A general modular framework for audio source separation

Alexey Ozerov, Emmanuel Vincent, Frédéric Bimbot

A General Modular Framework for Audio Source Separation

Alexey Ozerov[1]*, Emmanuel Vincent[1], and Frédéric Bimbot[2]

[1] INRIA, Rennes Bretagne Atlantique,
[2] IRISA, CNRS - UMR 6074, Campus de Beaulieu, 35042 Rennes cedex, France
{alexey.ozerov,emmanuel.vincent,frederic.bimbot}@irisa.fr

Abstract. Most of audio source separation methods are developed for a particular scenario characterized by the number of sources and channels and the characteristics of the sources and the mixing process. In this paper we introduce a general modular audio source separation framework based on a library of flexible source models that enable the incorporation of prior knowledge about the characteristics of each source. First, this framework generalizes several existing audio source separation methods, while bringing a common formulation for them. Second, it allows to imagine and implement new efficient methods that were not yet reported in the literature. We first introduce the framework by describing the flexible model, explaining its generality, and summarizing our modular implementation using a Generalized Expectation-Maximization algorithm. Finally, we illustrate the above-mentioned capabilities of the framework by applying it in several new and existing configurations to different source separation scenarios.

1 Introduction

Separating audio sources from multichannel mixtures is still challenging in most situations. The main difficulty is that audio source separation problems are usually mathematically ill-posed and to succeed one needs to incorporate additional knowledge about the mixing process and/or the source signals. Thus, efficient source separation methods are usually developed for a particular scenario characterized by *problem dimensionality* ((over)determined case, underdetermined case, and single-channel case), *mixing process characteristics* (synthetic instantaneous, anechoic, and convolutive mixtures, and live recorded mixtures), *source characteristics* (speech, singing voice, drums, bass, and noise (stationary or not, white or colored)). Moreover, there is often no common formulation describing methods applied for different scenarios, and this makes it difficult to reuse a method for a scenario it was not originally conceived for just by modifying some parameters.

The motivation of this work is to design a general audio source separation framework that can be easily applied to several separation scenarios just by

* This work was supported in part by the Quaero Programme, funded by OSEO.

selecting from a library of models a suitable model for each source incorporating *a priori* knowledge about that source. More precisely we wish such a framework to be

- *general*, i.e., generalizing existing methods and making it possible to combine them,
- *flexible*, allowing easy incorporation of the *a priori* information about a particular scenario considered,
- *modular*, allowing an implementation in terms of software blocks addressing the estimation of subsets of parameters.

To achieve the property of generality, we need to find some common formulation for methods we would like to generalize. Several recently proposed methods for source separation and/or characterization [12], [1], [7], [6], [5], [11], [10], [13], [3] (see also [14] and references therein) are based on the same zero-mean Gaussian model describing both the properties of the sources and of the mixing process, and only the global structure of Gaussian covariances differs from one method to another. These methods already cover several possible scenarios, including single-channel [6] or multichannel sources [11], instantaneous [7] or convolutive [11] mixtures of point [7] or diffuse [5] sources, and monophonic (e.g., speech [10]) or polyphonic sources (e.g., polyphonic music [6]). Moreover, a few of these methods have already been combined together, for example hidden Markov model (HMM) (a monophonic source model) and nonnegative matrix factorization (NMF) (a polyphonic source model) were combined in [10], NMF [6] was combined with point and diffuse source models in [11], [3]. We chose this local Gaussian model as the basis of our framework. To achieve flexibility, we leave the global structures of Gaussian covariances be specifiable in every particular case, allowing introduction of knowledge about every particular source and its mixing conditions. Thus, our framework generalizes all the above methods, and, thanks to its flexibility, it becomes applicable in many other scenarios one can imagine. We implement our framework using a Generalized Expectation-Maximization (GEM) algorithm, where the M-step is solved in a modular fashion by alternating between optimization of different parameter subsets.

Our approach is in line with the *library of components* by Cardoso and Martin [4] developed for the separation of components in astrophysical images. However, we consider advanced audio-specific structures (an overview of such structures can be found in [8], [13]) for source spectral power that generalize the structures considered in [4] that assume simply that source power is constant in some predefined region of time and space. In that sense our framework is more flexible than [4].

