matrix of orthogonal projection on the orthogonal range space of $\widetilde{\mathbf{A}}$, that is,

$$\mathbf{P}_{\widetilde{\mathbf{A}}}^{\perp} = \mathbf{I} - \widetilde{\mathbf{A}}\left(\widetilde{\mathbf{A}}^H\widetilde{\mathbf{A}}\right)^{-1}\widetilde{\mathbf{A}}^H, \tag{21}$$

where $\mathbf{I}$ is the identity matrix and $(\cdot)^H$ denotes the transpose conjugate. In practice, by taking into account the noise, one detects the columns j_1 and j_2 by minimizing

$$(j_1, j_2) = \arg\min_{(l,m)} \left\{ ||\mathbf{P}_{\widetilde{\mathbf{A}}}^{\perp}\hat{\mathbf{a}}_j^k|| \mid \widetilde{\mathbf{A}} = \left[\mathbf{a}_l\ \ \mathbf{a}_m\right] \right\}. \tag{22}$$

Once $\widetilde{\mathbf{A}}$ found, one estimates the weightings α_{j_1} and α_{j_2} by

$$\begin{bmatrix}\alpha_{j_1} \\ \alpha_{j_2}\end{bmatrix} = \widetilde{\mathbf{A}}^{\#}\hat{\mathbf{a}}_j^k. \tag{23}$$

In this paper, we treated all the components as being associated to two source signals. If ever a component is present only in one source, one of the two coefficients estimated in (23) should be zero or close to zero.

In what precedes, the mixing matrix $\mathbf{A}$ is supposed to be known. This means that it has to be estimated before applying a subspace projection. This is performed here by clustering all the spatial direction vectors in (14) as for the previous MD-UBSS algorithm. Then, the ith column vector of $\mathbf{A}$ is estimated as the centroid of $\mathcal{C}_i$ assuming implicitly that most modal components belong mainly to one source signal. This is confirmed by our simulation experiment shown in Figure 11.

4. MODIFIED MD-UBSS ALGORITHM

We propose here to improve the previous algorithm with respect to the computational cost and the estimation accuracy when Assumption 4 is poorly satisfied.[2] First, in order to avoid repeated estimation of modal components for each sensor output, we use all the observed data to estimate (only once) the poles of the source signals. Hence, we apply the ESPRIT technique on the averaged data covariance matrix $\mathbb{H}(\mathbf{x})$ define by

$$\mathbb{H}(\mathbf{x}) = \sum_{i=1}^{M} \mathcal{H}(x_i)\mathcal{H}(x_i)^H \tag{24}$$

and we apply steps 1 to 4 of Kung's algorithm described in Section 3.1.2 to obtain all the poles z_i^j, $i = 1,\dots,N$, $j = 1,\dots,l_i$. In this way, we reduce significantly the computational cost and avoid the problem of "best source estimate" selection of the previous algorithm.

Now, to relax Assumption 4, we can rewrite the data model as

$$\mathbf{\Gamma}\mathbf{z}(t) = \mathbf{x}(t), \tag{25}$$

where $\mathbf{\Gamma} \overset{\text{def}}{=} [\boldsymbol{\gamma}_1^1, \overline{\boldsymbol{\gamma}}_1^1, \dots, \boldsymbol{\gamma}_N^{l_N}, \overline{\boldsymbol{\gamma}}_N^{l_N}]$, $\boldsymbol{\gamma}_i^j = \beta_i^j e^{\jmath\phi_i^j}\mathbf{b}_i^j$ and $\overline{\boldsymbol{\gamma}}_i^j = \beta_i^j e^{-\jmath\phi_i^j}\mathbf{b}_i^j$, where $\mathbf{b}_i^j$ is a unit norm vector representing the spatial direction of the ith component (i.e., $\mathbf{b}_i^j = \mathbf{a}_k$ if the component $(z_i^j)^t$ belongs to the kth source signal) and $\mathbf{z}(t) \overset{\text{def}}{=} [(z_1^1)^t, (z_1^{1*})^t, \dots, (z_N^{l_N})^t, (z_N^{l_N*})^t]^T$.

The estimation of $\mathbf{\Gamma}$ using the least-squares fitting criterion leads to

$$\min_{\mathbf{\Gamma}} \|\mathbf{X} - \mathbf{\Gamma}\mathcal{Z}\|^2 \iff \mathbf{\Gamma} = \mathbf{X}\mathcal{Z}^{\#}, \tag{26}$$

where $\mathbf{X} = [\mathbf{x}(0), \dots, \mathbf{x}(T-1)]$ and $\mathcal{Z} = [\mathbf{z}(0), \dots, \mathbf{z}(T-1)]$. After estimating $\mathbf{\Gamma}$, we estimate the phase of each pole as

$$\phi_i^j = \frac{\arg(\overline{\boldsymbol{\gamma}}_i^{jH}\boldsymbol{\gamma}_i^j)}{2}. \tag{27}$$

The spatial direction of each modal component is estimated by

$$\widehat{\mathbf{a}}_i^j = \boldsymbol{\gamma}_i^j e^{-\jmath\phi_i^j} + \overline{\boldsymbol{\gamma}}_i^j e^{\jmath\phi_i^j} = 2\beta_i^j\mathbf{b}_i^j. \tag{28}$$

Finally, we group together these components by clustering the vectors $\widehat{\mathbf{a}}_i^j$ into N classes. After clustering, we obtain N classes with N unit-norm centroids $\widehat{\mathbf{a}}_1, \dots, \widehat{\mathbf{a}}_N$ corresponding to the estimates of the column vectors of the mixing matrix $\mathbf{A}$. If the pole z_i^j belongs to the kth class, then according to (28), its amplitude can be estimated by

$$\beta_i^j = \frac{\widehat{\mathbf{a}}_k^T\widehat{\mathbf{a}}_i^j}{2}. \tag{29}$$

One will be able to rebuild the initial sources up to a constant by adding the various modal components within a same class $\mathcal{C}_k$ as follows:

$$\widehat{s}_k(t) = \Re e\left\{\sum_{\mathcal{C}_k} \beta_i^j e^{\jmath\phi_i^j}(z_i^j)^t\right\}. \tag{30}$$

Note that one can also assign each component to two (or more) source signals as in Section 3.3 by using (20)–(23).

5. GENERALIZATION TO THE CONVOLUTIVE CASE

The instantaneous mixture model is, unfortunately, not valid in real-life applications where multipath propagation with large channel delay spread occurs, in which case convolutive mixtures are considered.

