Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation

Simon Arberet, Alexey Ozerov, Ngoc Duong, Emmanuel Vincent, Rémi Gribonval, Frédéric Bimbot, Pierre Vandergheynst

NONNEGATIVE MATRIX FACTORIZATION AND SPATIAL COVARIANCE MODEL FOR UNDER-DETERMINED REVERBERANT AUDIO SOURCE SEPARATION

Simon Arberet[1], *Alexey Ozerov*[2], *Ngoc Q.K. Duong*[2], *Emmanuel Vincent*[2], *Rémi Gribonval*[2], *Frédéric Bimbot*[3], *Pierre Vandergheynst*[1*]

[1] Signal Processing Institute, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
{simon.arberet, pierre.vandergheynst}@epfl.ch
[2] INRIA, Centre de Rennes - Bretagne Atlantique, France
{alexey.ozerov, Quang-Khanh-Ngoc.Duong, emmanuel.vincent, remi.gribonval}@inria.fr
[3] CNRS, IRISA - UMR 6074, France
frederic.bimbot@irisa.fr

ABSTRACT

We address the problem of blind audio source separation in the under-determined and convolutive case. The contribution of each source to the mixture channels in the time-frequency domain is modeled by a zero-mean Gaussian random vector with a full rank covariance matrix composed of two terms: a variance which represents the spectral properties of the source and which is modeled by a nonnegative matrix factorization (NMF) model and another full rank covariance matrix which encodes the spatial properties of the source contribution in the mixture. We address the estimation of these parameters by maximizing the likelihood of the mixture using an expectation-maximization (EM) algorithm. Theoretical propositions are corroborated by experimental studies on stereo reverberant music mixtures.

1. INTRODUCTION

In blind source separation (BSS), we observe a multichannel signal $\mathbf{x}(t) \in \mathbb{R}^M$ which is a mixture of N source signals $s_n(t) \in \mathbb{R}, 1 \leq n \leq N$. In the convolutive BSS case, each source $s_n(t)$ is convolved with M filters $\mathbf{h}_n(l) \in \mathbb{R}^M$ which model in the audio context the acoustic paths from source n to the M microphones. The mixture process can be expressed as:

$$\mathbf{x}(t) = \sum_{n=1}^{N} \mathbf{y}_n(t) + \mathbf{n}(t) \tag{1}$$

where $\mathbf{n}(t) \in \mathbb{R}^M$ is an additive noise and $\mathbf{y}_n(t) \in \mathbb{R}^M$ is the spatial image of source n which is expressed as:

$$\mathbf{y}_n(t) = \sum_{l=0}^{L-1} \mathbf{h}_n(l) s_n(t-l) \tag{2}$$

where L is the filter length. The BSS problem consists in recovering either the source signals $s_n(t)$ or their spatial images $\mathbf{y}_n(t)$ given the mixture signal $\mathbf{x}(t)$. In this paper we consider the later BSS problem formulation. When the number N of sources is larger than the number M of mixture channels, the mixture is said under-determined.

The BSS problem is often addressed in the time-frequency (TF) domain via the short time Fourier transform (STFT), and the convolutive mixing process is approximated by a complex-valued instantaneous mixing in each frequency bin. In other words each source spatial image in the STFT domain $\mathbf{Y}_n(t,f)$ is approximated by the following complex-valued multiplication:

$$\mathbf{Y}_n(t,f) \approx \widehat{\mathbf{h}}_n(f) S_n(t,f) \tag{3}$$

where $\widehat{\mathbf{h}}_n(f)$ is the Fourier transform of the mixing filters $\mathbf{h}_n(t)$ and $S_n(t,f)$ and $\mathbf{Y}_n(t,f)$ are respectively the STFT of $s_n(t)$ and $\mathbf{y}_n(t)$. Thus, according to the model (3), if $S_n(t,f)$ is a zero-mean random variable with variance $v_n(t,f)$, the covariance of $\mathbf{Y}_n(t,f)$ is given by:

$$\mathbf{R}_{y_n}(t,f) = v_n(t,f)\mathbf{R}_n(f) \tag{4}$$

where $\mathbf{R}_n(f) = \widehat{\mathbf{h}}_n(f)\widehat{\mathbf{h}}_n^H(f)$ (H denotes the matrix conjugate transposition) is a rank-1 matrix. This rank-1 model holds only when the filter length L is short compared to the STFT window size [1]. A particular case is when the mixture is instantaneous, *i.e.* the filters $\mathbf{h}_n(l)$ have length $L = 1$, then approximation (3) becomes an equality. However in an environment with realistic reverberation time, the filter length L is usually longer than the STFT window size.

Assuming the rank-1 model (3), BSS can be achieved by estimating a mixing matrix [2, 3, 4] in each frequency bin f (whose columns are the vectors $\widehat{\mathbf{h}}_n(f), 1 \leq n \leq N$) and then recovering the source coefficients $S_n(t,f)$ assuming independence of the sources and some sparse prior distributions [5]. However, if the mixing matrices are estimated independently in each frequency bin, the columns $\widehat{\mathbf{h}}_n(f)$ of these mixing matrices are arbitrary permuted in each frequency, leading to the well-known permutation problem. Recently, some *spectral* approaches using the rank-1 model (3) and modeling the structure of the source variances $v_n(t,f)$ in the TF plane, with a Gaussian mixture model (GMM) [6] or nonnegative matrix factorisation (NMF) [7, 8] have been proposed. These spectral approaches have shown to provide better performance [9] than classical sparse approaches like binary masking [2], l_1-norm [10] or l_p-norm [11] minimization. In the case of the NMF approach of [8], as there is a coupling of the frequency bins due to the structure of $v_n(t,f)$, and as $v_n(t,f)$ and $\widehat{\mathbf{h}}_n(f)$ are jointly estimated, we are able to avoid the permutation problem.

To model reverberation efficiently, Duong et al.[12] proposed recently to consider $\mathbf{R}_n(f)$ as a full-rank (unconstrainted) matrix. They showed that this model led to better results than the rank-1 model on reverberant mixtures in *oracle* context where

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

$\mathbf{R}_n(f)$ and $v_n(t,f)$ are known, in *semi-blind* context where $\mathbf{R}_n(f)$ is known but $v_n(t,f)$ is a free variance estimated from the mixture. They also formulated an expectation-maximization (EM) algorithm [13] to blindly estimate $\mathbf{R}_n(f)$ and $v_n(t,f)$ in each frequency bin. However as the parameters $\mathbf{R}_n(f)$ and $v_n(t,f)$ are estimated independently in each frequency bin, the permutation problem has to be solved a posteriori. Duong et al. [13] applied a DOA-based algorithm to solve the permutation problem. However, in order to deploy this DOA-based algorithm, it is imperative to know the inter-microphone distance beforehand.

Motivated by the effectiveness of the full rank *spatial* model of Duong et al.[13] and the NMF *spectral* model [7, 8], we investigate in this paper the modeling of each spatial source image with a combination of these two models. We describe the proposed source spatial image model in Section 2. Section 3 addresses the proposed inference method which consists of maximizing the likelihood of the mixture data using an EM algorithm [14]. In Section 4, we compare the source separation performance achieved by our full-rank NMF method with the rank-1 NMF method [8] and with other state-of-the-art algorithms over stereo music data. Finally, we conclude in Section 5.

