Under-Determined Reverberant Audio Source Separation Using a Full-Rank Spatial Covariance Model

Ngoc Q. K. Duong, Emmanuel Vincent, *Senior Member, IEEE*, and Rémi Gribonval, *Senior Member, IEEE*

***Abstract*—This paper addresses the modeling of reverberant recording environments in the context of under-determined convolutive blind source separation. We model the contribution of each source to all mixture channels in the time–frequency domain as a zero-mean Gaussian random variable whose covariance encodes the spatial characteristics of the source. We then consider four specific covariance models, including a full-rank unconstrained model. We derive a family of iterative expectation–maximization (EM) algorithms to estimate the parameters of each model and propose suitable procedures adapted from the state-of-the-art to initialize the parameters and to align the order of the estimated sources across all frequency bins. Experimental results over reverberant synthetic mixtures and live recordings of speech data show the effectiveness of the proposed approach.**

***Index Terms*—Convolutive blind source separation (BSS), expectation–maximization (EM) algorithm, permutation problem, spatial covariance models, under-determined mixtures.**

I. Introduction

In blind source separation (BSS), audio signals are generally mixtures of several sound sources such as speech, music, and background noise. The recorded multichannel signal $\mathbf{x}(t)$ is therefore expressed as

$$\mathbf{x}(t) = \sum_{j=1}^{J} \mathbf{c}_j(t) \tag{1}$$

where $\mathbf{c}_j(t) = [c_{1j}(t), \ldots, c_{Ij}(t)]^T$ is the spatial image of the jth source, that is the contribution of this source to all mixture channels, and I is number of mixture channels. For a point source in a reverberant environment, $\mathbf{c}_j(t)$ can be expressed via the convolutive mixing process

$$\mathbf{c}_j(t) = \sum_{\tau} \mathbf{h}_j(\tau) s_j(t-\tau) \tag{2}$$

Manuscript received November 21, 2009; revised April 29, 2010. Date of publication May 18, 2010; date of current version August 13, 2010. Part of this work was presented at the 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing. The work of R. Gribonval was supported in part by Agence Nationale de la Recherche (ANR), project ECHANGE (ANR-08-EMER-006) and in part by the European Union through the project Sparse Models, Algorithms and Learning for Large-Scale data (SMALL). The project SMALL acknowledges the financial support of the Future and Emerging Technologies (FET) program within the Seventh Framework Program for Research of the European Commission, under FET-Open Grant 225913. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tomohiro Nakatani.

The authors are with INRIA, Centre Inria Rennes–Bretagne Atlantique, 35042 Rennes Cedex, France (e-mail: qduong@irisa.fr; emmanuel.vincent@inria.fr; remi.gribonval@inria.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2010.2050716

where $s_j(t)$ is the jth source signal and $\mathbf{h}_j(\tau) = [h_{1j}(\tau), \ldots, h_{Ij}(\tau)]^T$ the vector of filter coefficients modeling the acoustic path from this source to all microphones. Source separation consists in recovering either the J original source signals or their spatial images given the I mixture channels. In the following, we focus on the separation of under-determined mixtures, i.e., such that $I < J$, assuming that J is known.

Most existing approaches operate in the time–frequency domain using the short-time Fourier transform (STFT) and rely on narrowband approximation of the convolutive mixture (2) by complex-valued multiplication in each frequency bin f and time frame n as

$$\mathbf{c}_j(n,f) \approx \mathbf{h}_j(f) s_j(n,f) \tag{3}$$

where the $I \times 1$ mixing vector $\mathbf{h}_j(f)$ is the Fourier transform of $\mathbf{h}_j(\tau)$, $s_j(n,f)$ are the STFT coefficients of the sources $s_j(t)$ and $\mathbf{c}_j(n,f) = [c_{1j}(n,f), \ldots, c_{Ij}(n,f)]^T$ the STFT coefficients of their spatial images $\mathbf{c}_j(t)$. The sources are typically estimated under the assumption that they are sparse in the STFT domain. For instance, the degenerate unmixing estimation technique (DUET) [2] uses binary masking to extract the predominant source in each time–frequency bin. Another popular technique known as ℓ_1-norm minimization extracts on the order of I sources per time–frequency bin by solving a constrained ℓ_1-minimization problem [3], [4]. The separation performance achievable by these techniques remains limited in reverberant environments [5], due in particular to the fact that the narrowband approximation does not hold because the mixing filters are much longer than the window length of the STFT.

Recently, a distinct framework has emerged whereby the STFT coefficients of the source images $\mathbf{c}_j(n,f)$ are modeled by a phase-invariant multivariate distribution whose parameters are functions of (n,f) [6]. One instance of this framework consists in modeling $\mathbf{c}_j(n,f)$ as a zero-mean Gaussian random variable with covariance matrix $\mathbf{R}_{\mathbf{c}_j}(n,f) = \mathbb{E}(\mathbf{c}_j(n,f)\mathbf{c}_j^H(n,f))$ factored as

$$\mathbf{R}_{\mathbf{c}_j}(n,f) = v_j(n,f)\,\mathbf{R}_j(f) \tag{4}$$

where $v_j(n,f)$ are scalar time-varying *variances* encoding the spectro-temporal power of the sources and $\mathbf{R}_j(f)$ are $I \times I$ time-invariant *spatial covariance matrices* encoding their spatial position and spatial spread [7]. The model parameters can then be estimated in the maximum likelihood (ML) sense and used to estimate the spatial images of all sources by Wiener filtering.

This framework was first applied to the separation of instantaneous audio mixtures in [8] and [9] and shown to provide better separation performance than ℓ_1-norm minimization. The instantaneous mixing process then translated into a rank-1 spatial covariance matrix for each source. In our preliminary paper [7], we extended this approach to convolutive mixtures and proposed to consider full-rank spatial covariance matrices modeling the spatial spread of the sources and circumventing the narrowband approximation to a certain extent. This approach was shown to improve separation performance of reverberant mixtures in both an *oracle* context, where all model parameters are known, and in a *semi-blind* context, where the spatial covariance matrices of all sources are known but their variances are blindly estimated from the mixture.

In [1] and the following, we extend this work to *blind* estimation of the model parameters as required for realistic BSS application. This paper provides three main contributions. First, we explain the appropriateness of full-rank spatial covariance models in the context of reverberant source separation and propose a new full-rank unconstrained model. Second, we design parameter estimation algorithms for these models by deriving the corresponding expectation–maximization (EM) [10] update rules and adapting the sparsity-based algorithms in [3], [11] for parameter initialization and permutation alignment. Third, we show that the proposed full-rank unconstrained model outperforms state-of-the-art algorithms on a wide range of data and estimation scenarios.

The structure of the rest of the paper is as follows. We introduce the general framework under study as well as four specific spatial covariance models in Section II. We then address the blind estimation of all model parameters from the observed mixture in Section III. We compare the source separation performance achieved by each model to that of state-of-the-art techniques in various experimental settings in Section IV. Finally, we conclude and discuss further research directions in Section V.

II. General Framework and Spatial Covariance Models

We start by describing the general probabilistic modeling framework adopted from now on. We then define four models with different degrees of flexibility resulting in rank-1 or full-rank spatial covariance matrices.