Blind separation of convolutive mixtures and multichannel deconvolution has received wide attention in various fields such as biomedical signal analysis and processing (EEG, MEG, ECG), speech enhancement, geophysical data processing, and data mining [2].

In particular, acoustic applications are considered in situations where signals, from several microphones in a sound field produced by several speakers (the so-called cocktail-party problem) or from several acoustic transducers in an underwater sound field produced by engine noises of several ships (sonar problem), need to be processed.

In this case, the signal can be modeled by the following equation:

$$\mathbf{x}(t) = \sum_{k=0}^{K} \mathbf{H}(k)\mathbf{s}(t-k) + \mathbf{w}(t), \tag{31}$$

where $\mathbf{H}(k)$ are $M \times N$ matrices for $k \in [0, K]$ representing the impulse response coefficients of the channel. We consider in this paper the underdetermined case ($M < N$). The sources are assumed, as in the instantaneous mixture case, to be decomposable in a sum of damped sinusoids satisfying approximately the quasiorthogonality Assumption 4. The channel satisfies the following diversity assumption.

Assumption 5. The channel is such that each column vector of

$$\mathbf{H}(z) \overset{\text{def}}{=} \sum_{k=0}^{K} \mathbf{H}(k)z^{-k} \overset{\text{def}}{=} [\mathbf{h}_1(z), \dots, \mathbf{h}_N(z)] \tag{32}$$

is irreducible, that is, the entries of $\mathbf{h}_i(z)$ denoted by $h_{ij}(z)$, $j = 1, \dots, M$, have no common zero for all i. Moreover, any two column vectors of $\mathbf{H}(z)$ form an irreducible polynomial matrix $\widetilde{\mathbf{H}}(z)$, that is, rank $(\widetilde{\mathbf{H}}(z)) = 2$ for all z.

Knowing that the convolution preserves the different modes of the signal, we can exploit this property to estimate the different modal components of the source signals using the ESPRIT method considered previously in the instantaneous mixture case. However, using the quasiorthogonality assumption, the correlation of a given modal component

[2] This is the case when the modal components are closely spaced or for modal components with strong damping factors.

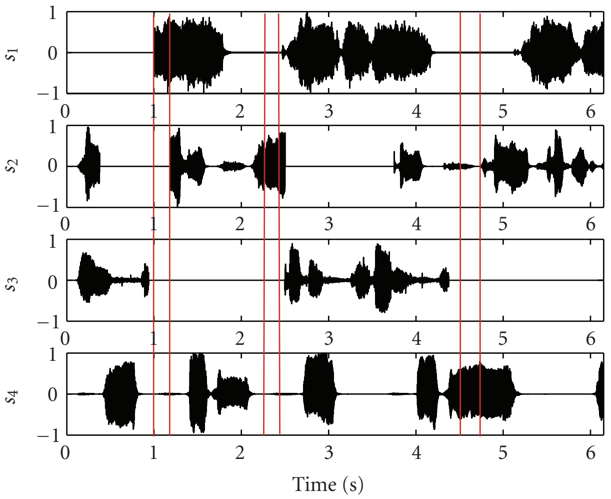

Figure 3: Time representation of 4 audio sources: this representation illustrates the audio signal sparsity (i.e., there exist time intervals where only one source is present).

corresponding to a pole z_i^j of source s_i with the observed signal $\mathbf{x}(t)$ leads to an estimate of vector $\mathbf{h}_i(z_i^j)$. Therefore, two components of respective poles z_i^j and z_i^k of the same source signal s_i will produce spatial directions $\mathbf{h}_i(z_i^j)$ and $\mathbf{h}_i(z_i^k)$ that are not colinear. Consequently, the clustering method used for the instantaneous mixture case cannot be applied in this context of convolutive mixtures.

In order to solve this problem, it is necessary to identify first the impulse response of the channels. This problem in overdetermined case is very difficult and becomes almost impossible in the underdetermined case without side information on the considered sources. In this work and similar to [28], we exploit the sparseness property of the audio sources by assuming that from time to time, only one source is present. In other words, we consider the following assumption.

Assumption 6. There exist, periodically, time intervals where only one source is present in the mixture. This occurs for all source signals of the considered mixtures (see Figure 3).

To detect these time intervals, we propose to use information criterion tests for the estimation of the number of sources present in the signal (see Section 5.1 for more details). An alternative solution would be to use the "frame selection" technique in [29] that exploits the structure of the spectral density function of the observations. The algorithm in convolutive mixture case is summarized in Algorithm 2.

5.1. Channel estimation

Based on Assumption 6, we propose here to apply SIMO-(single-input-multiple-output-) based techniques to blindly estimate the channel impulse response. Regarding the problem at hand, we have to solve 3 different problems: first, we have to select time intervals where only one source signal is effectively present; then, for each selected time interval one should apply an appropriate blind SIMO identification technique to estimate the channel parameters; finally, in the way we proceed, the same channel may be estimated several times and hence one has to group together (cluster) the channel estimates into N classes corresponding to the N source channels.

(1) Channel estimation; AIC criterion [30] to detect the number of sources and application of blind identification algorithm [31, 32] to estimate the channel impulse response.
(2) Extraction of all harmonic components from each sensor by applying parametric estimation algorithm (ESPRIT technique).
(3) Spatial direction estimation by (44).
(4) Source estimation by grouping together, using (45), the modal components corresponding to the same source (channel).
(5) Source grouping and source selection by (18).

Algorithm 2: MD-UBSS algorithm in convolutive mixture case using modal decomposition.

5.1.1. Source number estimation

Let define the spatiotemporal vector

$$\mathbf{x}_d(t) = \left[\mathbf{x}^T(t),\dots,\mathbf{x}^T(t-d+1)\right]^T = \sum_{k=1}^{N} \mathbf{H}_k\mathbf{s}_k(t) + \mathbf{w}_d(t), \tag{33}$$

where $\mathbf{H}_k$ are block-Sylvester matrices of size $dM \times (d+K)$ and $\mathbf{s}_k(t) \overset{\text{def}}{=} [s_k(t),\dots,s_k(t-K-d+1)]^T$. d is a chosen processing window size. Under the no-common zeros assumption and for large window sizes (see [30] for more details), matrices $\mathbf{H}_k$ are full column rank.

Hence, in the noiseless case, the rank of the data covariance matrix $\mathbf{R} \overset{\text{def}}{=} E[\mathbf{x}_d(t)\mathbf{x}_d^H(t)]$ is equal to $\min(p(d+K), dM)$, where p is the number of sources present in the considered time interval over which the covariance matrix is estimated. In particular, for $p = 1$, one has the minimum rank value equal to $(d+K)$.