2. MODEL

2.1. Source spatial image

We assume that each spatial source image $\mathbf{Y}_n(t,f)$ at TF point (t,f) is a zero-mean complex random vector:

$$\mathbf{Y}_n(t,f) \sim \mathcal{N}_c(0, \mathbf{R}_{y_n}(t,f)), \tag{5}$$

where $\mathcal{N}_c(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a proper complex distribution with probability density function (pdf):

$$N_c(\mathbf{Y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \triangleq |\pi\, \boldsymbol{\Sigma}|^{-1} \exp\left[-(\mathbf{Y}-\boldsymbol{\mu})^H \boldsymbol{\Sigma}^{-1} (\mathbf{Y}-\boldsymbol{\mu}) \right],$$

where $|\mathbf{A}|$ denotes the determinant of a square matrix $\mathbf{A}$. $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are respectively, the M-dimensional mean vector and the $M \times M$ covariance matrix of $\mathbf{Y}$. Covariance matrix $\mathbf{R}_{y_n}(t,f)$ is given by (4), where $\mathbf{R}_n(f)$ is a full-rank unconstrained time-invariant covariance matrix which encodes the spatial properties of the source [13] and $v_n(t,f)$ is a time-varying source variance which is an assumed sum of K components:

$$v_n(t,f) = \sum_{k=1}^{K} w_{f,k}^n h_{k,t}^n \tag{6}$$

where $w_{f,k}^n, h_{k,t}^n \in \mathbb{R}^+$. Thus, the power spectrum $\mathbf{V}_n = [v_n(t,f)]_{f,t}$ of each source n is structured as a product of two nonnegative matrices $\mathbf{W}_n = [w_{f,k}^n]_{f,k}$ and $\mathbf{H}_n = [h_{k,t}^n]_{k,t}$:

$$\mathbf{V}_n = \mathbf{W}_n \mathbf{H}_n.$$

According to (4), (5), (6), each source spatial image $\mathbf{Y}_n(t,f)$ can be seen as a sum of K independant zero-mean Gaussians $\mathbf{Y}_n(t,f) = \sum_{k=1}^K \mathbf{Y}_{n,k}(t,f)$ with the respective covariances:

$$\mathbf{R}_{y_{n,k}}(t,f) = w_{f,k}^n h_{k,t}^n \mathbf{R}_n(f). \tag{7}$$

2.2. Noise

Let $\mathbf{N}(t,f)$ be the STFT of $\mathbf{n}(t)$. We assume that the noise is a stationary zero-mean Gaussian process with covariance $\mathbf{R}_b(f)$:

$$\mathbf{N}(t,f) \sim \mathcal{N}_c(0, \mathbf{R}_b(f)). \tag{8}$$

Thus the noise can be seen as a particular source with only one component which is time-invariant. In other words relation (4) applied to the noise "source" becomes $\mathbf{R}_b(t,f) = \mathbf{R}_b(f)$.

2.3. Mixture

Since the STFT is a linear transform, the mixing process (1) can be rewritten as:

$$\mathbf{X}(t,f) = \sum_n \mathbf{Y}_n(t,f) + \mathbf{N}(t,f).$$

The sources and noise are assumed to be independent of each other. Thus, the model of the mixture STFT $\mathbf{X}(t,f)$ is a zero-mean Gaussian vector with covariance matrix:

$$\mathbf{R}_{\mathbf{x}}(t,f) = \sum_n \mathbf{R}_{y_n}(t,f) + \mathbf{R}_b(f). \tag{9}$$

3. ESTIMATION OF THE MODEL PARAMETERS

We wish to estimate in the maximum likelihood (ML) sense, the mixing parameters $\mathbf{R}_n(f)$ of each source, the source variances $v_n(t,f)$ under the constraint given by (6) and the noise covariance $\mathbf{R}_b(f)$.