A. General Framework

Let us assume that the vector $\mathbf{c}_j(n,f)$ of STFT coefficients of the spatial image of the jth source follows a zero-mean Gaussian distribution whose covariance matrix factors as in (4). Under the classical assumption that the sources are uncorrelated, the vector $\mathbf{x}(n,f)$ of STFT coefficients of the mixture signal is also zero-mean Gaussian with covariance matrix

$$\mathbf{R}_\mathbf{x}(n,f) = \sum_{j=1}^{J} v_j(n,f)\,\mathbf{R}_j(f). \tag{5}$$

In other words, the likelihood of the set of observed mixture STFT coefficients $\mathbf{x} = \{\mathbf{x}(n,f)\}_{n,f}$ given the set of variance parameters $v = \{v_j(n,f)\}_{j,n,f}$ and that of spatial covariance matrices $\mathbf{R} = \{\mathbf{R}_j(f)\}_{j,f}$ is given by

$$P(\mathbf{x}\,|\,v,\mathbf{R}) = \prod_{n,f} \frac{1}{\det(\pi\mathbf{R}_\mathbf{x}(n,f))} e^{-\mathbf{x}^H(n,f)\mathbf{R}_\mathbf{x}^{-1}(n,f)\mathbf{x}(n,f)} \tag{6}$$

where ^{H} denotes matrix conjugate transposition and $\mathbf{R}_\mathbf{x}(n,f)$ implicitly depends on v and $\mathbf{R}$ according to (5). In the following, we assume that the source variances are unconstrained and focus on modeling the covariance matrices by higher level spatial parameters.

Under this model, source separation can be achieved in two steps. The variance parameters v and the spatial parameters underlying $\mathbf{R}$ are first estimated in the ML sense. The spatial images of all sources are then obtained in the minimum mean square error (MMSE) sense by multichannel Wiener filtering

$$\hat{\mathbf{c}}_j(n,f) = v_j(n,f)\mathbf{R}_j(f)\mathbf{R}_\mathbf{x}^{-1}(n,f)\mathbf{x}(n,f). \tag{7}$$

B. Rank-1 Convolutive Model

Most existing approaches to audio source separation rely on narrowband approximation of the convolutive mixing process (2) by the complex-valued multiplication (3). The covariance matrix of $\mathbf{c}_j(n,f)$ is then given by (4), where $v_j(n,f)$ is the variance of $s_j(n,f)$ and $\mathbf{R}_j(f)$ is equal to the rank-1 matrix

$$\mathbf{R}_j(f) = \mathbf{h}_j(f)\mathbf{h}_j^H(f) \tag{8}$$

with $\mathbf{h}_j(f)$ denoting the Fourier transform of the mixing filters $\mathbf{h}_j(\tau)$. This *rank-1 convolutive model* of the spatial covariance matrices has recently been exploited in [12] together with a different model of the source variances.

C. Rank-1 Anechoic Model

For omnidirectional microphones in an anechoic recording environment without reverberation, each mixing filter boils down to the combination of a delay τ_{ij} and a gain κ_{ij} specified by the distance r_{ij} from the jth source to the ith microphone [13]

$$\tau_{ij} = \frac{r_{ij}}{c} \quad \text{and} \quad \kappa_{ij} = \frac{1}{\sqrt{4\pi}r_{ij}} \tag{9}$$

where c is sound velocity. The spatial covariance matrix of the jth source is hence given by the *rank-1 anechoic model*

$$\mathbf{R}_j(f) = \mathbf{a}_j(f)\mathbf{a}_j^H(f) \tag{10}$$

where the Fourier transform $\mathbf{a}_j(f)$ of the mixing filters is now parameterized as

$$\mathbf{a}_j(f) = \begin{pmatrix} \kappa_{1,j}e^{-2i\pi f\tau_{1,j}} \\ \vdots \\ \kappa_{I,j}e^{-2i\pi f\tau_{I,j}} \end{pmatrix}. \tag{11}$$

D. Full-Rank Direct+Diffuse Model

One possible interpretation of the narrowband approximation is that the sound of each source as recorded on the microphones

comes from a single spatial position at each frequency f, as specified by $\mathbf{h}_j(f)$ or $\mathbf{a}_j(f)$. This approximation is not valid in a reverberant environment, since reverberation induces some spatial spread of each source, due to echoes at many different positions on the walls of the recording room. This spread translates into full-rank spatial covariance matrices.

The theory of statistical room acoustics assumes that the spatial image of each source is composed of two uncorrelated parts: a direct part, which is modeled by $\mathbf{a}_j(f)$ in (11) for omnidirectional microphones, and a reverberant part. The spatial covariance $\mathbf{R}_j(f)$ of each source is then a full-rank matrix defined as the sum of the covariance of its direct part and the covariance of its reverberant part such that

$$\mathbf{R}_j(f) = \mathbf{a}_j(f)\mathbf{a}_j^H(f) + \sigma^2_{\text{rev}}\mathbf{\Psi}(f) \tag{12}$$

where σ^2_{rev} is the variance of the reverberant part and $\Psi_{il}(f)$ is a function of the distance d_{il} between the ith and the lth microphone such that $\Psi_{ii}(f) = 1$. This model assumes that the reverberation recorded at all microphones has the same power but is correlated as characterized by $\Psi_{il}(f)$. This model has been employed for single source localization in [13] but not for source separation yet.

Assuming that the reverberant part is diffuse, i.e., its intensity is uniformly distributed over all possible directions, its normalized cross-correlation can be shown to be real-valued and equal to [14]

$$\Psi_{il}(f) = \frac{\sin(2\pi f d_{il}/c)}{2\pi f d_{il}/c}. \tag{13}$$

Moreover, the power of the reverberant part within a parallelepipedic room with dimensions L_x, L_y, and L_z is given by

$$\sigma^2_{\text{rev}} = \frac{4\beta^2}{\mathcal{A}(1-\beta^2)} \tag{14}$$

where $\mathcal{A}$ is the total wall area and β the wall reflection coefficient computed from the room reverberation time T_{60} via Eyring's formula [13]

$$\beta = \exp\left\{-\frac{13.82}{\left(\frac{1}{L_x}+\frac{1}{L_y}+\frac{1}{L_z}\right)cT_{60}}\right\}. \tag{15}$$

Note that the covariance matrix $\mathbf{\Psi}(f)$ is usually employed for the modeling of diffuse background noise. For instance, the source separation algorithm in [15] assumes that the sources follow an anechoic model and represents the non-direct part of all sources by a shared diffuse noise component with covariance $\mathbf{\Psi}(f)$ and constant variance. Hence, this algorithm does not account for correlation between the variances of the direct part and the non-direct part. On the contrary, the direct+diffuse model scales the direct and non-direct part of $\mathbf{R}_j(f)$ by the same variance $v_j(n,f)$, which is more consistent with the physics of sound.

E. Full-Rank Unconstrained Model

In practice, the assumption that the reverberant part is diffuse is rarely satisfied in realistically reverberant environments. Indeed, early echoes accounting for most of its energy are not uniformly distributed on the boundaries of the recording room. When performing some simulations in a rectangular room, we observed that (13) is valid on average when considering a large number of sources distributed at different positions in a room, but generally not valid for each individual source.

Therefore, we also investigate the modeling of each source via a full-rank unconstrained spatial covariance matrix $\mathbf{R}_j(f)$ whose coefficients are unrelated *a priori*. This model is the most general possible model for a covariance matrix. It generalizes the above three models in the sense that any matrix taking the form of (8), (10) or (l2) can also be considered as an unconstrained matrix. Because of this increased flexibility, this unconstrained model better fits the data as measured by the likelihood. In particular, it improves the poor fit between the model and the data observed for rank-1 models due to the fact that the narrowband approximation underlying these models does not hold for reverberant mixtures. In that sense, it circumvents the narrowband approximation to a certain extent.

The entries of $\mathbf{R}_j(f)$ are not directly interpretable in terms of simple geometrical quantities. The principal component of the matrix can be interpreted as a beamformer [16] pointing towards the direction of maximum output power, while the ratio between its largest eigenvalue and its trace is equal to the ratio between the output and input power of that beamformer. In moderate reverberation conditions, the former is expected to be close to the source direction of arrival (DOA) while the latter is related to the ratio between the power of direct sound and that of reverberation. However, the strength of this model is precisely that it remains valid to a certain extent in more reverberant environments, since it is the most general possible model for a covariance matrix.

III. Blind Estimation of the Model Parameters

In order to use the above models for BSS, we need to estimate their parameters from the mixture signal only. In our preliminary paper [7], we used a quasi-Newton algorithm for semi-blind separation that converged in a very small number of iterations. However, due to the complexity of each iteration, we later found out that the EM algorithm, which is a popular choice for Gaussian models [15], [17], [18], provided faster convergence despite a larger number of iterations.