Therefore, our approach consists in estimating the rank of the sample averaged covariance matrix $\mathbf{R}$ over several time slots (intervals) and selecting those corresponding to the smallest rank value $r = d+K$.

In the case where p sources are active (present) in the considered time slot, the rank would be $r = p(d+K)$, and hence p can be estimated by the closest integer value to $r/(d+K)$.

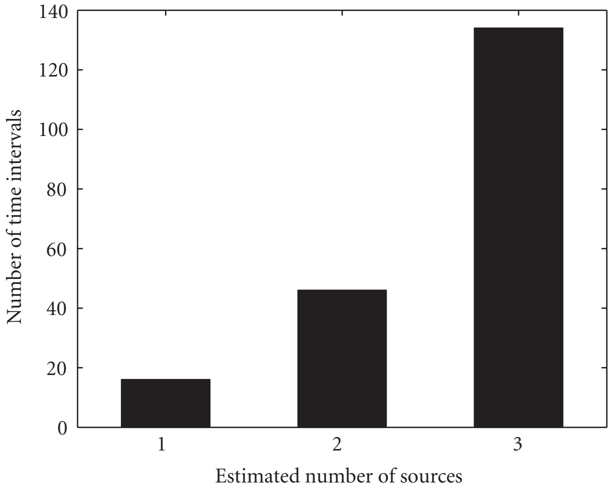

Figure 4: Histogram representing the number of time intervals for each estimated number of sources for 4 audio sources and 3 sensors in convolutive mixture case.

The estimation of the rank value is done here by Akaike's information criterion (AIC) [30] according to

$$r = \arg\min_{k}\left[-2\log\left(\frac{\prod_{i=k+1}^{Md}\lambda_i^{1/(Md-k)}}{(1/(Md-k))\sum_{i=k+1}^{Md}\lambda_i}\right)^{(Md-k)T_s} + 2k(2Md-k)\right], \tag{34}$$

where $\lambda_1 \geq \cdots \geq \lambda_{Md}$ represent the eigenvalues of $\mathbf{R}$ and T_s is the time slot size. Note that it is not necessary at this stage to know exactly the channel degree K as long as $d > K$ (i.e., an overestimation of the channel degree is sufficient) in which case the presence of one source signal is characterized by

$$d < r < 2d. \tag{35}$$

Figure 4 illustrates the effectiveness of the proposed method where a recording of 6 seconds of $M = 3$ convolutive mixtures of $N = 4$ sources is considered. The sampling frequency is 8 KHz and the time slot size is $T_s = 200$ samples. The filter coefficients are chosen randomly and the channel order is $K = 6$. One can observe that the case $p = 1$ (one source signal) occurs approximatively 10% of the time in the considered context.

5.1.2. Blind channel identification

To perform the blind channel identification, we have used in this paper the cross-relation (CR) technique described in [31, 32]. Consider a time interval where we have only the source s_i present. In this case, we can consider a SIMO system of M outputs given by

$$\mathbf{x}(t) = \sum_{k=0}^{K}\mathbf{h}_i(k)s_i(t-k) + \mathbf{w}(t), \tag{36}$$

where $\mathbf{h}_i(k) = [h_{i1}(k)\cdots h_{iM}(k)]^T$, $k = 0,\ldots,K$. From (36), the noise-free outputs $x_j(k)$, $1 \leq j \leq M$, are given by

$$x_j(k) = h_{ij}(k) * s_i(k), \quad 1 \leq j \leq M, \tag{37}$$

where "$*$" denotes the convolution. Using commutativity of convolution, it follows that

$$h_{il}(k) * x_j(k) = h_{ij}(k) * x_l(k), \quad 1 \leq j < l \leq M. \tag{38}$$

This is a linear equation satisfied by every pair of channels. It was shown that reciprocally the previous $M(M-1)/2$ cross-relations characterize uniquely the channel parameters. We have the following theorem [31].

Theorem 1. *Under the no-common zeros assumption, the set of cross-relations (in the noise free case):*

$$x_l(k) * h_j'(k) - x_j(k) * h_l'(k) = 0, \quad 1 \leq l < j \leq M, \tag{39}$$

where $\mathbf{h}'(z) = [h_1'(z)\cdots h_M'(z)]^T$ is an $M\times 1$ polynomial vector of degree K, is satisfied if and only if $\mathbf{h}'(z) = \alpha\mathbf{h}_i(z)$ for a given scalar constant α.

By collecting all possible pairs of M channels, one can easily establish a set of linear equations. In matrix form, this set of equations can be expressed as

$$\mathcal{X}_M\mathbf{h}_i = \mathbf{0}, \tag{40}$$

where $\mathbf{h}_i \overset{\text{def}}{=} [h_{i1}(0)\cdots h_{i1}(K),\ldots,h_{iM}(0)\cdots h_{iM}(K)]^T$ and $\mathcal{X}_M$ is defined by

$$\begin{gathered}\mathcal{X}_2 = [\mathcal{X}_{(2)}, -\mathcal{X}_{(1)}],\\ \mathcal{X}_n = \left[\begin{array}{ccc|c} & \mathcal{X}_{n-1} & & \mathbf{0}\\ \hline \mathcal{X}_{(n)} & & \mathbf{0} & -\mathcal{X}_{(1)}\\ & \ddots & & \vdots\\ \mathbf{0} & & \mathcal{X}_{(n)} & -\mathcal{X}_{(n-1)}\end{array}\right],\end{gathered} \tag{41}$$

with $n = 3,\ldots,M$, and

$$\mathcal{X}_{(n)} = \begin{bmatrix} x_n(K) & \cdots & x_n(0)\\ \vdots & & \vdots\\ x_n(T-1) & \cdots & x_n(T-K-1)\end{bmatrix}. \tag{42}$$

In the presence of noise, (40) can be naturally solved in the least-squares (LS) sense according to

$$\hat{\mathbf{h}}_i = \arg\min_{\|\mathbf{h}\|=1}\mathbf{h}^H\mathcal{X}_M^H\mathcal{X}_M\mathbf{h}, \tag{43}$$

which solution is given by the least eigenvector of matrix $\mathcal{X}_M^H\mathcal{X}_M$.

Remark 1. We have presented here a basic version of the CR method. In [33], an improved version of the method (introduced in the adaptive scheme) is proposed exploiting the quasisparse nature of acoustic impulse responses.

5.1.3. Clustering of channel vector estimates

The first step of our channel estimation method consists in detecting the time slots where only one single source signal is "effectively" present. However, the same source signal s_i may be present in several time intervals (see Figures 3 and 4) leading to several estimates of the same channel vector $\mathbf{h}_i$.