3.1. Criterion and indeterminacies

Let $\boldsymbol{\Theta} = \{\mathbf{R}_n(f), \mathbf{W}_n, \mathbf{H}_n, \mathbf{R}_b(f),\ \forall f, n\}$ be the set of all the parameters we wish to estimate. As $P(\mathbf{X}(t,f)|\boldsymbol{\Theta})$ is a zero-mean Gaussian according to section 2.3, maximizing the log-likelihood $\sum_{t,f} \log P(\mathbf{X}(t,f)|\boldsymbol{\Theta})$ is equivalent to minimizing the cost:

$$C(\boldsymbol{\Theta}) = \sum_{t,f} \mathbf{X}(t,f)^H \mathbf{R}_{\mathbf{x}}^{-1}(t,f) \mathbf{X}(t,f) + \log |\mathbf{R}_{\mathbf{x}}(t,f)|$$

where $\mathbf{R}_{\mathbf{x}}(t,f)$ is defined in (9). Thus, the ML criterion suffers from scaling indeterminacies because for any $\alpha \in \mathbb{R}^{+,*}$: $\mathbf{R}_{y_n}(t,f) = (\alpha v_n(t,f)) \left(\frac{1}{\alpha}\mathbf{R}_n(f)\right)$, and also for any $\alpha_k \in \mathbb{R}^{+,*}$: $v_n(t,f) = \sum_k (\alpha_k w_{f,k}^n) \left(\frac{1}{\alpha_k} h_{k,t}^n\right)$. In order to remove these scaling indeterminacies, we normalize $\mathbf{R}_n(f)$ according to the Frobenius norm $\|\mathbf{R}_n(f)\|_F = 1$ (and scale $v_n(t,f)$ accordingly) and impose the condition, $\sum_f w_{f,k}^n = 1$ (and scaling $h_{k,t}^n$ accordingly) as in [8].

3.2. Algorithm

We derive an EM algorithm [14] based on the *complete data* $\{\mathbf{Y}_{n,k}(t,f), \mathbf{N}(t,f)\ \forall t,f,n,k\}$, that is the set of the STFT coefficients of all the source spatial image components and the noise. Each iteration of the EM algorithm is composed of two steps: the E-step and the M-step. The E-step consists of computing the expectation of the natural statistics $\widehat{\mathbf{R}}_{y_n}(t,f), \widehat{\mathbf{R}}_{y_{n,k}}(t,f), \widehat{\mathbf{R}}_b(t,f)$, that is, the covariances of $\mathbf{Y}_n(t,f)$, $\mathbf{Y}_{n,k}(t,f)$ and $\mathbf{N}(t,f)$, conditionally on the mixture data and the current parameter estimates $\boldsymbol{\Theta}$. The M-step consists in re-estimating the parameters $\boldsymbol{\Theta}$ using the updated natural statistics.

3.2.1. ***E-step***: *Conditional expectation of natural statistics*

$$\widehat{\mathbf{R}}_{y_n}(t,f) = \widehat{\mathbf{Y}}_n(t,f)\widehat{\mathbf{Y}}_n^H(t,f) + (\mathbf{I} - \mathbf{G}_n(t,f))\, \mathbf{R}_{y_n}(t,f) \tag{10}$$

$$\widehat{\mathbf{R}}_{y_{n,k}}(t,f) = \widehat{\mathbf{Y}}_{n,k}(t,f)\widehat{\mathbf{Y}}_{n,k}^H(t,f) + (\mathbf{I} - \mathbf{G}_{n,k}(t,f))\, \mathbf{R}_{y_{n,k}}(t,f) \tag{11}$$

$$\widehat{\mathbf{R}}_b(t,f) = \widehat{\mathbf{N}}(t,f)\widehat{\mathbf{N}}^H(t,f) + (\mathbf{I} - \mathbf{G}_b(t,f))\, \mathbf{R}_b(f)$$

where

$$\widehat{\mathbf{Y}}_n(t,f) = \mathbf{G}_n(t,f)\mathbf{X}(t,f) \quad (12)$$
$$\widehat{\mathbf{Y}}_{n,k}(t,f) = \mathbf{G}_{n,k}(t,f)\mathbf{X}(t,f)$$
$$\widehat{\mathbf{N}}(t,f) = \mathbf{G}_b(t,f)\mathbf{X}(t,f)$$