The rest of this paper is organized as follows. The audio source separation problem considered here is described in section 2. Section 3 is devoted to the presentation of the model at the heart of our framework, and some information about the employed GEM algorithm is given in section 4. The results of several source separation experiments are given in section 5 to illustrate the flexibility of our framework and the resulting performance improvement compared to individual approaches. Conclusions are drawn in section 6.

2 Audio Source Separation

Since we would like to address the separation of both point and diffuse sources, the standard point source based convolutive blind source separation (BSS) problem formulation (see e.g., [11], Eq. (1)) is not suitable in our case. Thus, we rather assume that the observed multichannel time-domain signal, called *mixture*, $\tilde{\mathbf{x}}(t) \in \mathbb{R}^I$ (I being the number of channels, and $t = 1, \ldots, T$) is a sum of J multichannel signals $\tilde{\mathbf{y}}_j(t) \in \mathbb{R}^I$, called *spatial source images* [4], [14]:

$$\tilde{\mathbf{x}}(t) = \sum\nolimits_{j=1}^{J} \tilde{\mathbf{y}}_j(t), \tag{1}$$

and the goal is to estimate the spatial source images $\tilde{\mathbf{y}}_j(t)$, given the mixture $\tilde{\mathbf{x}}(t)$.

Audio signals are usually processed in the time-frequency domain, due to sparsity property of audio signals in such a representation. Thus, we convert all the signals in the short-time Fourier transform (STFT) domain, and equation (1) becomes:

$$\mathbf{x}_{fn} = \sum\nolimits_{j=1}^{J} \mathbf{y}_{j,fn} \tag{2}$$

where $\mathbf{x}_{fn} \in \mathbb{C}^I$ and $\mathbf{y}_{j,fn} \in \mathbb{C}^I$ are I-dimensional complex-valued vectors of STFT coefficients of the corresponding time-domain signals; and $f = 1, \ldots, F$ and $n = 1, \ldots, N$ denote respectively STFT frequency and time indices.

3 Flexible Model

We assume that every vector $\mathbf{y}_{j,fn} \in \mathbb{C}^I$ is a proper complex-valued Gaussian random vector with zero mean and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{y},j,fn} = v_{j,fn}\mathbf{R}_{j,fn}$:

$$\mathbf{y}_{j,fn} \sim \mathcal{N}_c\left(\bar{0}, v_{j,fn}\mathbf{R}_{j,fn}\right) \tag{3}$$

where the matrix $\mathbf{R}_{j,fn} \in \mathbb{C}^{I\times I}$ called *spatial covariance matrix*, represents the spatial characteristics of the source and of the mixing setup, and the non-negative scalar $v_{j,fn} \in \mathbb{R}_+$ called *spectral power* represents the spectral characteristics of the source. Moreover, the random vectors $\mathbf{y}_{j,fn}$ are assumed to be independent given $\boldsymbol{\Sigma}_{\mathbf{y},j,fn}$. The model of the j-th source can be parametrized as $\theta_j = \{v_{j,fn}, \mathbf{R}_{j,fn}\}_{f,n=1}^{F,N}$, and the overall model writes $\theta = \{\theta_j\}_{j=1}^{J}$.

Given the model parameters θ, the sources can be estimated in the minimum mean square error (MMSE) sense via Wiener filtering:

$$\hat{\mathbf{y}}_{j,fn} = v_{j,fn}\mathbf{R}_{j,fn}\boldsymbol{\Sigma}_{\mathbf{x},fn}^{-1}(\theta)\mathbf{x}_{fn}, \tag{4}$$

where $\boldsymbol{\Sigma}_{\mathbf{x},fn}(\theta) \triangleq \sum_{j=1}^{J} v_{j,fn}\mathbf{R}_{j,fn}$.

