Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation

Alexey Ozerov, *Member, IEEE*, and Cédric Févotte, *Member, IEEE*

***Abstract*—We consider inference in a general data-driven object-based model of multichannel audio data, assumed generated as a possibly underdetermined convolutive mixture of source signals. We work in the short-time Fourier transform (STFT) domain, where convolution is routinely approximated as linear instantaneous mixing in each frequency band. Each source STFT is given a model inspired from nonnegative matrix factorization (NMF) with the Itakura–Saito divergence, which underlies a statistical model of superimposed Gaussian components. We address estimation of the mixing and source parameters using two methods. The first one consists of maximizing the exact joint likelihood of the multichannel data using an expectation-maximization (EM) algorithm. The second method consists of maximizing the sum of individual likelihoods of all channels using a multiplicative update algorithm inspired from NMF methodology. Our decomposition algorithms are applied to stereo audio source separation in various settings, covering blind and supervised separation, music and speech sources, synthetic instantaneous and convolutive mixtures, as well as professionally produced music recordings. Our EM method produces competitive results with respect to state-of-the-art as illustrated on two tasks from the international Signal Separation Evaluation Campaign (SiSEC 2008).**

***Index Terms*—Expectation-maximization (EM) algorithm, multichannel audio, nonnegative matrix factorization (NMF), nonnegative tensor factorization (NTF), underdetermined convolutive blind source separation (BSS).**

I. Introduction

NONNEGATIVE matrix factorization (NMF) is an unsupervised data decomposition technique with effervescent popularity in the fields of machine learning and signal/image processing [1]. Much research about this topic has been driven by applications in audio, where the data matrix is taken as the magnitude or power spectrogram of a sound signal. NMF was for example applied with success to automatic music transcription [2], [3] and audio source separation [4], [5]. The factorization amounts to decomposing the spectrogram data into a sum of rank-1 spectrograms, each of which being the expression of an elementary spectral pattern amplitude-modulated in time. However, while most music recordings are available in multichannel format (typically, stereo), NMF in its standard setting is only suited to single-channel data. Extensions to multichannel data have been considered, either by stacking up the spectrograms of each channel into a single matrix [6] or by considering nonnegative tensor factorization (NTF) under a parallel factor analysis (PARAFAC) structure, where the channel spectrograms form the slices of a 3-valence tensor [7]. These approaches inherently assume that the original sources have been mixed instantaneously, which in modern music mixing is not realistic, and they require a posterior binding step so as to group the elementary components into instrumental sources. Furthermore they do not exploit the redundancy between the channels in an optimal way, as will be shown later.

The aim of this work is to remedy these drawbacks. We formulate a multichannel NMF model that accounts for convolutive mixing. The source spectrograms are modeled through NMF and the mixing filters serve to identify the elementary components pertaining to each source. We consider more precisely I sampled signals $\tilde{x}_i(t)$ $(i = 1,\ldots,I,\ t = 1,\ldots,T)$ generated as convolutive noisy mixtures of J point source signals $\tilde{s}_j(t)$ $(i = 1,\ldots,J)$ such that

$$\tilde{x}_i(t) = \sum_{j=1}^{J} \sum_{\tau=0}^{L-1} \tilde{a}_{ij}(\tau)\tilde{s}_j(t-\tau) + \tilde{b}_i(t) \tag{1}$$

where $\tilde{a}_{ij}(\tau)$ is the finite-impulse response of some (causal) filter and $\tilde{b}_i(t)$ is some additive noise. The time-domain mixing given by (1) can be approximated in the short-time Fourier transform (STFT) domain as