As any iterative optimization algorithm, EM is sensitive to initialization [12] so that a suitable parameter initialization scheme is necessary. Also, the well-known source permutation problem must be addressed when the model parameters are independently estimated at different frequencies [11]. We hence adopt the following three-step procedure as depicted in Fig. 1: initialization of $\mathbf{h}_j(f)$ or $\mathbf{R}_j(f)$ by hierarchical clustering, iterative ML estimation of all model parameters via EM, and DOA-based permutation alignment. The latter step is needed only for the rank-1 convolutive model and the full-rank unconstrained model whose parameters are estimated independently in each frequency bin. It is conducted after EM parameter estimation, since optimized parameters provide better DOA information than initial ones and EM does not always preserve the order of the sources.

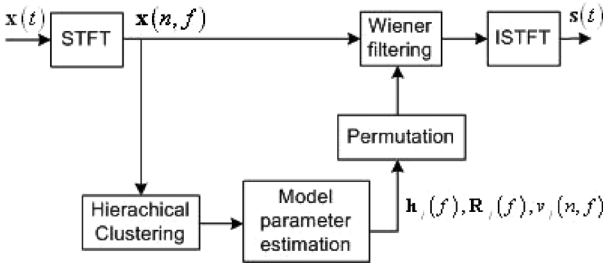

Fig. 1. Flow of the proposed blind source separation approach.

A. Initialization by Hierarchical Clustering

Preliminary experiments showed that the initialization of the model parameters greatly affects the separation performance resulting from the EM algorithm. Yet, the parameter initialization schemes previously proposed for rank-1 Gaussian models are either restricted to instantaneous mixtures [18] or non-blind [7], [12]. By contrast, a number of clustering algorithms have been proposed for blind estimation of the mixing vectors in the context of sparsity-based convolutive source separation. In the following, we use up to minor improvements the hierarchical clustering-based algorithm in [3] for the purpose of parameter initialization of rank-1 models and introduce a modified version of this algorithm for parameter initialization of full-rank models.

The algorithm in [3] relies on the assumptions that at each frequency f the sounds of all sources come from disjoint regions of space and that a single source predominates in most time–frequency bins. The vectors $\mathbf{x}(n,f)$ of mixture STFT coefficients then cluster around the direction of the associated mixing vector $\mathbf{h}_j(f)$ in the time frames n where the jth source is predominant. It is well known that the validity of the latter sparsity assumption decreases with increasing reverberation. Nevertheless, this algorithm was explicitly developed for reverberant mixtures.

In order to estimate these clusters, the vectors of mixture STFT coefficients are first normalized as

$$\bar{\mathbf{x}}(n,f) \leftarrow \frac{\mathbf{x}(n,f)}{\|\mathbf{x}(n,f)\|_2} e^{-i\arg(x_1(n,f))} \tag{16}$$

where $\arg(\cdot)$ denotes the phase of a complex number and $\|\cdot\|_2$ the Euclidean norm. We then define the distance between two clusters C_1 and C_2 by the average distance between the associated normalized mixture STFT coefficients

$$d(C_1,C_2) = \frac{1}{|C_1||C_2|} \sum_{\bar{\mathbf{x}}_{c1}\in C_1} \sum_{\bar{\mathbf{x}}_{c2}\in C_2} \|\bar{\mathbf{x}}_{c1} - \bar{\mathbf{x}}_{c2}\|_2. \tag{17}$$

In a given frequency bin f, each normalized vector of mixture STFT coefficients $\bar{\mathbf{x}}(n,f)$ at a time frame n is first considered as a cluster containing a single item. The distance between each pair of clusters is computed and the two clusters with the smallest distance are merged. This "bottom up" process called linking is repeated until the number of clusters is smaller than a predetermined threshold K. This threshold is usually much larger than the number of sources J [3], so as to eliminate outliers. We finally choose the J clusters with the largest number of samples and compute the initial mixing vector and spatial covariance matrix for each source as

$$\mathbf{h}_j^{\text{init}}(f) = \frac{1}{|C_j|} \sum_{\bar{\mathbf{x}}(n,f)\in C_j} \tilde{\mathbf{x}}(n,f) \tag{18}$$

$$\mathbf{R}_j^{\text{init}}(f) = \frac{1}{|C_j|} \sum_{\bar{\mathbf{x}}(n,f)\in C_j} \tilde{\mathbf{x}}(n,f)\tilde{\mathbf{x}}(n,f)^H \tag{19}$$

where $\tilde{\mathbf{x}}(n,f) = \mathbf{x}(n,f)e^{-i\arg(x_1(n,f))}$, and $|C_j|$ denotes the total number of samples in cluster C_j, which depends on the considered frequency bin f.

Note that, contrary to the algorithm in [3], we define the distance between clusters as the average distance between the normalized mixture STFT coefficients instead of the minimum distance between them. Besides, the mixing vector $\mathbf{h}_j^{\text{init}}(f)$ is computed from the phase-normalized mixture STFT coefficients $\tilde{\mathbf{x}}(n,f)$ instead of both phase and amplitude normalized coefficients $\bar{\mathbf{x}}(n,f)$. This increases the weight of time–frequency bins of large amplitude where the modeled source is more likely to be prominent, in a way similar to [2]. These modifications were found to provide better initial approximation of the mixing parameters in our experiments. We also tested random initialization and DOA-based initialization, i.e., where the mixing vectors $\mathbf{h}_j^{\text{init}}(f)$ are derived from known source and microphone positions assuming no reverberation. Both schemes were found to result in slower convergence and poorer separation performance than the chosen scheme.

The source variances were initialized to $v_j^{\text{init}}(n,f) = 1$. This basic initialization scheme did not significantly affect performance compared to the slower advanced scheme consisting of finding the $v_j^{\text{init}}(n,f)$ most consistent with $\mathbf{h}_j^{\text{init}}(f)$ and $\mathbf{R}_j^{\text{init}}(f)$ by running EM without updating the mixing vectors or the spatial covariance matrices.

B. EM Updates for the Rank-1 Convolutive Model

The derivation of the EM parameter estimation algorithm for the rank-1 convolutive model is strongly inspired from the study in [12]. Indeed it relies on the same model of spatial covariance matrices but on a distinct unconstrained model of the source variances. Similarly to [12], EM cannot be directly applied to the mixture model (1) since the estimated mixing vectors remain fixed to their initial value. This issue can be addressed by considering the noisy mixture model

$$\mathbf{x}(n,f) = \mathbf{H}(f)\mathbf{s}(n,f) + \mathbf{b}(n,f) \tag{20}$$

where $\mathbf{H}(f)$ is the mixing matrix whose jth column is the mixing vector $\mathbf{h}_j(f)$, $\mathbf{s}(n,f)$ is the vector of source STFT coefficients $s_j(n,f)$ and $\mathbf{b}(n,f)$ some additive zero-mean Gaussian noise. We denote by $\mathbf{R}_{\mathbf{s}}(n,f)$ the diagonal covariance matrix of $\mathbf{s}(n,f)$. Following [12], we assume that $\mathbf{b}(n,f)$ is stationary and spatially uncorrelated and denote by $\mathbf{R}_{\mathbf{b}}(f)$ its time-invariant diagonal covariance matrix. This matrix is initialized to a small value related to the average empirical channel variance as discussed in [12].