We end up, finally, with several estimates of each source channel that we need to group together into N classes. This is done by clustering the estimated vectors using k-means algorithm. The ith channel estimate is evaluated as the centroid of the ith class.

5.2. Component grouping and source estimation

For the synthesis of the source signals, one observes that the quasiorthogonality assumption leads to

$$\widehat{\mathbf{h}}_i^j = \frac{\langle \mathbf{x} \mid \hat{c}_i^j \rangle}{\|\hat{c}_i^j\|^2} \propto \mathbf{h}_i(z_i^j), \tag{44}$$

where $z_i^j = e^{d_i^j + \jmath\omega_i^j}$ is the pole of the component $\hat{c}_i^j$, that is, $\hat{c}_i^j(t) = \Re e\{\alpha_i^j (z_i^j)^t\}$. Therefore, we propose to gather these components by minimizing the criterion[3]:

$$\hat{c}_i^j \in \mathcal{C}_i \Longleftrightarrow i = \arg\min_l \left(\min_\alpha \|\widehat{\mathbf{h}}_i^j - \alpha \mathbf{h}_l(z_i^j)\|^2\right), \tag{45}$$

$$i = \arg\min_l \left\{ \|\widehat{\mathbf{h}}_i^j\|^2 - \frac{|\mathbf{h}_l^H(z_i^j)\widehat{\mathbf{h}}_i^j|^2}{\|\mathbf{h}_l(z_i^j)\|^2} \right\}, \tag{46}$$

where $\mathbf{h}_l$ is the lth column of $\mathbf{H}$ estimated in Section 5.1 and $\mathbf{h}_l(z_j^k)$ is computed by

$$\mathbf{h}_l(z_i^j) = \sum_{k=0}^{K} \mathbf{h}_l(k)(z_i^j)^{-k}. \tag{47}$$

One will be able to rebuild the initial sources up to a constant by adding the various components within a same class using (17).

Similar to the instantaneous mixture case, one modal component can be assigned to two or more source signals, which relaxes the quasiorthogonality assumption and improves the estimation accuracy at moderate and high SNRs (see Figure 9).

6. DISCUSSION

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

(i) Overdetermined case

In that case, one is able to separate the sources by left inversion of matrix $\mathbf{A}$ (or matrix $\mathbf{H}$ in the convolutive case). The latter can be estimated from the centroids of the N clusters (i.e., the centroid of the ith cluster represents the estimate of the ith column of $\mathbf{A}$).

(ii) Estimation of the number of sources

This is a difficult and challenging task in the underdetermined case. Few approaches exist based on multidimensional tensor decomposition [34] or based on the clustering with joint estimation of the number of classes [24]. 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 assumed that the number of sources is known (or correctly estimated).

(iii) Number of modal components

In the parametric approach, we have to choose the number of modal components L_{tot} needed to well-approximate the audio signal. Indeed, small values of L_{tot} lead to poor signal representation while large values of L_{tot} increase the computational cost. In fact, L_{tot} depends on the "signal complexity," and in general musical signals require less components (for a good modeling) than speech signals [35]. In Section 7, we illustrate the effect of the value of L_{tot} on the separation quality.

(iv) Hybrid separation approach

It is most probable that the separation quality can be further improved using signal analysis in conjunction with spatial filtering or interference cancelation as in [28]. 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.

(v) SIMO versus MIMO channel estimation

We have opted here to estimate the channels using SIMO techniques. However, it is also possible to estimate the channels using overdetermined blind MIMO techniques by considering the time slots where the number of sources is smaller than $(M-1)$ instead of using only those where the number of "effective" sources is one. The advantage of doing so would be the use of a larger number of time slots (see Figure 4). The drawback resides in the fact that blind identification of MIMO systems is more difficult compared to the SIMO case and leads in particular to higher estimation error (see Figure 12 for a comparative performance evaluation).

[3] We minimize over the scalar α because of the inherent indeterminacy of the blind channel identification, that is, $\mathbf{h}_i(z)$ is estimated up to a scalar constant as shown by Theorem 1.

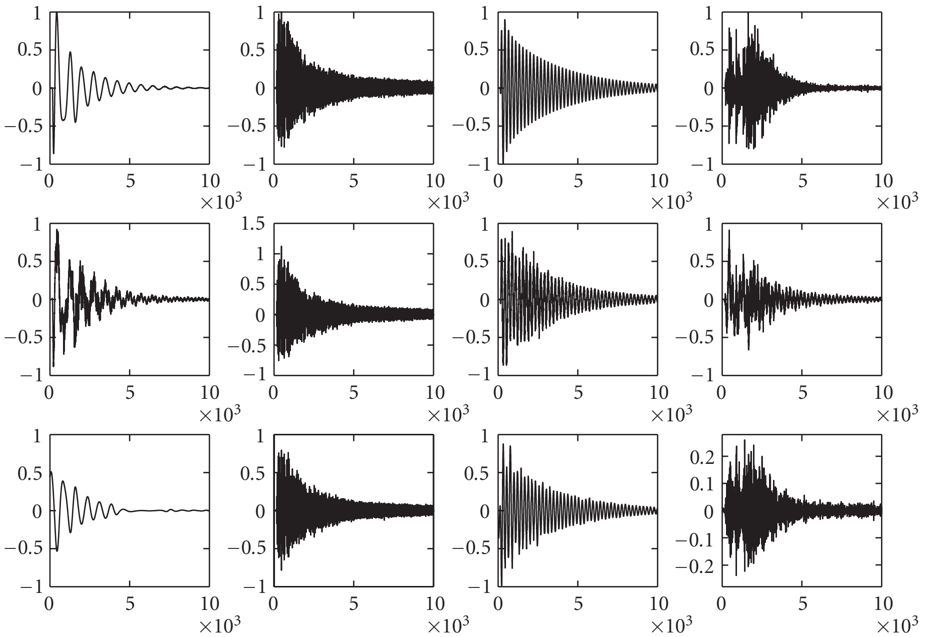

Figure 5: Blind source separation example for 4 audio sources and 3 sensors in instantaneous mixture case: the upper line represents the original source signals, the second line represents the source estimation by pseudoinversion of mixing matrix **A** assumed exactly known and the bottom one represents estimates of sources by our algorithm using EMD.

(vi) Noiseless case

In the noiseless case (with perfect modelization of the sources as sums of damped sinusoids), the estimation of the modal components using ESPRIT would be perfect. This would lead to perfect (exact) estimation of the mixing matrix column vectors using least-squares filtering, and hence perfect clustering and source restoration.