$$\mathbf{G}_n(t,f) = \mathbf{R}_{y_n}(t,f)\left(\mathbf{R}_{\mathbf{x}}(t,f)\right)^{-1}$$
$$\mathbf{G}_{n,k}(t,f) = \mathbf{R}_{y_{n,k}}(t,f)\left(\mathbf{R}_{\mathbf{x}}(t,f)\right)^{-1}$$
$$\mathbf{G}_b(t,f) = \mathbf{R}_b(f)\left(\mathbf{R}_{\mathbf{x}}(t,f)\right)^{-1}.$$

3.2.2. ***M-step**: Update of the parameters*

The re-estimation of $\mathbf{R}_n(f)$ in the ML sense is equivalent to minimizing the sum over all the time frames t of the Kullback-Leibler (KL) divergence $D_{\text{KL}}\left(\widehat{\mathbf{R}}_{y_n}(t,f)\middle|\,\mathbf{R}_{y_n}(t,f)\right)$, with respect to (w.r.t.) $\mathbf{R}_n(f)$, between two zero-mean Gaussian distributions with covariances matrices $\widehat{\mathbf{R}}_{y_n}(t,f)$ and $\mathbf{R}_{y_n}(t,f)$ defined in (10) and (4) respectively and with:

$$D_{\text{KL}}\left(\mathbf{R}_1|\,\mathbf{R}_2\right) = \frac{1}{2}\left(\text{tr}\left(\mathbf{R}_1\mathbf{R}_2^{-1}\right) - \log\det\left(\mathbf{R}_1\mathbf{R}_2^{-1}\right) - M\right).$$

Given $\mathbf{W}_n$ and $\mathbf{H}_n$, this minimization has a closed-form representation which is [13]

$$\mathbf{R}_n(f) = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{v_n(t,f)}\widehat{\mathbf{R}}_{y_n}(t,f) \quad (13)$$

where $v_n(t,f) = \sum_{k=1}^{K} w_{f,k}^n h_{k,t}^n$ as defined in (6).

The re-estimation of $w_{f,k}^n$ and $h_{k,t}^n$ in the ML sense is equivalent to minimizing $\sum_{t,f} D_{\text{KL}}\left(\widehat{\mathbf{R}}_{y_{n,k}}(t,f)\middle|\,\mathbf{R}_{y_{n,k}}(t,f)\right)$ w.r.t. $w_{f,k}^n$ and $h_{k,t}^n$, where $\widehat{\mathbf{R}}_{y_{n,k}}(t,f)$ and $\mathbf{R}_{y_{n,k}}(t,f)$ are defined in (11) and (7) respectively. Given $\mathbf{R}_n(f)$, these minimizations have closed-form representations which are:

$$w_{f,k}^n = \frac{1}{T}\sum_{t=1}^{T}\frac{\hat{v}_{n,k}(t,f)}{h_{k,t}^n}, \quad h_{k,t}^n = \frac{1}{F}\sum_{f=1}^{F}\frac{\hat{v}_{n,k}(t,f)}{w_{f,k}^n} \quad (14)$$

with:

$$\hat{v}_{n,k}(t,f) = \frac{1}{M}\,\text{tr}\left(\mathbf{R}_n^{-1}(f)\widehat{\mathbf{R}}_{y_{n,k}}(t,f)\right) \quad (15)$$

where tr(.) denotes the trace of a square matrix.

To remove the scaling ambiguity between the components $w_{f,k}^n$ and $h_{k,t}^n$, we normalize them as explained in section 3.1.

The process of re-estimating of the noise covariance involves a similar set of steps as in (13) and is given by:

$$\mathbf{R}_b(f) = \frac{1}{T}\sum_{t=1}^{T}\widehat{\mathbf{R}}_b(t,f).$$

Assuming that the noise is spatially uncorrelated, we set the off-diagonal coefficients of $\mathbf{R}_b(f)$ to zero.