The model parameters θ are usually not given and should be estimated. It is clear that estimating model parameters in the maximum likelihood (ML) sense would not lead to any consistent estimation, since there are more free parameters

in θ than data samples in $[\mathbf{x}_{fn}]_{f,n}$. We hence assume that θ belongs to a subset of admissible parameters Θ (*structural constraints*) and/or we consider that θ follows some *a priori* distribution $p(\theta|\eta)$, where η denotes some hyperparameters. With these assumptions we use the maximum *a posteriori* (MAP) criterion that can be rewritten as [11]:

$$\theta^*, \eta^* = \arg\min_{\theta\in\Theta,\eta} \sum_{f,n} \left[\mathrm{tr}\left(\boldsymbol{\Sigma}_{\mathbf{x},fn}^{-1}(\theta)\mathbf{x}_{fn}\mathbf{x}_{fn}^H\right) + \log|\boldsymbol{\Sigma}_{\mathbf{x},fn}(\theta)|\right] - \log p(\theta|\eta). \quad (5)$$

In this paper we focus on structural constraints (overviewed in the following sections) allowing the incorporation of additional knowledge about the mixing process and the audio source signals, and we leave aside the prior $p(\theta|\eta)$.

3.1 Spatial Covariance Structures

In this work we first assume that the spatial covariances are time invariant, i.e., $\mathbf{R}_{j,fn} = \mathbf{R}_{j,f}$. In the case of audio, it is mostly interesting to consider either rank-1 covariances representing instantaneously mixed (or convolutively mixed with weak reverberation) point sources (see e.g., [11]) or full rank covariances modeling diffuse or reverberated sources [5]. We assume in the rank-1 case that $\mathbf{R}_{j,f} = \mathbf{a}_{j,f}\mathbf{a}_{j,f}^H$, where $\mathbf{a}_{j,f} \in \mathbb{C}^I$ is a column vector, and in the full rank case that $\mathbf{R}_{j,f}$ is a positive definite Hermitian matrix.

Moreover, we assume that for every source j the spatial covariances $\{\mathbf{R}_{j,f}\}_f$ are either linear instantaneous (i.e., constant over frequency: $\mathbf{a}_{j,f} = \mathbf{a}_j$ or $\mathbf{R}_{j,f} = \mathbf{R}_j$) or convolutive (i.e., varying with frequency), and either fixed (i.e., not updated during model estimation) or adaptive.

3.2 Spectral Power Structures

To model spectral power we use non-negative tensor factorization (NTF)-like audio-specific decompositions [8], thus all variables introduced in this section are implicitly assumed to be non-negative. We first model spectral power $v_{j,fn}$ as the product of *excitation spectral power* $v_{j,fn}^{\mathrm{excit}}$ (e.g., representing the excitation of the glottal source for voice or the plucking of the string of a guitar) and *filter spectral power* $v_{j,fn}^{\mathrm{filt}}$ (e.g., representing the vocal tract or the impedance of the guitar body) [8]:

$$v_{j,fn} = v_{j,fn}^{\mathrm{excit}} \times v_{j,fn}^{\mathrm{filt}} \quad (6)$$

The excitation spectral power $[v_{j,fn}^{\mathrm{excit}}]_f$ is modeled as the sum of K_{excit} characteristic spectral patterns $[e_{j,fk}^{\mathrm{excit}}]_f$ modulated in time by $p_{j,kn}^{\mathrm{excit}}$, i.e., $v_{j,fn}^{\mathrm{excit}} = \sum_{k=1}^{K_{\mathrm{excit}}} p_{j,kn}^{\mathrm{excit}} e_{j,fk}^{\mathrm{excit}}$ [6]. In order to further constrain the fine spectral structure of the spectral patterns, they can be represented as linear combinations of L_{excit} *elementary narrowband spectral patterns* $[w_{j,fl}^{\mathrm{excit}}]_f$ [13], i.e., $e_{j,fk}^{\mathrm{excit}} = \sum_{l=1}^{L_{\mathrm{excit}}} u_{j,lk}^{\mathrm{excit}} w_{j,fl}^{\mathrm{excit}}$, where $u_{j,lk}^{\mathrm{excit}}$ are some non-negative weights. These narrowband patterns may be for instance harmonic, inharmonic or noise-like with a smooth spectral envelope.