7. SIMULATION RESULTS

We present here some simulation results to illustrate the performance of our blind separation algorithms. For that, we consider first an instantaneous mixture with a uniform linear array of $M = 3$ sensors receiving the signals from $N = 4$ audio sources (except for the third experiment where N varies in the range $[2 \cdots 6]$). The angle of arrivals (AOAs) of the sources is chosen randomly.[4] In the convolutive mixture case, the filter coefficients are chosen randomly and the channel order is $K = 6$. The sample size is set to $T = 10000$ samples (the signals are sampled at a rate of 8 KHz). 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 (NMSEs) of the sources evaluated over $N_r = 100$ Monte Carlo runs. The plots represent the averaged NMSE over the N sources:

$$\begin{aligned}
\text{NMSE}_i &\overset{\text{def}}{=} \frac{1}{N_r}\sum_{r=1}^{N_r} \min_{\alpha}\left(\frac{\left\|\alpha\hat{\mathbf{s}}_{i,r} - \mathbf{s}_i\right\|^2}{\left\|\mathbf{s}_i\right\|^2}\right),\\
\text{NMSE}_i &= \frac{1}{N_r}\sum_{r=1}^{N_r} 1 - \left(\frac{\hat{\mathbf{s}}_{i,r}\mathbf{s}_i^T}{\left\|\hat{\mathbf{s}}_{i,r}\right\|\left\|\mathbf{s}_i\right\|}\right)^2, \qquad (48)\\
\text{NMSE} &= \frac{1}{N}\sum_{i=1}^{N}\text{NMSE}_i,
\end{aligned}$$

where $\mathbf{s}_i \overset{\text{def}}{=} [s_i(0), \dots, s_i(T-1)]$, $\hat{\mathbf{s}}_{i,r}$ (defined similarly) is the rth estimate of source $\mathbf{s}_i$, and α is a scalar factor that compensates for the scale indeterminacy of the BSS problem.

In Figure 5, we present a simulation example with $N = 4$ audio sources. The upper line represents the original source signals, the second line represents the source estimation by pseudoinversion of mixing matrix **A** assumed exactly known, and the bottom one represents estimates of the sources by our algorithm.

In Figure 6, we compare the separation performance obtained by our algorithm using EMD and the parametric technique with $L = 30$ modal components per source signal ($L_{\text{tot}} = NL$). As a reference, we plot also the NMSE obtained by pseudoinversion of matrix **A** [36] (assumed exactly known). It is observed that both EMD and parametric-based separation provide better results than those obtained by pseudoinversion of the exact mixing matrix.

The plots in Figure 7 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.

[4] This is used here just for the simulation to generate the mixture matrix **A**. We do not consider a parametric model using sources AOAs in our separation algorithm.

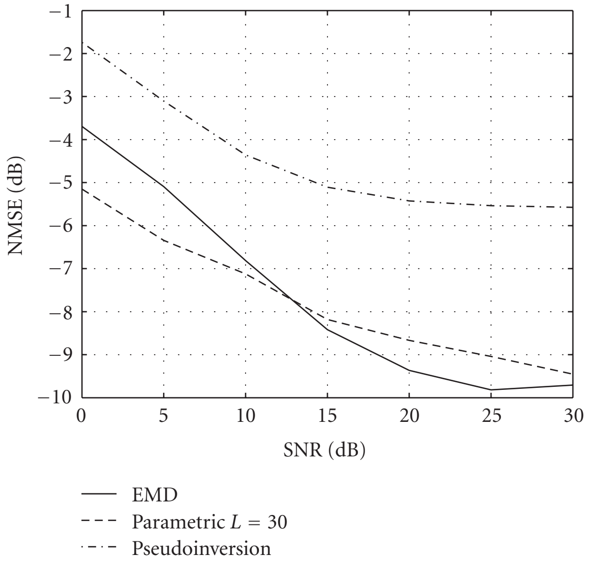

Figure 6: NMSE versus SNR for 4 audio sources and 3 sensors in instantaneous mixture case: comparison of the performance of our algorithm (EMD and ESPRIT) with those given by the pseudoinversion of mixing matrix **A** (assumed exactly known).

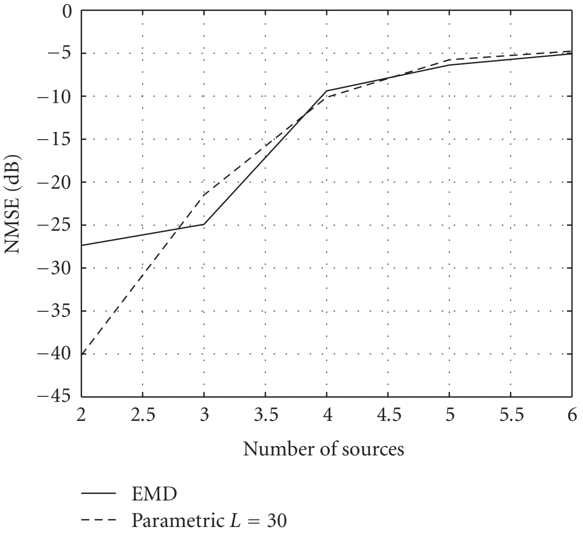

Figure 8: NMSE versus N for 3 sensors in instantaneous mixture case: comparison of the performance of our algorithm (EMD and ESPRIT) for $N \in [2, \ldots, 6]$.

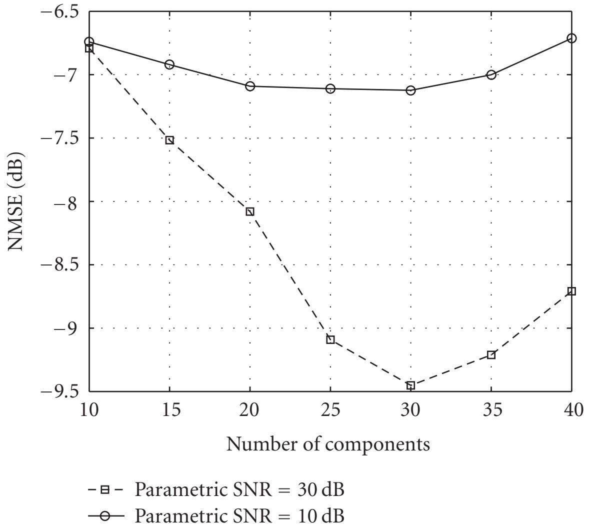

Figure 7: NMSE versus L for 4 audio sources and 3 sensors in instantaneous mixture case: comparison of the performance of our algorithm (ESPRIT) for L varying in the range $[10, \ldots, 40]$ with SNR = 10 dB and SNR = 30 dB.