EM is separately derived for each frequency bin f for the *complete data* $\{\mathbf{x}(n,f), s_j(n,f)\}_{j,n}$ that is the set of observed mixture STFT coefficients and hidden source STFT coefficients

of all time frames. The details of one iteration are as follows. In the E-step, the Wiener filter $\mathbf{W}(n,f)$ and the conditional mean $\hat{\mathbf{s}}(n,f)$ and covariance $\hat{\mathbf{R}}_{\mathbf{ss}}(n,f)$ of the sources are computed as

$$\mathbf{R}_{\mathbf{s}}(n,f) = \mathrm{diag}(v_1(n,f),\ldots,v_J(n,f)) \quad (21)$$

$$\mathbf{R}_{\mathbf{x}}(n,f) = \mathbf{H}(f)\mathbf{R}_{\mathbf{s}}(n,f)\mathbf{H}^H(f) + \mathbf{R}_{\mathbf{b}}(f) \quad (22)$$

$$\mathbf{W}(n,f) = \mathbf{R}_{\mathbf{s}}(n,f)\mathbf{H}^H(f)\mathbf{R}_{\mathbf{x}}^{-1}(n,f) \quad (23)$$

$$\hat{\mathbf{s}}(n,f) = \mathbf{W}(n,f)\mathbf{x}(n,f) \quad (24)$$

$$\hat{\mathbf{R}}_{\mathbf{ss}}(n,f) = \hat{\mathbf{s}}(n,f)\hat{\mathbf{s}}^H(n,f) + (\mathbf{I} - \mathbf{W}(n,f)\mathbf{H}(f))\mathbf{R}_{\mathbf{s}}(n,f) \quad (25)$$

where $\mathbf{I}$ is the $I \times I$ identity matrix and $\mathrm{diag}(\cdot)$ the diagonal matrix whose entries are given by its arguments. Conditional expectations of multichannel statistics are also computed by averaging over all N time frames as

$$\hat{\mathbf{R}}_{\mathbf{ss}}(f) = \frac{1}{N}\sum_{n=1}^{N}\hat{\mathbf{R}}_{\mathbf{ss}}(n,f) \quad (26)$$

$$\hat{\mathbf{R}}_{\mathbf{xs}}(f) = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}(n,f)\hat{\mathbf{s}}^H(n,f) \quad (27)$$

$$\hat{\mathbf{R}}_{\mathbf{xx}}(f) = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}(n,f)\mathbf{x}^H(n,f). \quad (28)$$

In the M-step, the source variances, the mixing matrix and the noise covariance are updated via

$$v_j(n,f) = \hat{\mathbf{R}}_{\mathbf{ss}\,jj}(n,f) \quad (29)$$

$$\mathbf{H}(f) = \hat{\mathbf{R}}_{\mathbf{xs}}(f)\hat{\mathbf{R}}_{\mathbf{ss}}^{-1}(f) \quad (30)$$

$$\mathbf{R}_{\mathbf{b}}(f) = \mathrm{Diag}(\hat{\mathbf{R}}_{\mathbf{xx}}(f) - \mathbf{H}(f)\hat{\mathbf{R}}_{\mathbf{xs}}^H(f) - \hat{\mathbf{R}}_{\mathbf{xs}}\mathbf{H}^H(f) + \mathbf{H}(f)\hat{\mathbf{R}}_{\mathbf{ss}}(n,f)\mathbf{H}^H(f)) \quad (31)$$

where $\mathrm{Diag}(\cdot)$ projects a matrix onto its diagonal.

C. EM Updates for the Full-Rank Unconstrained Model

The derivation of EM for the full-rank unconstrained model is much easier since the above issue does not arise. We hence stick with the exact mixture model (1), which can be seen as an advantage of full-rank versus rank-1 models. EM is again separately derived for each frequency bin f. Since the mixture can be recovered from the spatial images of all sources, the complete data reduces to $\{\mathbf{c}_j(n,f)\}_{n,f}$, that is the set of hidden STFT coefficients of the spatial images of all sources on all time frames. The details of one iteration are as follows. In the E-step, the Wiener filter $\mathbf{W}_j(n,f)$ and the conditional mean $\hat{\mathbf{c}}_j(n,f)$ and covariance $\hat{\mathbf{R}}_{\mathbf{c}_j}(n,f)$ of the spatial image of the jth source are computed as

$$\mathbf{W}_j(n,f) = \mathbf{R}_{\mathbf{c}_j}(n,f)\mathbf{R}_{\mathbf{x}}^{-1}(n,f) \quad (32)$$

$$\hat{\mathbf{c}}_j(n,f) = \mathbf{W}_j(n,f)\mathbf{x}(n,f) \quad (33)$$

$$\hat{\mathbf{R}}_{\mathbf{c}_j}(n,f) = \hat{\mathbf{c}}_j(n,f)\hat{\mathbf{c}}_j^H(n,f) + (\mathbf{I} - \mathbf{W}_j(n,f))\mathbf{R}_{\mathbf{c}_j}(n,f) \quad (34)$$

where $\mathbf{R}_{\mathbf{c}_j}(n,f)$ is defined in (4) and $\mathbf{R}_{\mathbf{x}}(n,f)$ in (5). In the M-step, the variance and the spatial covariance of the jth source are updated via

$$v_j(n,f) = \frac{1}{I}\mathrm{tr}\left(\mathbf{R}_j^{-1}(f)\hat{\mathbf{R}}_{\mathbf{c}_j}(n,f)\right) \quad (35)$$

$$\mathbf{R}_j(f) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{v_j(n,f)}\hat{\mathbf{R}}_{\mathbf{c}_j}(n,f) \quad (36)$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a square matrix. Note that, strictly speaking, this algorithm is a generalized form of EM [19], since the M-step increases but does not maximize the likelihood of the complete data due to the interleaving of (35) and (36). The increase of the log-likelihood and of the separation performance resulting from these updates is illustrated in Section IV-C.

D. EM Updates for the Rank-1 Anechoic Model and the Full-Rank Direct + Diffuse Model

The derivation of EM for the two remaining models is more complex since the M-step cannot be expressed in closed form. The complete data and the E-step for the rank-1 anechoic model and the full-rank direct+diffuse model are identical to those for the rank-1 convolutive model and the full-rank unconstrained model, respectively. The M-step, which consists of maximizing the likelihood of the complete data given their natural statistics computed in the E-step, could be addressed, e.g., via a quasi-Newton technique or by sampling possible parameter values from a grid [15]. In the following, we do not attempt to derive the details of these algorithms since these two models appear to provide lower performance than the rank-1 convolutive model and the full-rank unconstrained model in a semi-blind context, as discussed in Section IV-B.

E. Permutation Alignment

Since the parameters of the rank-1 convolutive model and the full-rank unconstrained model, i.e., $\mathbf{h}_j(f)$, $\mathbf{R}_j(f)$ and $v_j(n,f)$, are estimated independently in each frequency bin f, they should be ordered so as to correspond to the same source across all frequency bins. This so-called permutation problem has been widely studied in the context of sparsity-based source separation. In the following, we apply the DOA-based algorithm in [11] to the rank-1 model and explain how to adapt this algorithm to the full-rank model.

The principle of this algorithm is as follows. Given the geometry of the microphone array, a critical frequency is determined above which spatial aliasing may occur. The mixing vectors $\mathbf{h}_j(f)$ are each unambiguously related to a certain DOA below that frequency while phase wrapping may occur at higher frequencies. The algorithm first estimates the source DOAs and the permutations at low frequencies by clustering the mixing vectors after suitable normalization assuming no phase wrapping and then re-estimates them at all frequencies by taking phase wrapping into account. Note that the order of source variances $v_j(n,f)$ in each frequency bin is permuted identically to that of the mixing vectors $\mathbf{h}_j(f)$.