Following exactly the same idea, we propose to represent the series of time activation coefficients $p_{j,kn}^{\text{excit}}$ as sums of M_{excit} *time localized patterns* to ensure their continuity or some other structure, i.e., $p_{j,kn}^{\text{excit}} = \sum_{m=1}^{M_{\text{excit}}} h_{j,mn}^{\text{excit}} g_{j,km}^{\text{excit}}$. Altogether we have:

$$v_{j,fn}^{\text{excit}} = \sum\nolimits_{k=1}^{K_{\text{excit}}} \sum\nolimits_{m=1}^{M_{\text{excit}}} h_{j,mn}^{\text{excit}} g_{j,km}^{\text{excit}} \sum\nolimits_{l=1}^{L_{\text{excit}}} u_{j,lk}^{\text{excit}} w_{j,fl}^{\text{excit}}, \tag{7}$$

and, introducing matrices $\mathbf{V}_j^{\text{excit}} \triangleq [v_{j,fn}^{\text{excit}}]_{f,n}$, $\mathbf{H}_j^{\text{excit}} \triangleq [h_{j,mn}^{\text{excit}}]_{m,n}$, $\mathbf{G}_j^{\text{excit}} \triangleq [g_{j,km}^{\text{excit}}]_{k,m}$, $\mathbf{U}_j^{\text{excit}} \triangleq [u_{j,lk}^{\text{excit}}]_{l,k}$ and $\mathbf{W}_j^{\text{excit}} \triangleq [w_{j,fl}^{\text{excit}}]_{f,l}$, this equation can be rewritten in matrix form as $\mathbf{V}_j^{\text{excit}} = \mathbf{W}_j^{\text{excit}} \mathbf{U}_j^{\text{excit}} \mathbf{G}_j^{\text{excit}} \mathbf{H}_j^{\text{excit}}$.

Filter spectral power $[v_{j,fn}^{\text{filt}}]_f$ is represented with exactly the same structure as (7), so that to allow modeling time-varying filters as linear combination of some characteristic spectral patterns $[e_{j,fk}^{\text{filt}}]_f$ constrained to be continuous using some smooth narrowband elementary spectral patterns $[w_{j,fl}^{\text{filt}}]_f$.

Altogether spectral power structure can be represented by the following matrix decomposition ($\odot$ denotes element-wise matrix multiplication):

$$\mathbf{V}_j = \mathbf{V}_j^{\text{excit}} \odot \mathbf{V}_j^{\text{filt}} = \left(\mathbf{W}_j^{\text{excit}} \mathbf{U}_j^{\text{excit}} \mathbf{G}_j^{\text{excit}} \mathbf{H}_j^{\text{excit}}\right) \odot \left(\mathbf{W}_j^{\text{filt}} \mathbf{U}_j^{\text{filt}} \mathbf{G}^{\text{filt}} \mathbf{H}_j^{\text{filt}}\right), \tag{8}$$

where each matrix in this decomposition is assumed to be either fixed or adaptive. To cover Gaussian mixture models (GMM), HMM, and scaled versions of these models (SGMM, HSMM) [10], every column $\mathbf{g}_{j,m}^{\text{excit}} = [g_{j,km}^{\text{excit}}]_k$ of matrix $\mathbf{G}_j^{\text{excit}}$ (and similarly for matrix $\mathbf{G}_j^{\text{filt}}$) may further be constrained to have either a single nonzero entry (for SGMM, HSMM) or a single nonzero entry equal to 1 (for GMM, HMM). Mixture component probabilities of GMM and transition probabilities for HMM should be included in hyperparameters η of (5).

3.3 Generality

It can be easily shown that the model structures considered in [12], [1], [7], [6], [5], [11], [10], [13], [3], [14] are particular instances of the proposed general formulation. Let us give some examples.