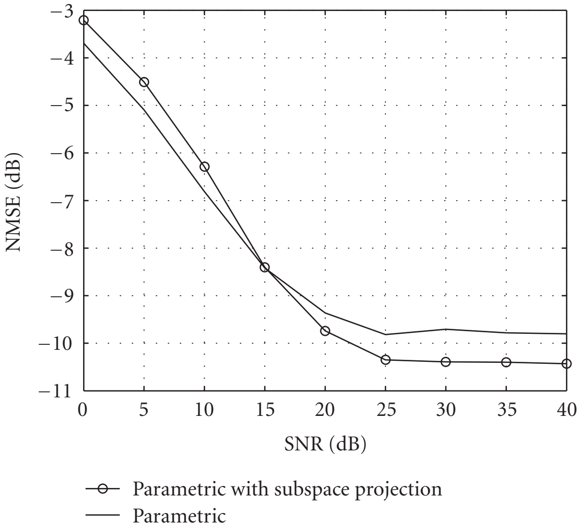

Figure 9: NMSE versus SNR for 4 audio sources and 3 sensors in instantaneous mixture case: comparison of the performance of our algorithm (EMD) and the same algorithm with subspace projection.

In other words, there exists an optimal choice of L that depends on the signal type.

In Figure 8, we compare 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 $N = 3$ (overdetermined case), we estimate the sources by left inversion of the estimate of matrix **A**. In the underdetermined case, the EMD and parametric-based algorithms present similar performance. However, the latter method is better in the overdetermined case.

In Figure 9, we compare the performance of our algorithm using ESPRIT with and without subspace projection. One can observe that using the subspace projection leads to a performance gain at moderate and high SNRs. At low SNRs, the performance is slightly degraded due to the noise effect. Indeed, when a given component belongs “effectively”

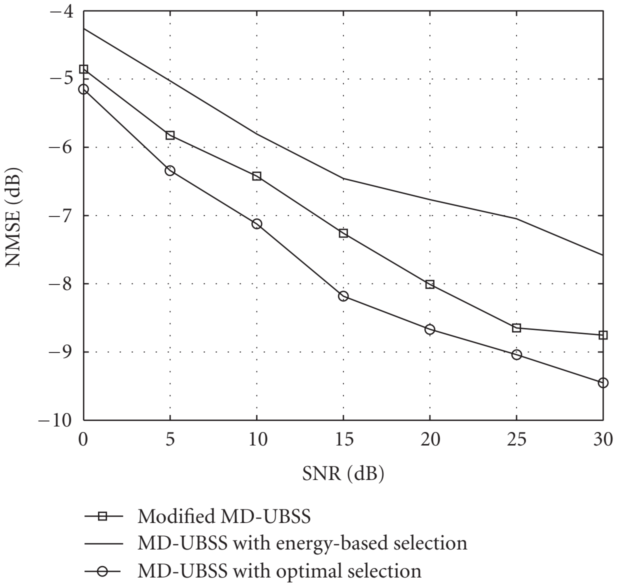

Figure 10: NMSE versus SNR for 4 audio sources and 3 sensors: comparison of the performance of MD-UBSS algorithms with and without quasiorthogonality assumption.

to only one source signal, (23) would provide a nonzero amplitude coefficient for the second source due to noise effect which explains the observed degradation.

In Figure 10, we compare the separation performance obtained by our MD-UBSS algorithm and the modified MD-UBSS algorithm. We observe a performance gain in favor of the modified MD-UBSS mainly due to the fact that it does not rely on the quasiorthogonality assumption. This plot also highlights the problem of "best source estimate" selection related to the MD-UBSS as we observe a performance loss between the results given by the proposed energy-based selection procedure and the optimal[5] one using the exact source signals.

Figure 11 illustrates the estimation performance of the mixing matrix **A** using proposed clustering method. The observed good estimation performance translates the fact that most modal components belong "effectively" to one single source signal.

In Figure 12, we present the performance of channel identification obtained by using SIMO identification algorithm (in this case, we choose only the time intervals where only one source is present using AIC criterion) with SIMO and MIMO identification algorithms (in this case, we choose the time intervals where we are in the overdetermined case; i.e., where $p = 1$ or $p = 2$). It is observed that SIMO-based identification provides better results than those obtained by SIMO and MIMO identification algorithms.

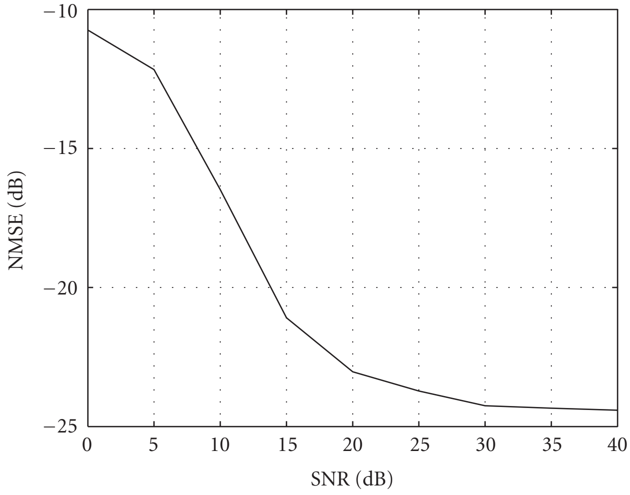

Figure 11: Mixing matrix estimation: NMSE versus SNR for 4 speech sources and 3 sensors in instantaneous mixture case.

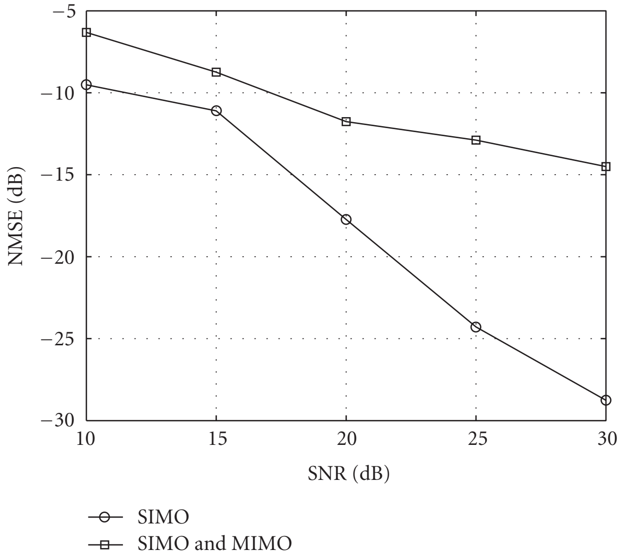

Figure 12: NMSE versus SNR for 4 audio sources and 3 sensors in convolutive mixture case: comparison of the performance of identification algorithm using only SIMO system and the algorithm using SIMO and MIMO systems.

