Underdetermined audio source separation using fast parametric decomposition

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

UNDERDETERMINED AUDIO SOURCE SEPARATION USING FAST PARAMETRIC DECOMPOSITION

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

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

ABSTRACT

In this paper, we consider the problem of underdetermined blind source separation 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. Our contributions in this paper are a new separation method with relaxed assumption and reduced computational cost compared to other existing algorithms. Simulation results are given to assess the performance of the proposed algorithm.

1. INTRODUCTION

The objective of blind source separation (BSS) is to extract the original source signals from their mixtures using only the information within the observed mixtures with no, or very limited knowledge about the source signals and the mixing matrix. BSS problem arises in many fields, such as noise reduction, radar and sonar processing, speech enhancement, separation of rotating machine noises, biomedical signal processing and even in optical tracking system [1]. This problem has been intensively studied in the literature and many effective solutions have been proposed so far [1]. In the particular case where the number of sources is larger than the number of observed mixtures (underdetermined BSS case (UBSS)), the separation can be achieved only if side information about the sources is available (sparseness, W-disjointness, finite alphabet sources, etc). In the case of non-stationary signals (including the audio signals), certain solutions using time-frequency (TF) analysis of the observations and the sources TF-orthogonality exist for the underdetermined case [2, 3].

In this paper, we propose an alternative approach using modal decomposition (MD) of the received signals [4]. More precisely we propose to decompose the signal into its various modes. The audio signals and more particularly the musical signals can be modeled by a sum of damped sinusoids [5] 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. To start, we review the MD-UBSS approach presented in [4], then we propose an improved algorithm that reduces the computational cost and relax some of the working assumptions.

In this paper, we use bold upper and lower case letters for matrices and vectors, respectively. The remaining notational conventions and major symbols are listed as follows:

$(\cdot)^*$	Complex conjugation.
$(\cdot)^T$	Transpose.
$(\cdot)^H$	Transpose conjugate.
$\|\cdot\|$	Frobenius norm.
$\mathbf{I}$	Identity matrix.

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 $x_1(t)$, $\ldots, x_M(t)$ which 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 assume that for any pair (i, j) with $i \neq j$, the vectors $\mathbf{a}_i$ and $\mathbf{a}_j$ are linearly independent. It is known that BSS is only possible up to some scaling and permutation [6]. We take advantage of these indeterminacies to assume without loss of generality that the column vectors of $\mathbf{A}$ are of unit norm i.e. $\|\mathbf{a}_i\| = 1$ for $1 \leq i \leq N$.

The considered 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

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

3. MD-UBSS ALGORITHM

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 the sensor outputs in order to extract all the harmonic components from them.
- *A synthesis step*: this is to 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. Parametric signal analysis

The source signal and hence the observations are modeled as sum of damped sinusoids:

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

where $\alpha_{l,k}$ represents the complex amplitude and $z_l = e^{d_l + \jmath\omega_l}$ is the l^{th} pole where d_l is the negative damping factor and ω_l is the angular-frequency. $\Re e(\cdot)$ 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 Kung's algorithm given in [7] 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.

3.2. 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[1]. One will be able to rebuild the initial sources up to a constant by adding the various components within a same class.

3.3. Existing MD-UBSS algorithm [4]

This algorithm applies the previous analysis and synthesis steps to each signal output $x_k(t)$. By doing so, one obtains M estimates of each source signal with an estimation quality varying significantly from one sensor to another. Indeed, this latter depends strongly on the mixing matrix coefficients and, in particular, on the signal to interference ratio (SIR) of the desired source. Consequently, a blind selection method to choose a 'good' estimate among the M available is proposed in [4]. First, the source estimates are paired together 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, the source estimate of maximal energy is selected, i.e.

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

where E_i^j represents the energy of the i^{th} source extracted from the j^{th} sensor $\widehat{s}_i^j(t)$.