Pham *et al* [12] assume rank-1 spatial covariances and constant spectral power over time-frequency regions of size (1 frequency bin $\times$ L frames). This structure can be implemented in our framework by choosing rank-1 adaptive spatial covariances and constraining spectral power to $\mathbf{V}_j = \mathbf{W}_j^{\text{excit}} \mathbf{G}^{\text{excit}} \mathbf{H}_j^{\text{excit}}$ [3] with $\mathbf{W}_j^{\text{excit}}$ being the identity ($F \times F$) matrix, $\mathbf{G}^{\text{excit}}$ being ($F \times \lceil N/L \rceil$) adaptive, and $\mathbf{H}_j^{\text{excit}}$ being ($\lceil N/L \rceil \times N$) fixed with entries $h_{j,mn}^{\text{excit}} = 1$ for $n \in \mathcal{L}_m$ and $h_{j,mn}^{\text{excit}} = 0$ for $n \notin \mathcal{L}_m$, where $\mathcal{L}_m$ is the set of time indices of the L-length block.

Multichannel NMF structures with point [11] or diffuse source model [3] can be represented within our framework as $\mathbf{V}_j = \mathbf{W}_j^{\text{excit}} \mathbf{H}_j^{\text{excit}}$ with $\mathbf{W}_j^{\text{excit}}$ of size ($F \times K_{\text{excit}}$) and $\mathbf{H}_j^{\text{excit}}$ of size ($K_{\text{excit}} \times F$), both being adaptive, and rank-1 or full rank adaptive spatial covariances.

[3] The power structure in (8) can be easily reduced to $\mathbf{V}_j = \mathbf{W}_j^{\text{excit}} \mathbf{G}^{\text{excit}} \mathbf{H}_j^{\text{excit}}$ by assuming that all the other matrices $\mathbf{U}_j^{\text{excit}}$, $\mathbf{W}_j^{\text{filt}}$, $\mathbf{U}_j^{\text{filt}}$, $\mathbf{G}^{\text{filt}}$ and $\mathbf{H}_j^{\text{filt}}$ are of sizes ($1 \times K_{\text{excit}}$), ($F \times 1$), ($1 \times 1$), ($1 \times 1$), and ($1 \times N$), fixed, and composed of 1.

4 Modular Implementation

Due to lack of space we here give a very brief overview of the framework implementation via a GEM algorithm that consists in iterating the expectation (E) and maximization (M) steps. The E-step consists in computing conditional expectation $\widehat{\mathbf{T}}$ of a *natural sufficient statistics* $\mathbf{T}$, given the observations $\mathbf{X} = \{\mathbf{x}_{fn}\}_{f,n}$ and current model parameters. The M-step consists in updating model parameters θ so as to increase the conditional expectation of the log-likelihood of the *complete data*. We assume that the J-th source with full rank spatial covariance represents a controllable additive noise needed for *simulated annealing* as in [11]. Let $\mathcal{J}_{\mathrm{r1}}$ and $\mathcal{J}_{\mathrm{rf}}$ be the subsets of the remaining source indices $\{1,\ldots,J-1\}$ corresponding respectively to rank-1 and full rank spatial covariances, and we assume that each source with rank-1 spatial covariance $\mathbf{R}_{j,f} = \mathbf{a}_{j,f}\mathbf{a}_{j,f}^H$ writes $\mathbf{y}_{j,fn} = \mathbf{a}_{j,f}s_{j,fn}$, where $s_{j,fn}$ are the STFT coefficients of a single-channel signal. With these conventions we choose $\mathbf{Z} = \{\mathbf{x}_{fn}, \{\mathbf{y}_{j,fn}\}_{j\in\mathcal{J}_{\mathrm{rf}}}, \{s_{j,fn}\}_{j\in\mathcal{J}_{\mathrm{r1}}}\}_{f,n}$ as the complete data set of the proposed GEM algorithm. The model $\theta = \{\theta_j\}_{j=1}^J$ being a set of source models, each source model is further represented as a set of 9 parameter subsets $\theta_j = \{\theta_j^m\}_{m=1}^9 = \{\mathbf{R}_j, \mathbf{W}_j^{\mathrm{excit}}, \mathbf{U}_j^{\mathrm{excit}}, \mathbf{G}_j^{\mathrm{excit}}, \mathbf{H}_j^{\mathrm{excit}}, \mathbf{W}_j^{\mathrm{filt}}, \mathbf{U}_j^{\mathrm{filt}}, \mathbf{G}_j^{\mathrm{filt}}, \mathbf{H}_j^{\mathrm{filt}}\}$. We implement the M-step via a loop over all $J \times 9$ parameter subsets. Each subset, depending whether it is adaptive or fixed, is updated or not in turn using existing spatial covariance update rules [11], [5] and multiplicative NMF update rules [13] that guarantee that the log-likelihood of the complete data is non-decreasing. Finally, particular constraints (see Sec. 3.1 and 3.2) are applied, if specified.