Regarding the full-rank model, we first apply principal component analysis (PCA) to summarize the spatial covariance matrix $\mathbf{R}_j(f)$ of each source in each frequency bin by its first

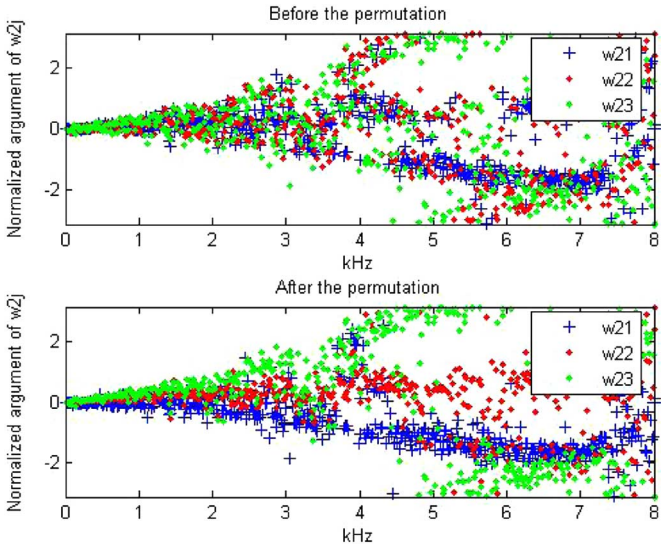

Fig. 2. Normalized argument of $w_{2j}(f)$ before and after permutation alignment from a real-world stereo recording of three sources with $\mathrm{T}_{60} = 250$ ms.

principal component $\mathbf{w}_j(f)$ that points to the direction of maximum variance. This vector is conceptually equivalent to the mixing vector $\mathbf{h}_j(f)$ of the rank-1 model. Thus, we can apply the same procedure to solve the permutation problem. Fig. 2 depicts the phase of the second entry $w_{2j}(f)$ of $\mathbf{w}_j(f)$ before and after solving the permutation for a real-world stereo recording of three female speech sources with room reverberation time $T_{60} = 250$ ms, where $\mathbf{w}_j(f)$ has been normalized as in (16). The critical frequency below which this phase is unambiguously related to the source DOAs is here equal to 5 kHz [11]. The source order appears globally aligned for most frequency bins after solving the permutation.

IV. Experimental Evaluation

We evaluate the above models and algorithms under three different experimental settings. First, we compare all four models in a semi-blind setting so as to estimate an upper bound of their separation performance. Based on these results, we select two models for further study, namely the rank-1 convolutive model and the full-rank unconstrained model. Second, we evaluate these models in a blind setting over synthetic reverberant speech mixtures and compare them to state-of-the-art algorithms over the real-world speech mixtures of the 2008 Signal Separation Evaluation Compaign (SiSEC 2008) [5]. Finally, we assess the robustness of these two models to source movements in a semi-blind setting.

A. Common Parameter Settings and Performance Criteria

The common parameter setting for all experiments are summarized in Table I. In order to evaluate the separation performance of the algorithms, we use the signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), signal-to-artifact ratio (SAR) and source image-to-spatial distortion ratio (ISR) criteria expressed in decibels (dB), as defined in [20]. These criteria account respectively for overall distortion of the target source, residual crosstalk from other sources, musical noise, and spatial or filtering distortion of the target.

TABLE I
Common Experimental Parameter Setting

Signal duration	10 s
Number of channels	$I = 2$
Sampling rate	16 kHz
Window type	sine window
STFT frame size	1024
STFT frame shift	512
Propagation velocity	343 m/s
Number of EM iterations	10
Cluster threshold	$K = 30$

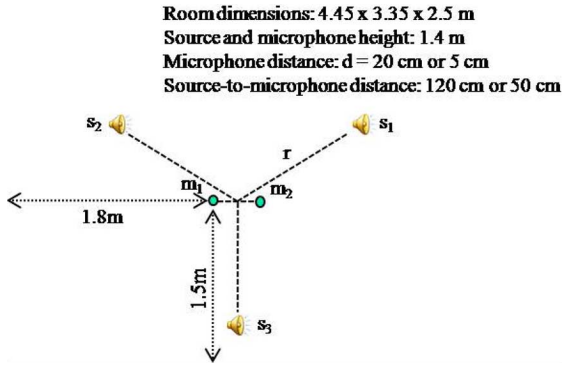

Fig. 3. Room geometry setting for synthetic convolutive mixtures.

B. Potential Source Separation Performance of All Models

The first experiment is devoted to the investigation of the potential source separation performance achievable by each model in a semi-blind context, i.e., assuming knowledge of the true spatial covariance matrices. We generated ten stereo synthetic mixtures of three speech sources, i.e., two mixtures with male voices only, two mixtures with female voices only, and six mixtures with mixed male and female voices, by convolving different sets of speech signals with room impulse responses simulated via the source image method. The positions of the sources and the microphones are illustrated in Fig. 3. The distance from each source to the center of the microphone pair was 120 cm and the microphone spacing was 20 cm. The reverberation time was set to $\mathrm{T}_{60} = 250$ ms.

The true spatial covariance matrices $\mathbf{R}_j(f)$ of all sources were computed either from the positions of the sources and the microphones and other room parameters or from the mixing filters. More precisely, we used the equations in Sections II-B–II-D for rank-1 models and the full-rank direct+diffuse model and ML estimation from the spatial images of the true sources for the full-rank unconstrained model. The source variances were then estimated from the mixture using the quasi-Newton technique in [7], for which an efficient initialization exists when the spatial covariance matrices are fixed. Binary masking, ℓ_1-norm minimization and ℓ_0-norm minimization were also evaluated for comparison using the reference software in [5][1] with the same mixing vectors as for the rank-1 convolutive model. The SDR obtained with ℓ_1-norm

[1]Binary masking is achieved in each time–frequency bin by projecting the mixture STFT coefficients $\mathbf{x}(n, f)$ onto the subspace spanned by each mixing vector $\mathbf{h}_j(f)$ and selecting the source j_0 whose projection has largest ℓ_2 norm. The spatial image $\hat{\mathbf{c}}_{j_0}(n, f)$ of this source is then set to the projected mixture STFT coefficients, while that of the other sources is set to zero.

TABLE II
AVERAGE POTENTIAL SOURCE SEPARATION PERFORMANCE IN A SEMI-BLIND SETTING OVER STEREO MIXTURES OF THREE SOURCES WITH $T_{60} = 250$ ms

Covariance models	Number of spatial parameters	SDR	SIR	SAR	ISR
Rank-1 anechoic	6	0.9	2.4	7.8	5.1
Rank-1 convolutive	3078	4.0	7.9	5.4	9.5
Full-rank direct+diffuse	8	3.2	6.7	5.3	7.9
Full-rank unconstrained	6156	5.7	10.8	7.3	11.0
Binary masking	3078	4.5	10.1	5.0	9.4
ℓ_0-norm minimization	3078	4.1	7.7	5.9	9.5

minimization was about 0.2 dB below that given by ℓ_0-norm minimization; therefore, only the latter is considered for comparison hereafter. The results are averaged over all sources and all set of mixtures and shown in Table II, together with the number of spatial parameters of each model, i.e., the number of parameters encoding the spatial characteristics of the sources.

The rank-1 anechoic model has lowest performance in terms of SDR, SIR, and ISR because it only accounts for the direct path. By contrast, the full-rank unconstrained model has highest performance in terms of SDR and ISR. It improves the SDR by 1.7, 1.2, and 1.6 dB when compared to the rank-1 convolutive model, binary masking, and ℓ_0-norm minimization, respectively. The full-rank direct+diffuse model results in a SDR decrease of 0.8 dB compared to the rank-1 convolutive model. This decrease appears surprisingly small when considering the fact that the former involves only eight spatial parameters (six distances r_{ij}, plus σ^2_{rev} and d) instead of 3078 parameters (six mixing coefficients per frequency bin) for the latter. Nevertheless, we focus on the two best models, namely the rank-1 convolutive model and the full-rank unconstrained model in subsequent experiments.

C. Blind Source Separation Performance as a Function of the Reverberation Time

The second experiment aims to investigate two things: first, the blind source separation performance achieved via these two models as well as via binary masking and ℓ_0-norm minimization; and second, the convergence property of EM iterations for the proposed full-rank unconstrained model in different reverberant conditions. Ten synthetic speech mixtures were generated in the same as in the first experiment for each reverberant condition, except that the microphone spacing was changed to 5 cm in order to reduce permutation alignment errors caused by spatial aliasing and the distance from the sources to the microphones to 50 cm. The reverberation time was varied in the range from 50 to 500 ms.