3.4. Proposed MD-UBSS algorithm

We propose here to improve the previous algorithm w.r.t the computational cost and the estimation accuracy when Assumption 4 (Equation (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-like technique on the averaged data covariance matrix $\mathbb{H}(\mathbf{x})$ define by:

$$\mathbb{H}(\mathbf{x}) = \sum_{i=1}^{M} \mathcal{H}(x_i)\mathcal{H}(x_i)^H \tag{11}$$

and we apply steps 1 to 4 of Kung's algorithm described in Section 3.1 to obtain all the poles $z_i,\ \ i = 1, \dots, L$. 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 re-write the data model as:

$$\boldsymbol{\Gamma}\mathbf{z}(t) = \mathbf{x}(t) \tag{12}$$

where $\boldsymbol{\Gamma} \stackrel{\text{def}}{=} [\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_1^*, \cdots, \boldsymbol{\gamma}_L, \boldsymbol{\gamma}_L^*]$, $\boldsymbol{\gamma}_i = \beta_i e^{\jmath\phi_i}\mathbf{b}_i$, where $\mathbf{b}_i$ is a unit norm vector representing the spatial direction of the i^{th} component (i.e. $\mathbf{b}_i = \mathbf{a}_k/\|\mathbf{a}_k\|$ if the i^{th} component belongs to the k^{th} source signal) and $\mathbf{z}(t) \stackrel{\text{def}}{=} [z_1^t, (z_1^*)^t, \cdots, z_L^t, (z_L^*)^t]^T$.

[1] In the simulation, we have used the k-means algorithm in [8] for vector clustering.

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

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

$$\min_{\mathbf{\Gamma}} \|\mathbf{X} - \mathbf{\Gamma}\boldsymbol{\mathcal{Z}}\|^2 \Leftrightarrow \mathbf{\Gamma} = \mathbf{X}\boldsymbol{\mathcal{Z}}^{\#} \tag{13}$$

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

$$\phi_i = \frac{\arg(\boldsymbol{\gamma}_{2i}^H \boldsymbol{\gamma}_{2i-1})}{2} \tag{14}$$

The spatial direction of each modal component is estimated by:

$$\widehat{\mathbf{v}}_i = \boldsymbol{\gamma}_{2i-1} e^{-\jmath\phi_i} + \boldsymbol{\gamma}_{2i} e^{\jmath\phi_i} = 2\beta_i \mathbf{b}_i \; . \tag{15}$$

Finally, we group together these components by clustering the vectors $\widehat{\mathbf{v}}_i$ into N classes. After clustering, we obtain N classes with N centroids $\widehat{\mathbf{a}}_1, \ldots, \widehat{\mathbf{a}}_N$ corresponding to the estimates of the column vectors of the mixing matrix $\mathbf{A}$. If the pole z_i belongs to the j^{th} class, then according to (15), its amplitude can be estimated by:

$$\beta_i = \frac{\widehat{\mathbf{a}}_j^H \widehat{\mathbf{v}}_i}{2} \; . \tag{16}$$

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

$$\widehat{s}_i(t) = \Re e \left\{ \sum_{j \in \mathcal{C}_i} \beta_j e^{\jmath\phi_j} z_j^t \right\}. \tag{17}$$

4. NON-DISJOINT SOURCES CASE

We consider here the case where a given component $c_j^k(t)$ can be shared by several sources. This is the case, for example, for certain musical signals such as those treated in [9]. To simplify, we suppose that a component belongs to at most two sources. Thus, let us suppose that the component $c_j^k(t)$ 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. It follows that the spatial direction associated with this component as estimated by (15), is given by:

$$\widehat{\mathbf{v}}_i \approx \alpha_{j_1} \mathbf{a}_{j_1} + \alpha_{j_2} \mathbf{a}_{j_2}. \tag{18}$$

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 assume that $M > 2$ and matrix $\mathbf{A}$ is known and satisfies the condition that any triplet of its column vectors are linearly independent. Consequently, we have:

$$\mathbf{P}_{\widetilde{\mathbf{A}}}^{\perp} \mathbf{v}_i = 0, \tag{19}$$

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 $\widetilde{\mathbf{A}}$, i.e.