5 Experimental Illustrations

To illustrate the flexibility, we have evaluated four instances of our framework on the development data of the second community-based Signal Separation Evaluation Campaign (SiSEC 2010) [4] "Underdetermined-speech and music mixtures" task. The former two instances considered are NMF spectral power structures with rank-l [11] and full rank [3] spatial covariances (see Sec. 3.3). The later two instances are similar, except that the spectral power is structured as $\mathbf{V}_j = \mathbf{W}_j^{\mathrm{excit}}\,\mathbf{U}_j^{\mathrm{excit}}\,\mathbf{H}_j^{\mathrm{excit}}$ with adaptive $\mathbf{W}_j^{\mathrm{excit}}$ and $\mathbf{H}_j^{\mathrm{excit}}$ of sizes $(F \times L_{\mathrm{excit}})$ and $(K_{\mathrm{excit}} \times F)$, and fixed $\mathbf{U}_j^{\mathrm{excit}}$ of size $(K_{\mathrm{excit}} \times F)$ being composed of harmonic (e.g., to represent voiced speech) and noise-like and smooth (e.g., to represent non-voiced speech) narrowband spectral patterns [13]. Such a spectral structure is simply referred hereafter as *harmonic NMF*. In line with [11], parameter estimation via GEM is very sensitive to initialization for all the configurations we consider. To provide our GEM algorithm with a "good initialization" we used the DEMIX mixing matrix estimation algorithm [2], followed by l_0 norm minimization (see e.g., [14]) to initialize the source spectra, for the instantaneous mixtures. For synthetic convolutive and live recorded mixtures we used Cumulative

[4] `http://sisec.wiki.irisa.fr/tiki-index.php`

State Coherence (CSC) transform-based Time Differences Of Arrival (TDOAs) estimation algorithm [9] to initialize anechoic spatial covariances, followed by binary masking to initialize the source spectra. Source separation results in terms of average Source to Distortion Ratio (SDR) after 200 iterations of the proposed GEM algorithm are summarized in table 1 together with results of the *baseline* used for initialization. As expected, rank-1 spatial covariances perform the best for instantaneous mixtures and full rank spatial covariances perform the best for synthetic convolutive and live recorded mixtures. Moreover, as compared to the NMF spectral power, the harmonic NMF spectral power improves results for speech sources in almost all cases. Thus, we see that each tested configuration is performs the best for some setting. For each setting the configuration preforming the best on the development data was entered to the SiSEC 2010.

Table 1. Average SDRs on subsets of SiSEC 2010 development data.

Mixing	instantaneous		synth. convolutif				live recorded			
Sources	speech	music	speech		music		speech		music	
Microphone distance	-	-	5 cm	1 m	5 cm	1 m	5 cm	1 m	5 cm	1 m
baseline (l_0 min. or bin. mask.)	8.6	12.4	0.3	1.4	-0.8	-0.9	1.0	1.4	2.3	0.0
NMF / rank-1 [11]	9.6	**18.4**	1.0	2.3	-0.6	-0.6	2.0	2.4	**3.6**	0.3
NMF / full-rank [3]	8.7	17.9	1.2	2.9	-2.3	-0.5	2.2	2.9	3.3	**0.7**
harmonic NMF / rank-1	**10.6**	15.1	1.0	2.7	**-0.1**	**0.0**	2.2	3.4	2.2	0.6
harmonic NMF / full-rank	10.5	14.3	**1.5**	**3.5**	-1.8	-0.2	**2.5**	**3.9**	1.5	0.4