The resulting source separation performance in terms of SDR, SIR, SAR, and ISR is depicted in Fig. 4. Interestingly, the rank-1 convolutive model and ℓ_0-norm minimization results in a very similar SIR, ISR, and even SDR. Besides, we observe that in a low reverberant environment, i.e., $T_{60} = 50$ ms, the rank-1 convolutive model provides a very similar SDR and SAR to the full-rank model. This is consistent with the fact that the direct part contains most of the energy received at the microphones, so that the rank-1 spatial covariance matrix provides similar modeling accuracy to the full-rank model with fewer parameters. However, in an environment with realistic reverberation time, i.e., $T_{60} \geq 130$ ms, the full-rank unconstrained model outperforms both the rank-1 model and binary masking in terms of SDR and SAR and results in a SIR not very far below that of binary masking. For instance, with $T_{60} = 500$ ms, the SDR achieved via the full-rank unconstrained model is 2.0, 1.1, and 2.3 dB larger than that of the rank-1 convolutive model, binary masking, and ℓ_0-norm minimization, respectively. These results confirm the effectiveness of our proposed model parameter estimation scheme and also show that full-rank spatial covariance matrices better approximate the mixing process in a reverberant room.

The convergence property of EM iteration in different reverberant conditions for the proposed full-rank unconstrained model is evaluated in terms of the log-likelihood convergence as well as the averaged SDR improvement of separation results, and is shown in Fig. 5. The log-likelihood values were computed as the average logarithm of $P(\mathbf{x}(n,f)|\{v_j(n,f)\}_j, \{\mathbf{R}_j(f)\}_j)$ for all time frame n and frequency bin f after each EM iteration. The first SDR and log-likelihood value was computed from the initialization values of $v_j(n,f)$ and $\mathbf{R}_j(f)$. It can be seen that both the SDR and log-likelihood value are gradually increased after each EM iteration and the SDR increase fast during the first 3 EM iterations. After 10 EM iterations, SDR is improved by 2.2 dB and 1.7 dB in the reverberation time of 130 ms and 250 ms, respectively. These figures prove the effectiveness of the derived EM algorithm in the overall proposed system.

D. Blind Source Separation Performance as a Function of the Angle Between Sources

The third experiment is devoted to the assessment of the robustness of the considered blind source separation algorithms to a challenging condition where the source directions become closer. For that purpose, we simulated room impulse responses via the source image method for the same room and the same microphone positions as in Fig. 3, with a distance of 50 cm from the sources to the center of the microphone pair and a reverberation time of $T_{60} = 250$ ms, but changed the DOAs of the three sources to $90^0 - \alpha$, 90^0 and $90^0 + \alpha$, respectively, where $\alpha = 5^0, 10^0, 15^0, 30^0$ or 60^0 is the angular distance between sources. We then generated ten synthetic convolutive mixtures in the same way as for previous experiments for each value of α.

The average SDR achieved by the full-rank unconstrained model, the rank-1 convolutive model as well as binary masking and ℓ_0-norm minimization is depicted in Fig. 6. As expected, all algorithms result in lower separation performance when the source directions are closer due in particular to poorer estimation of the spatial parameters, i.e., spatial covariance matrices $\mathbf{R}_j(f)$ for the full-rank unconstrained model and mixing vectors $\mathbf{h}_j(f)$ for other algorithms, but the full-rank unconstrained model still outperforms other algorithms in all cases. For instance, when the sources are very close to each other where $\alpha = 5^0$, the full-rank unconstrained model offers 0.9 dB SDR while binary masking only provides 0.2 dB SDR and both rank-1 convolutive model and ℓ_0-norm minimization results in negative SDR. This supports the benefit of the full-rank unconstrained model regardless of the source directions.

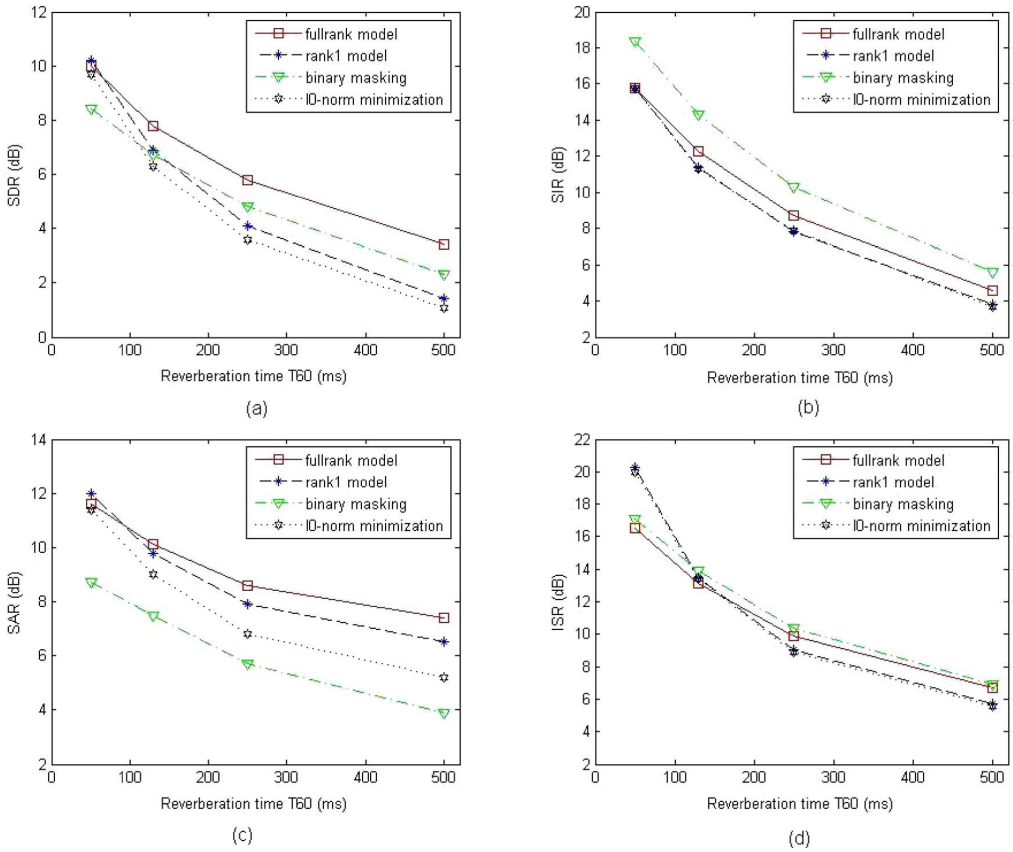

Fig. 4. Average blind source separation performance over stereo mixtures of three sources as a function of the reverberation time, measured in terms of (a) SDR, (b) SIR, (c) SAR, and (d) ISR.

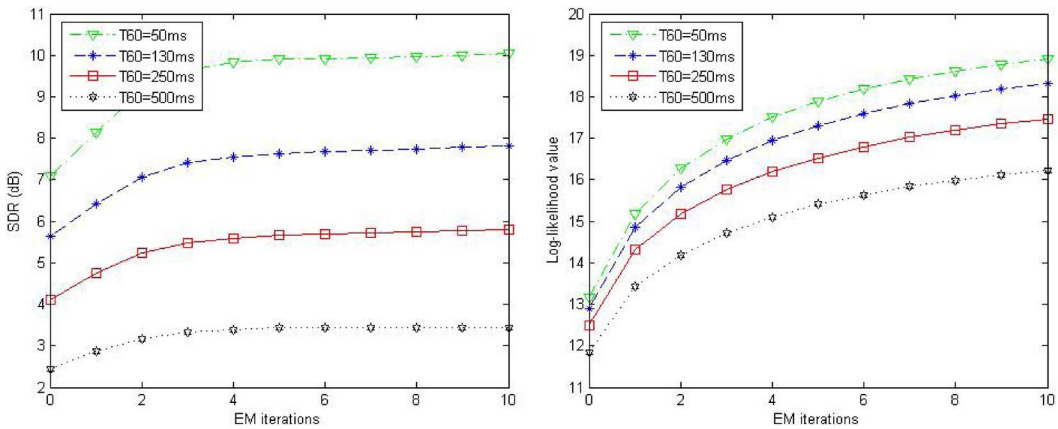

Fig. 5. Convergence properties of EM iteration for the full-rank unconstrained model.