$$\mathbf{P}_{\widetilde{\mathbf{A}}}^{\perp} = \mathbf{I} - \widetilde{\mathbf{A}} \left(\widetilde{\mathbf{A}}^H \widetilde{\mathbf{A}} \right)^{-1} \widetilde{\mathbf{A}}^H. \tag{20}$$

In practice, by taking into account the noise, one detects the columns j_1 and j_2 by minimizing:

$$(j_1, j_2) = \arg\min_{(l,m)} \left\{ \|\mathbf{P}_{\widetilde{\mathbf{A}}}^{\perp} \widehat{\mathbf{v}}_i\| \mid \widetilde{\mathbf{A}} = [\mathbf{a}_l \; \mathbf{a}_m] \right\}. \tag{21}$$

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

$$\begin{bmatrix} \alpha_{j_1} \\ \alpha_{j_2} \end{bmatrix} = \widetilde{\mathbf{A}}^{\#} \widehat{\mathbf{v}}_i. \tag{22}$$

In the simulation, the optimization problem of (21) is solved using exhaustive search. This is computationally tractable for small vector array sizes but would be prohibitive if M is very large. 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 (22) should be zero or close to zero. Also, in what precedes, the mixing matrix $\mathbf{A}$ is supposed to be known. This means, it has to be estimated before applying a subspace projection. This is performed here by clustering all the spatial direction vectors in (15) as for the preview MD-UBSS algorithm. Then, the i^{th} column vector of $\mathbf{A}$ is estimated as the centroid of C_i assuming implicitly that most modal components belong mainly to one source signal. This is confirmed by our simulation experiment shown in Figure 2.

5. 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. The angles of arrival of the sources are chosen randomly. The sample size is set to $T = 10000$ samples (the signals are sampled at a rate of 22 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 (NMSE) of the sources evaluated over 200 Monte-Carlo runs and defined as:

$$NMSE_i \quad \stackrel{\text{def}}{=} \quad \frac{1}{N_r} \sum_{r=1}^{N_r} \min_{\alpha} \left(\frac{\|\alpha \widehat{\mathbf{s}}_{i,r} - \mathbf{s}_i\|^2}{\|\mathbf{s}_i\|^2} \right) \tag{23}$$

$$NMSE_i \quad = \quad \frac{1}{N_r} \sum_{r=1}^{N_r} 1 - \left(\frac{\widehat{\mathbf{s}}_{i,r} \mathbf{s}_i^H}{\|\widehat{\mathbf{s}}_{i,r}\| \|\mathbf{s}_i\|} \right)^2 \tag{24}$$

$$NMSE \quad = \quad \frac{1}{N} \sum_{i=1}^{N} NMSE_i \; . \tag{25}$$

where $\mathbf{s}_i \stackrel{\text{def}}{=} [s_i(0), \ldots, s_i(T-1)]$, $\widehat{\mathbf{s}}_{i,r}$ (defined similarly) represents the r^{th} estimate of source $\mathbf{s}_i$ and α is a scalar factor that compensate for the scale indeterminacy of the BSS problem. In Figure 1, we compare the separation performance obtained

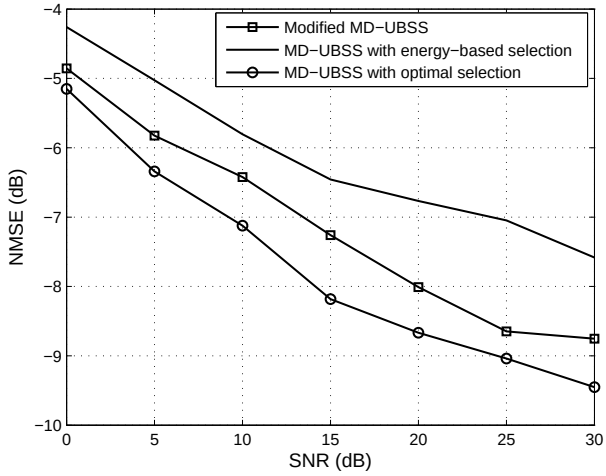

Fig. 1. NMSE versus SNR for 4 audio sources and 3 sensors: comparison of the performance of MD-UBSS algorithms with and without quasi-orthogonality assumption.