3.3. Estimation of the sources

After the convergence of the EM algorithm, the source spatial images are estimated in the TF domain with the Wiener estimator as in (12). The estimation of source spatial images in the time domain are then obtained via inversion of the STFT map using the overlappadd technique.

4. EXPERIMENTAL RESULTS

4.1. Setting

4.1.1. *Datasets*

We evaluated our NMF full rank algorithm with the NMF rank-1 algorithm of Ozerov and Févotte [8] and the rank-1 and full rank EM algorithms of Duong et al [13], over stereo noiseless music mixtures under various mixing conditions. For each experiment 10 mixtures were generated by convolving different 10 seconds source signals sampled at 16 kHz with room impulse responses obtained with the Roomsim toolbox [15]. The microphones are omnidirectional and the room dimensions are 4.45 m x 3.55 m x 2.5 m. The number of sources is set to either 3 or 4, the reverberation time (RT) to either 130 ms or 250 ms and the distance between the two microphones to either 5 cm or 1 m, resulting in 8 mixing conditions overall.

4.1.2. *Algorithms setting and evaluation criterion*

The STFT was computed with a sine window of length 1024. The number of components per source of the NMF models was set to $K = 5$, the number of iterations for each EM algorithm was 50. Separation performance was evaluated using the signal-to-distortion ratio (SDR) criterion averaged over all sources.

4.1.3. *Initialization*

As the EM algorithm is very sensitive to the initialization and to be sure to have a "good initialization", we provide it with *perturbed oracle initializations*, where the parameters $\mathbf{R}_n(f)$ and $v_n(t,f)$ are estimated form the original source spatial images as in [12] and then perturbed with a high level additive noise (SNR of 3 dB) as in [8]. Parameters $w_{f,k}^n, h_{k,t}^n$ of the NMF approaches are then computed with NMF decomposition using multiplicative update (MU) rules and KL divergence as in [8]. For the rank-1 methods (i.e. binary masking, Duong et al. rank-1 and NMF rank-1) we compute $\widehat{\mathbf{h}}_n(f)$ by calculating the first principal component of $\mathbf{R}_n(f)$ using the principal component analysis (PCA).

4.1.4. *Noise*

When the noise tends to zero, the estimation of the mixture parameters using the NMF rank-1 algorithm gets stuck [8] and when the noise is small, the convergence of this EM algorithm is very slow. Thus, the authors of [8] proposed a strategy called *noise annealing with noise injection* where the noise covariance $\mathbf{N}(t,f) = \sigma_b^2(f)\mathbf{I}$ is initialized with a large value of $\sigma_b^2(f)$ and instead of being re-estimated at each iteration, is gradually decreased through iterations to a small value. Noise injection means that a random noise with a covariance $\mathbf{N}(t,f)$ is added to $\mathbf{X}(t,f)$ at each EM iteration. This technique accelerates the overall global convergence [8]. Although this stuck problem doesn't hold in our full rank NMF algorithm, we used this *noise annealing with noise injection* scheme for both the NMF rank-1 algorithm and our NMF full rank algorithm[1].

4.2. Results

The results corresponding to reverberation times of 130 ms and 250 ms are respectively shown in Table 1 and Table 2.

[1] We also noticed that there is a marginal increase of the performance (between 0 and 0.5 dB of the SDR) of NMF full rank when using this *noise annealing with noise injection* scheme.