The plots in Figure 13 present the separation performance in convolutive mixture case when using the exact channel impulse response **H** compared to that obtained with an approximate channel $\widehat{\mathbf{H}} = \mathbf{H} + \delta\mathbf{H}$, where the entries of $\delta\mathbf{H}$ are i.i.d. Gaussian distributed. This is done for different values of channel normalized mean-squares error (CNMSE) defined by

$$\text{CNMSE} = 10\log\frac{\|\mathbf{H} - \widehat{\mathbf{H}}\|^2}{\|\mathbf{H}\|^2}. \tag{49}$$

[5] Clearly, the optimal selection procedure is introduced here just for performance comparison and not as an alternative selection method since it relies on the exact source signals that are unavailable in our context.

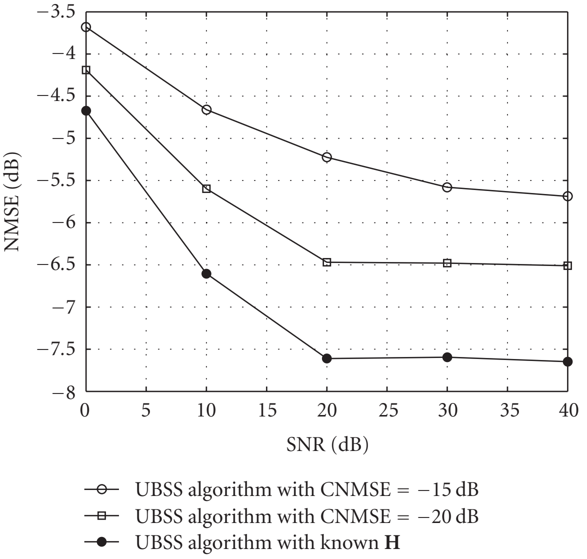

Figure 13: NMSE versus SNR for 4 audio sources and 3 sensors in convolutive mixture case: comparison, for the MD-UBSS algorithm in convolutive mixture case, when the channel response **H** is known or disturbed by Gaussian noise for different values of CNMSE.

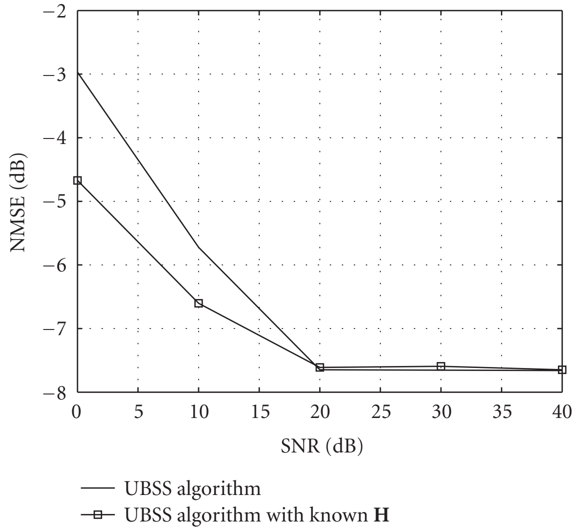

Figure 14: NMSE versus SNR for 4 audio sources and 3 sensors in convolutive mixture case: comparison, for the MD-UBSS algorithm in convolutive mixture case, when the channel response **H** is known or estimated using the CR technique.

Clearly, the separation quality depends strongly on the quality of channel estimation.

In Figure 14, we present the separation performance when using the exact channel response **H** compared to that obtained with the proposed estimate $\widehat{\mathbf{H}}$ using SIMO approach. For SNRs larger than 20 dB, the channel estimation is good enough for the proposed method to achieve almost the same performance as if the channel is exactly known. Surprisingly, at SNR = 20 dB, the channel estimate NMSE is approximately equal to −18 dB (see Figure 12), an error level corresponding to a nonnegligible degradation shown in Figure 13. This seemingly contradiction comes from the fact that in the experiment of Figure 13, the channel is disturbed "artificially" using spatially white Gaussian noise, while the real channel estimation error is spatially colored (see, e.g., [37] where explicit expression of the asymptotic channel covariance error is given) which seems to be favorable to our separation method.

8. 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 pseudoinversion of the mixture matrix (even if the latter is known exactly) in the instantaneous mixture case. The separation method proceeds in two steps: an analysis step where all modal components are estimated followed by a synthesis step to group (cluster) together the modal components and reconstruct the source signals. For the signal analysis step, two algorithms are used and compared based, respectively, on the EMD and on the ESPRIT techniques. A modified MD-UBSS as well as a subspace projection approach are also proposed to relax the "quasiorthogonality" assumption and allow the source signals to share common modal components, respectively. This approach leads to a performance improvement of the separation quality. For the convolutive mixture case, we propose to use again modal decomposition based on ESPRIT technique, but the signal synthesis is more complex and requires the prior identification of the channel impulse response, which is done here using the sparsity of the audio sources.

ACKNOWLEDGMENT

Part of this work has been published in conferences [38, 39].

REFERENCES

[1] A. K. Nandi, Ed., *Blind Estimation Using Higher-Order Statistics*, Kluwer Academic, Boston, Mass, USA, 1999.

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

[3] J.-F. Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2009–2025, 1998.

[4] P. Sugden and N. Canagarajah, "Underdetermined noisy blind separation using dual matching pursuits," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '04)*, vol. 5, pp. 557–560, Montreal, Que, Canada, May 2004.

[5] P. Sugden and N. Canagarajah, "Underdetermined blind separation using learned basis function sets," *Electronics Letters*, vol. 39, no. 1, pp. 158–160, 2003.

[6] P. Comon, "Blind identification and source separation in 2 × 3 under-determined mixtures," *IEEE Transactions on Signal Processing*, vol. 52, no. 1, pp. 11–22, 2004.

[7] A. Belouchrani and J. F. Cardoso, "A maximum likelihood source separation for discrete sources," in *Proceedings of the 7th European Signal Processing Conference (EUSIPCO '94)*, vol. 2, pp. 768–771, Scotland, UK, September 1994.

[8] J. M. Peterson and S. Kadambe, "A probabilistic approach for blind source separation of underdetermined convolutive mixtures," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '03)*, vol. 6, pp. 581–584, Hong Kong, April 2003.

[9] S. Y. Low, S. Nordholm, and R. Togneri, "Convolutive blind signal separation with post-processing," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 539–548, 2004.