E. Blind Source Separation With the SiSEC 2008 Test Data

We conducted a fourth experiment to compare the proposed full-rank unconstrained model-based algorithm with state-of-the-art BSS algorithms submitted for evaluation to SiSEC 2008 over real-world mixtures of three or four speech sources. Two mixtures were recorded for each given number of sources, using either male or female speech signals. The room reverberation time was either 130 or 250 ms and the microphone spacing 5 cm [5]. The average SDR achieved by each algorithm is listed in Table III for comparison since it provides the overall distortion of the system. The SDR results of all algorithms besides the proposed full-rank unconstrained model-based algorithm were taken from the website of SiSEC 2008,[2] except for Izumi's algorithm [15] whose results were provided by its author.

For three-source mixtures, the proposed algorithm provides 0.4 dB and 0.1 dB SDR improvement compared to the best current results given by Araki's algorithm [24] with $T_{60} = 130$ ms and $T_{60} = 250$ ms, respectively. For four-source mixtures, it

[2]http://sisec2008.wiki.irisa.fr/tiki-index.php?page=Under-determined+speech+and+music+mixtures

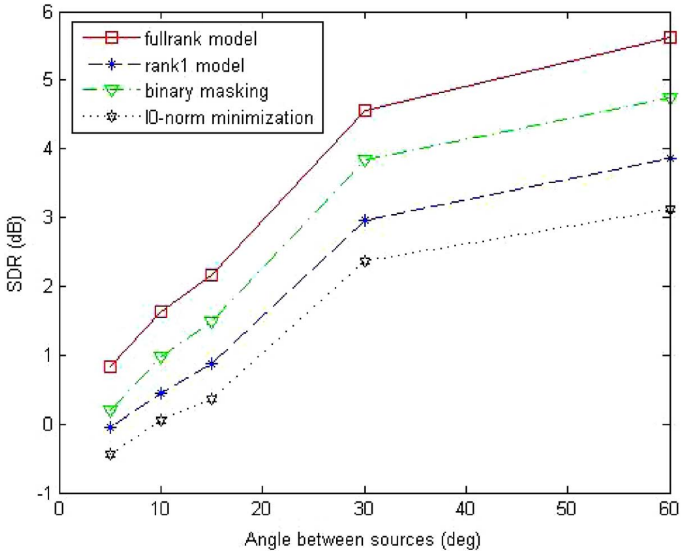

Fig. 6. Average blind source separation performance over stereo mixtures of three sources as a function of the DOA difference between sources.

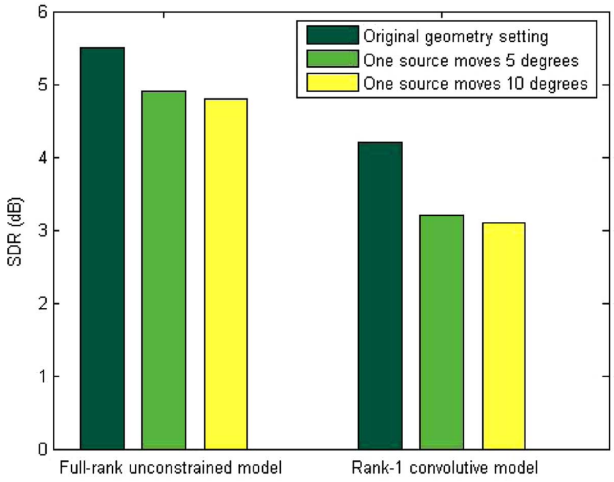

Fig. 7. SDR results in the small source movement scenarios.

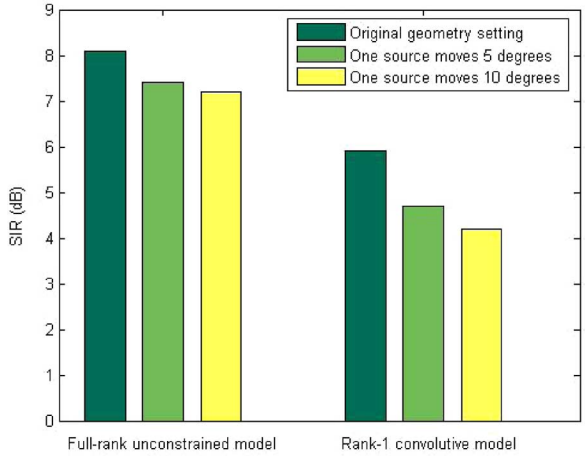

Fig. 8. SIR results in the small source movement scenarios.

TABLE III
AVERAGE SDR OVER THE REAL-WORLD TEST DATA OF SISEC 2008 WITH 5-cm MICROPHONE SPACING

T_{60}	Algorithms	3 source mixtures	4 source mixtures
130 ms	full-rank unconstrained	3.3	2.8
	M. Cobos [21]	2.3	2.1
	M. Mandel [22]	0.1	-3.7
	R. Weiss [23]	2.9	2.3
	S. Araki [24]	2.9	-
	Z. El Chami [25]	2.3	2.1
250 ms	full-rank unconstrained	3.8	2.0
	M. Cobos [21]	2.2	1.0
	M. Mandel [22]	0.8	1.0
	R. Weiss [23]	2.3	1.5
	S. Araki [24]	3.7	-
	Y. Izumi [15]	-	1.6
	Z. El Chami [25]	3.1	1.4

provides even higher SDR improvements of 0.5 and 0.4 dB, respectively compared to the best current results given by Weiss's [23] and Izumi's algorithms [15]. More detailed comparison (not shown in the Table) indicates that the proposed algorithm also outperforms most others in terms of SIR, SAR, and ISR. For instance, it achieves higher SIR than all other algorithms on average except Weiss's. Compared to Weiss's, it achieves the same average SIR but a higher SAR.

F. Investigation of the Robustness to Small Source Movements

Our last experiment aims to examine the robustness of the rank-1 convolutive model and the full-rank unconstrained model to small source movements. We made several recordings of three speech sources s_1, s_2, s_3 in a meeting room with 250 ms reverberation time using omnidirectional microphones spaced by 5 cm. The distance from the sources to the microphones was 50 cm. For each recording, the spatial images of all sources were separately recorded and then added together to obtain a test mixture. After the first recording, we kept the same positions for s_1 and s_2 and successively moved s_3 by 5° and 10° both clockwise and counterclockwise resulting in four new positions of s_3. We then applied the same procedure to s_2 while the positions of s_1 and s_3 remained identical to those in the first recording. Overall, we collected nine mixtures: one from the first recording, four mixtures with 5° movement of either s_2 or s_3, and four mixtures with 10° movement of either s_2 or s_3. We performed source separation in a semi-blind setting: the source spatial covariance matrices were estimated from the spatial images of all sources recorded in the first recording while the source variances were estimated from the nine mixtures using the same algorithm as in Section IV-B. The average SDR and SIR obtained for the first mixture and for the mixtures with 5° and 10° source movement are depicted in Figs. 7 and 8, respectively. This procedure simulates errors encountered by online source separation algorithms in moving source environments, where the source separation parameters learnt at a given time are not applicable anymore at a later time.