by existing MD-UBSS algorithm and the new MD-UBSS algo-

rithm. We observe a performance gain in favor of the new MD-UBSS due mainly to the fact that it does not rely on the quasi-orthogonality 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[3] one using the exact source signals. Figure 2 illustrates the estimation

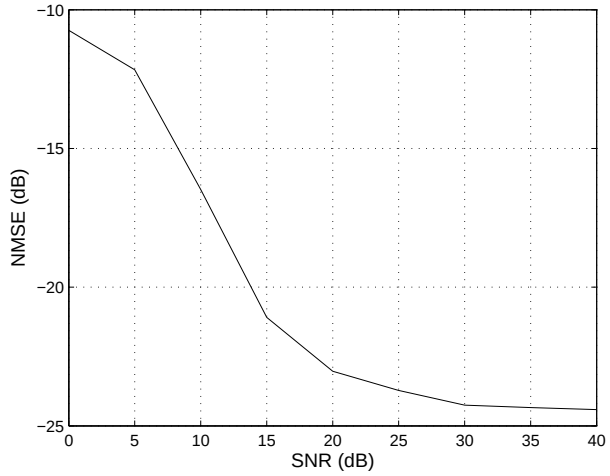

Fig. 2. Mixing matrix estimation: NMSE versus SNR for 4 speech sources and 3 sensors.

performance of the mixing matrix $\mathbf{A}$ 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 3, we compare the performance of

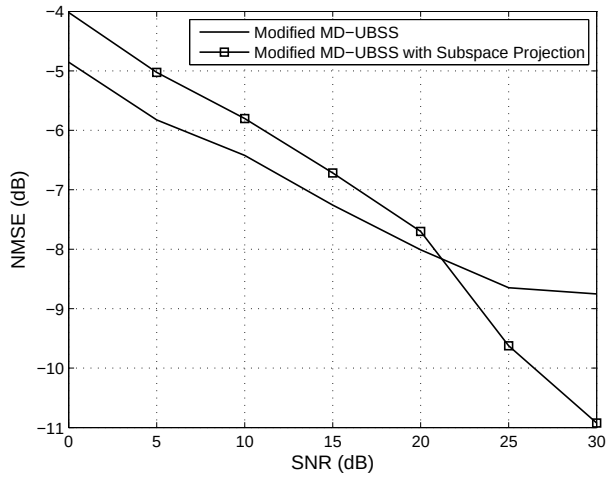

Fig. 3. NMSE versus SNR for 4 audio sources and 3 sensors: comparison of the performance of modified MD-UBSS algorithm and the same algorithm with subspace projection.

new MD-UBSS algorithm and the same algorithm with 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' to only one source signal, equation (22) would provide a non zero amplitude coefficient for the second source due to noise effect which explains the observed degradation.

[3]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.

6. CONCLUSION

This paper introduces a new MD-UBSS algorithm of audio sources. The main advantages over the proposed MD-UBSS algorithm are, a reduced the computational cost and relaxed a quasi-orthogonality assumption. Moreover, this algorithm is extended to the non-disjoint sources case using an approximate subspace projection technique. Simulation results illustrate the effectiveness of our algorithm compared to the one [4].

7. REFERENCES

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

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

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

[4] A. Aïssa-El-Bey, K. Abed-Meraim, and Y. Grenier, "Blind separation of audio sources using modal decomposition," in *Proc. ISSPA*, Sydney, Australia, August 2005, vol. 2, pp. 451–454.

[5] 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, March 2004.

[6] J. F. Cardoso, "Blind signal separation: statistcal principles," *Proceeding of the IEEE special issue on blind identification and estimation*, vol. 10, pp. 2009–2025, October 1998.

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

[8] I. E. Frank and R. Todeschini, *The data analysis handbook*, Elsevier Science, Amsterdam, September 1994.

[9] J. Rosier and Y. Grenier, "Unsupervised classification techniques for multipitch estimation," in *116th Convention of the Audio Engineering Society*, Berlin, 2004.