[10] L. C. Khor, W. L. Woo, and S. S. Dlay, "Non-sparse approach to underdetermined blind signal estimation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '05)*, vol. 5, pp. 309–312, Philadelphia, Pa, USA, March 2005.

[11] P. Georgiev, F. Theis, and A. Cichocki, "Sparse component analysis and blind source separation of underdetermined mixtures," *IEEE Transactions on Neural Networks*, vol. 16, no. 4, pp. 992–996, 2005.

[12] I. Takigawa, M. Kudo, and J. Toyama, "Performance analysis of minimum ℓ_1-norm solutions for underdetermined source separation," *IEEE Transactions on Signal Processing*, vol. 52, no. 3, pp. 582–591, 2004.

[13] N. Linh-Trung, A. Belouchrani, K. Abed-Meraim, and B. Boashash, "Separating more sources than sensors using time-frequency distributions," *EURASIP Journal on Applied Signal Processing*, vol. 2005, no. 17, pp. 2828–2847, 2005.

[14] Ö. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1846, 2004.

[15] Y. Li, S.-I. Amari, A. Cichocki, D. W. C. Ho, and S. Xie, "Underdetermined blind source separation based on sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 2, pp. 423–437, 2006.

[16] N. E. Huang, Z. Shen, S. R. Long, et al., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A*, vol. 454, no. 1971, pp. 903–995, 1998.

[17] P. Flandrin, G. Rilling, and P. Gonçalvès, "Empirical mode decomposition as a filter bank," *IEEE Signal Processing Letters*, vol. 11, no. 2, part 1, pp. 112–114, 2004.

[18] R. Boyer and K. Abed-Meraim, "Audio modeling based on delayed sinusoids," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 2, pp. 110–120, 2004.

[19] J. Nieuwenhuijse, R. Heusens, and Ed. F. Deprettere, "Robust exponential modeling of audio signals," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '98)*, vol. 6, pp. 3581–3584, Seattler, Wash, USA, May 1998.

[20] D. Nuzillard and J.-M. Nuzillard, "Application of blind source separation to 1-D and 2-D nuclear magnetic resonance spectroscopy," *IEEE Signal Processing Letters*, vol. 5, no. 8, pp. 209–211, 1998.

[21] H. Park, S. Van Huffel, and L. Elden, "Fast algorithms for exponential data modeling," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '94)*, vol. 4, pp. 25–28, Adelaide, SA, Australia, April 1994.

[22] C. Serviere, V. Capdevielle, and J.-L. Lacoume, "Separation of sinusoidal sources," in *Proceedings of IEEE Signal Processing Workshop on Higher-Order Statistics*, pp. 344–348, Banff, Canada, July 1997.

[23] P. D. O'Grady, B. A. Pearlmutter, and S. T. Rickard, "Survey of sparse and non-sparse methods in source separation," *International Journal of Imaging Systems and Technology*, vol. 15, no. 1, pp. 18–33, 2005.

[24] I. E. Frank and R. Todeschini, *The Data Analysis Handbook*, Elsevier Science, Amsterdam, The Netherlands, 1994.

[25] G. Rilling, P. Flandrin, and P. Gonçalvès, "Empirical mode decomposition," http://perso.ens-lyon.fr/patrick.flandrin/emd.html.

[26] S. Y. Kung, K. S. Arun, and D. V. Bhaskar Rao, "State space and singular value decomposition based on approximation methods for harmonic retrieval," *Journal of the Optical Society of America*, vol. 73, no. 12, pp. 1799–1811, 1983.

[27] J. Rosier and Y. Grenier, "Unsupervised classification techniques for multipitch estimation," in *Proceedings of the 116th Convention of the Audio Engineering Society (AES '04)*, Berlin, Germany, May 2004.

[28] Y. Huang, J. Benesty, and J. Chen, "A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, part 2, pp. 882–895, 2005.

[29] B. Albouy and Y. Deville, "Alternative structures and power spectrum criteria for blind segmentation and separation of convolutive speech mixtures," in *Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA '03)*, pp. 361–366, Nara, Japan, April 2003.

[30] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 387–392, 1985.

[31] G. Xu, H. Liu, L. Tong, and T. Kailath, "A least-squares approach to blind channel identification," *IEEE Transactions on Signal Processing*, vol. 43, no. 12, pp. 2982–2993, 1995.

[32] A. Aïssa-El-Bey, M. Grebici, K. Abed-Meraim, and A. Belouchrani, "Blind system identification using cross-relation methods: further results and developments," in *Proceedings of the 7th International Symposium on Signal Processing and Its Applications (ISSPA '03)*, vol. 1, pp. 649–652, Paris, France, July 2003.

[33] R. Ahmad, A. W. H. Khong, and P. A. Naylor, "Proportionate frequency domain adaptive algorithms for blind channel identification," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '06)*, vol. 5, pp. 29–32, Toulouse, France, May 2006.

[34] L. De Lathauwer, B. De Moor, and J. Vandewalle, "ICA techniques for more sources than sensors," in *Proceedings of the IEEE Signal Processing Workshop on Higher-Order Statistics*, pp. 121–124, Caesarea, Israel, June 1999.

[35] J. Jensen and R. Heusdens, "A comparison of sinusoidal model variants for speech and audio representation," in *Proceedings of the 11th European Signal Processing Conference (EUSIPCO '02)*, vol. 1, pp. 479–482, Toulouse, France, September 2002.

[36] M. Z. Ikram, "Blind separation of delayed instantaneous mixtures: a cross-correlation based approach," in *Proceedings of the 2nd IEEE International Symposium on Signal Processing and Information Technology (ISSPIT '02)*, Marrakesh, Morocco, December 2002.

[37] W. Qiu and Y. Hua, "Performance comparison of subspace and cross-relation methods for blind channel identification," *Signal Processing*, vol. 50, no. 1-2, pp. 71–81, 1996.

[38] A. Aïssa-El-Bey, K. Abed-Meraim, and Y. Grenier, "Blind separation of audio sources using modal decomposition," in *Proceedings of the 8th International Symposium on Signal Processing and Its Applications (ISSPA '05)*, vol. 2, pp. 451–454, Sydney, Australia, August 2005.

[39] A. Aïssa-El-Bey, K. Abed-Meraim, and Y. Grenier, "Séparation aveugle sous-déterminée de sources audio par la méthode EMD (Empirical Mode Decomposition)," in *Actes 20e Colloque GRETSI sur le Traitement du Signal et des Images*, vol. 2, pp. 1233–1236, Louvain-La-Neuve, Belgium, September 2005.