The separation performance of the rank-1 convolutive model degrades more than that of the full-rank unconstrained model both with 5° and 10° source rotation. For instance, the SDR

drops by 0.6 dB for the full-rank unconstrained model based algorithm when a source moves by 5° while the corresponding drop for the rank-1 convolutive model equals 1 dB. This result can be explained when considering the fact that the full-rank model accounts for the spatial spread of each source as well as its spatial direction. Therefore, small source movements remaining in the range of the spatial spread do not affect much separation performance. This result indicates that, besides its numerous advantages presented in the previous experiments, this model could also offer a promising approach to the separation of moving sources due to its greater robustness to parameter estimation errors.

V. Conclusion and Discussion

In this paper, we presented a general probabilistic framework for convolutive source separation based on the notion of spatial covariance matrix. We proposed four specific models, including rank-1 models based on the narrowband approximation and full-rank models that overcome this approximation, and derived an efficient algorithm to estimate their parameters from the mixture. Experimental results indicate that the proposed full-rank unconstrained spatial covariance model better accounts for reverberation and therefore improves separation performance compared to rank-1 models and state-of-the-art algorithms in realistic reverberant environments.

Let us now mention several further research directions. Short-term work will be dedicated to the application of the full-rank unconstrained model to the modeling and separation of diffuse and semi-diffuse sources or background noise. Contrary to the rank-1 model in [12] which involves an explicit spatially uncorrelated noise component, this model implicitly represents noise as any other source and can account for multiple noise sources as well as spatially correlated noises with various spatial spreads. We also aim to complete the probabilistic framework by defining a prior distribution for the model parameters across all frequency bins so as to improve the robustness of parameter estimation with small amounts of data and to address the permutation problem in a probabilistically relevant fashion. Finally, a promising way to improve source separation performance is to combine the spatial covariance models investigated in this paper with models of the source spectra such as Gaussian mixture models [18] or nonnegative matrix factorization [12].

Acknowledgment

The authors would like to thank Y. Izumi for providing the results of his algorithm [15] on some test data.

References

[1] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined convolutive blind source separation using spatial covariance models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Mar. 2010.

[2] O. Yılmaz and S. T. Rickard, "Blind separation of speech mixtures via time–frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[3] 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 J. Adv. Signal Process.*, vol. 2007, 2007, article ID 24717.

[4] P. Bofill, "Underdetermined blind separation of delayed sound sources in the frequency domain," *Neurocomputing*, vol. 55, no. 3-4, pp. 627–641, 2003.

[5] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," in *Proc. Int. Conf. Ind. Compon. Anal. Signal Separat. (ICA)*, 2009, pp. 734–741.

[6] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, "Probabilistic modeling paradigms for audio source separation," in *Machine Audition: Principles, Algorithms and Systems*, W. Wang, Ed. Hershey, PA: IGI Global, 2010.

[7] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Spatial covariance models for under-determined reverberant audio source separation," in *Proc. 2009 IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, 2009, pp. 129–132.

[8] C. Févotte and J.-F. Cardoso, "Maximum likelihood approach for blind audio source separation using time–frequency Gaussian models," in *Proc. 2005 IEEE Workshop Applicat.Signal Process. Audio Acoust. (WASPAA)*, 2005, pp. 78–81.

[9] E. Vincent, S. Arberet, and R. Gribonval, "Underdetermined instantaneous audio source separation via local Gaussian modeling," in *Proc. Int. Conf. Ind. Compon. Anal. Signal Separat. (ICA)*, 2009, pp. 775–782.

[10] A. P. Dempster, N. M. Laird, and B. D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc.*, ser. B, vol. 39, pp. 1–38, 1977.

[11] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1592–1604, Jul. 2007.

[12] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 550–563, 2010.

[13] T. Gustafsson, B. D. Rao, and M. Trivedi, "Source localization in reverberant environments: Modeling and statistical analysis," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 791–803, Nov. 2003.

[14] H. Kuttruff, *Room Acoustics*, 4th ed. New York: Spon, 2000.

[15] Y. Izumi, N. Ono, and S. Sagayama, "Sparseness-based 2CH BSS using the EM algorithm in reverberant environment," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, 2007, pp. 147–150.

[16] B. D. van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[17] J.-F. Cardoso, H. Snoussi, J. Delabrouille, and G. Patanchon, "Blind separation of noisy Gaussian stationary sources. Application to cosmic microwave background imaging," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2002, vol. 1, pp. 561–564.

[18] S. Arberet, A. Ozerov, R. Gribonval, and F. Bimbot, "Blind spectral—GMM estimation for underdetermined instantaneous audio source separation," in *Proc. Int. Conf. Ind. Compon. Anal. Signal Separat. (ICA)*, 2009, pp. 751–758.

[19] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. New York: Wiley, 1997.

[20] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. Rosca, "First stereo audio source separation evaluation campaign: Data, algorithms and results," in *Proc. Int. Conf. Ind. Compon. Anal. Signal Separat. (ICA)*, 2007, pp. 552–559.

[21] M. Cobos and J. López, "Blind separation of underdetermined speech mixtures based on DOA segmentation," *IEEE Trans. Audio, Speech, Lang. Process.*, submitted for publication.

[22] M. Mandel and D. Ellis, "EM localization and separation using interaural level and phase cues," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, 2007, pp. 275–278.

[23] R. Weiss and D. Ellis, "Speech separation using speaker-adapted eigenvoice speech models," *Comput. Speech Lang.*, vol. 24, no. 1, pp. 16–20, Jan. 2010.

[24] S. Araki, T. Nakatani, H. Sawada, and S. Makino, "Stereo source separation and source counting with MAP estimation with Dirichlet prior considering spatial aliasing problem," in *Proc. Int. Conf. Ind. Compon. Anal. Signal Separat. (ICA)*, 2009, pp. 742–750.

[25] Z. El Chami, D. T. Pham, C. Servière, and A. Guerin, "A new model based underdetermined source separation," in *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC)*, 2008.

Ngoc Q. K. Duong received the B.S. degree from Posts and Telecommunications Institute of Technology (PTIT), Ha Noi City, Vietnam, in 2004, and the M.S. degree in electronic engineering from Paichai University, Daejeon, Korea, in 2008. He is currently pursuing the Ph.D. degree at the French National Institute for Research in Computer Science and Control (INRIA), Rennes, France.

From 2004 to 2006, he was with Visco JSC as a System Engineer for the audio/video conferencing system. He was also a Research Engineer for the acoustic echo/noise cancelation system at Emersys Company, Korea in 2008. His current research interest concerns adaptive spectral and spatial covariance models for blind source separation.

Emmanuel Vincent (M'07–SM'10) received the mathematics degree from the École Normale Supérieure, Paris, France, in 2001 and the Ph.D. degree in acoustics, signal processing, and computer science applied to music from the University of Paris-VI Pierre and Marie Curie, Paris, in 2004.

From 2004 to 2006, he was a Research Assistant with the Centre for Digital Music at Queen Mary, University of London, London, U.K. He is now a Permanent Researcher with the French National Institute for Research in Computer Science and Control (INRIA), Rennes, France. His research focuses on probabilistic modeling of audio signals applied to blind source separation, indexing, and object coding of musical audio.

Rémi Gribonval (M'01–SM'08) graduated from École Normale Supérieure, Paris, France in 1997 and received the Ph.D. degree in applied mathematics from the University of Paris-IX Dauphine, Paris, in 1999, and the Habilitation à Diriger des Recherches in applied mathematics from the University of Rennes I, Rennes, France, in 2007.

From 1999 until 2001, he was a Visiting Scholar at the Industrial Mathematics Institute (IMI), Department of Mathematics, University of South Carolina, Columbia. He is now a Senior Research Scientist (Directeur de Recherche) with INRIA (the French National Center for Computer Science and Control) at IRISA, Rennes, France, in the METISS group. His research focuses on sparse approximation, mathematical signal processing, and applications to multichannel audio signal processing, with a particular emphasis in blind audio source separation and compressed sensing. Since 2002, he has been the coordinator of several national, bilateral, and European research projects, and in 2008 he was elected a member of the steering committee for the international conference ICA on independent component analysis and source separation.