Table 1. Source separation performance, RT = 130 ms

Reverberation Time	130 ms			
Microphone distance	5 cm		1 m	
Number of sources	3	4	3	4
Appoaches	**SDR in dB**			
Binary masking	0.4	0.0	3.6	0.9
Duong et al. rank-1	1.2	0.2	2.2	0.8
Duong et al. full rank	**9.5**	**8.3**	9.8	8.3
NMF rank-1	8.7	7.1	7.3	4.6
NMF full rank	9.1	7.5	**10.2**	**8.5**

Table 2. Source separation performance, RT = 250 ms

Reverberation Time	250 ms			
Microphone distance	5 cm		1 m	
Number of sources	3	4	3	4
Appoaches	**SDR in dB**			
Binary masking	0.8	-0.3	2.9	0.7
Duong et al. rank-1	0.1	-0.1	1.5	0.5
Duong et al. full rank	8.1	7.2	8.1	7.1
NMF rank-1	7.7	6.4	5.5	3.8
NMF full rank	**8.8**	**7.5**	**9.6**	**8.0**

Unsurprisingly, when the number of sources increases as well as when the reverberation time increases, the performance of all the tested algorithms degrades. NMF full rank outperforms NMF rank-1 and NMF rank-1 outperforms Duong et al. rank-1 by between 3 dB and 7 dB. In the "low" reverberant setting (RT = 130 ms), Duong et al. full rank performs better than NMF full rank when the microphones distance is 5 cm, but less than NMF full rank when the microphone distance is 1 m. When the reverberation time is longer (RT = 250 ms), the NMF full rank outperforms all the other tested methods. Thus it shows that combining the full rank *spatial* covariance model with the NMF *spectral* model improves the separation in realistic reverberant environment.

5. CONCLUSION

In this paper we have introduced a new model for convolutive blind source separation that combines the advantages of the two existing models. The source spectrum is modeled via nonnegative matrix factorization (NMF) and the convolutive mixing process is modeled using a full rank spatial covariance instead of a rank-1. We addressed the estimation of the model parameters by maximizing the likelihood of the observed mixture using an EM algorithm.

Experimental results over music data in different settings (number of sources, microphone distance, reverberation time) validate that our model outperfoms the NMF rank-1 approach [8], the full rank method of Duong et al. [13] and binary masking, when the reverberation time is realistic (RT of 250 ms). Future works include the extension of other spectral models (e.g. GMM) to the full rank model and validation of the proposed method over real-word recordings with more than four sources. As the EM algorithm is sensitive to parameter initialization, it is important to investigate blind initialization procedures of the model parameters and particularly, the initialization of the spatial covariance matrices.

6. REFERENCES

[1] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, 2000.

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

[3] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a multichannel underdetermined mixture," *IEEE Transactions on Signal Processing*, vol. 58, no. 1, January 2010.

[4] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "Map-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and l 1-norm minimization," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, pp. 1–12, 2007.

[5] B. Pearlmutter P. O'Grady and S. Rickard, "Survey of sparse and non-sparse methods in source separation," *IJIST*, March 2005.

[6] S. Arberet, A. Ozerov, R. Gribonval, and F. Bimbot, "Blind spectral-GMM estimation for underdetermined instantaneous audio source separation," in *ICA*, 2009.

[7] C. Févotte, N. Bertin, and J.L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[8] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[9] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," in *ICA*, 2009, pp. 734–741.

[10] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal Processing*, vol. 81, no. 11, 2001.

[11] E. Vincent, "Complex nonconvex l_p norm minimisation for underdetermined source separation," in *ICA*, 2007.

[12] N.Q.K. Duong, E. Vincent, and R. Gribonval, "Spatial covariance models for under-determined reverberant audio source separation," in *WASPAA*, 2009.

[13] N.Q.K. Duong, E. Vincent, and R. Gribonval, "Underdetermined convolutive blind source separation using spatial covariance models," in *ICASSP*, 2010.

[14] A.P. Dempster, N.M. Laird, D.B. Rubin, et al., "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[15] D. Campbell, K. Palomäki, and G. Brown, "A matlab simulation of 'shoebox' room acoustics for use in research and teaching," *Computing and Information Systems Journal*, vol. 9, no. 3, pp. 48–51, October 2005.