6 Conclusion

We have introduced a general flexible and modular audio source separation framework that generalizes several existing source separation methods, brings them into a common framework, and allows to imagine and implement new efficient methods. The framework capabilities were illustrated in the experimental part, where we have reproduced two existing methods, namely NMF / rank-1 [11] and NMF / full-rank [3], but also we have tested two new methods, namely harmonic NMF / rank-1 and harmonic NMF / full-rank, that in our best knowledge were not yet reported in the literature. We have observed that adding harmonic and noise-like smooth constraints to NMF allows improving separation results for speech signals. Note also that the proposed framework can also be seen as a statistical implementation of Computational Auditory Scene Analysis (CASA) principles, whereby primitive grouping cues and learned grouping cues are simultaneously used to segregate the sources, thereby avoiding error propagation due to sequential use of grouping cues. Examples primitive grouping cues accounted by our model include harmonicity, spectral smoothness, time continuity, common onset, common amplitude modulation, spectral similarity and spatial similarity.

Acknowledgments

The authors would like to thank S. Arberet and F. Nesta for kindly sharing their implementations of DEMIX [2] and a TDOAs estimation [9] algorithms.

References

1. Abdallah, S.A., Plumbley, M.D.: Polyphonic transcription by nonnegative sparse coding of power spectra. In: Proc. 5th International Symposium Music Information Retrieval (ISMIR'04). pp. 318–325 (Oct 2004)
2. Arberet, S., Gribonval, R., Bimbot, F.: A robust method to count and locate audio sources in a multichannel underdetermined mixture. Signal Processing, IEEE Transactions on 58(1), 121 –133 (jan 2010)
3. Arberet, S., Ozerov, A., Duong, N., Vincent, E., Gribonval, R., Bimbot, F., Vandergheynst, P.: Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation. In: 10th Int. Conf. on Information Sciences, Signal Proc. and their applications (ISSPA'10) (2010)
4. Cardoso, J.F., Martin, M.: A flexible component model for precision ICA. In: Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA'07). pp. 1–8 (2007)
5. Duong, N.Q.K., Vincent, E., Gribonval, R.: Under-determined convolutive blind source separation using spatial covariance models. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (Mar 2010)
6. Févotte, C., Bertin, N., Durrieu, J.L.: Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. Neural Computation 21(3), 793–830 (Mar 2009)
7. Févotte, C., Cardoso, J.F.: Maximum likelihood approach for blind audio source separation using time-frequency Gaussian models. In: WASPAA'05. Mohonk, NY, USA (Oct 2005)
8. FitzGerald, D., Cranitch, M., Coyle, E.: Extended nonnegative tensor factorisation models for musical sound source separation. Computational Intelligence and Neuroscience. Hindawi Publishing Corp 2008 (2008)
9. Nesta, F., Svaizer, P., Omologo, M.: Cumulative state coherence transform for a robust two-channel multiple source localization. In: Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA'09) (2009)
10. Ozerov, A., Févotte, C.: Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. IEEE Trans. on Audio, Speech and Lang. Proc. 18(3), 550–563 (March 2010)
11. Ozerov, A., Févotte, C., Charbit, M.: Factorial scaled hidden Markov model for polyphonic audio representation and source separation. In: WASPAA '09. pp. 121–124 (Oct 18–21, 2009)
12. Pham, D.T., Servière, C., Boumaraf, H.: Blind separation of speech mixtures based on nonstationarity. In: Proceedings of the 7th International Symposium on Signal Processing and its Applications. pp. II–73–76 (2003)
13. Vincent, E., Bertin, N., Badeau, R.: Adaptive harmonic spectral decomposition for multiple pitch estimation. IEEE Trans. on Audio, Speech and Language Processing 18(3), 528–537 (2010)
14. Vincent, E., Jafari, M., Abdallah, S.A., Plumbley, M.D., Davies, M.E.: Probabilistic modeling paradigms for audio source separation. In: Machine Audition: Principles, Algorithms and Systems. IGI Global (2010), to appear