$$x_{i,fn} = \sum_{j=1}^{J} a_{ij,f} s_{j,fn} + b_{i,fn} \tag{2}$$

where $x_{i,fn}$, $s_{j,fn}$ and $b_{i,fn}$ are the complex-valued STFTs of the corresponding time signals, $a_{ij,f}$ is the complex-valued discrete Fourier transform of filter $\tilde{a}_{ij}(\tau)$, $f = 1,\ldots,F$ is a frequency bin index, and $n = 1,\ldots,N$ is a time frame index. Equation (2) holds when the filter length L is assumed "significantly" shorter than the STFT window size $(2F-2)$ [8]. Equation (2) can be rewritten in matrix form, such that

$$\mathbf{x}_{fn} = \mathbf{A}_f \mathbf{s}_{fn} + \mathbf{b}_{fn} \tag{3}$$

where $\mathbf{x}_{fn} = [x_{1,fn},\ldots,x_{I,fn}]^T$, $\mathbf{s}_{fn} = [s_{1,fn},\ldots,s_{J,fn}]^T$, $\mathbf{b}_{fn} = [b_{1,fn},\ldots,b_{I,fn}]^T$, and $\mathbf{A}_f = [a_{ij,f}]_{ij} \in \mathbb{C}^{I\times J}$.

Manuscript received December 24, 2008; revised August 17, 2009. Current version published February 10, 2010. This work was supported in part by the French ANR project SARAH (StAndardisation du Remastering Audio Haute-Définition). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Paris Smaragdis.

A. Ozerov was with the Institut Telecom, Telecom ParisTech, CNRS LTCI, 75014 Paris, France. He is now with the METISS Team of IRISA/INRIA, 35042 Rennes Cedex, France (e-mail: alexey.ozerov@irisa.fr).

C. Févotte is with CNRS LTCI, Telecom ParisTech, 75014 Paris, France (e-mail: cedric.fevotte@telecom-paristech.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.2009.2031510

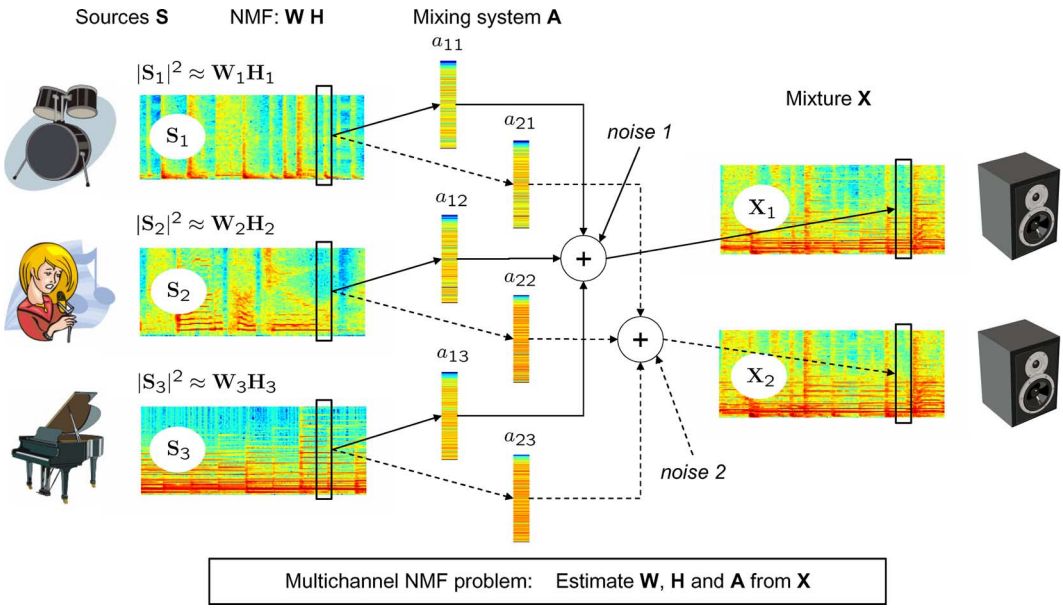

Fig. 1. Representation of convolutive mixing system and formulation of Multichannel NMF problem.

A key ingredient of this work is to model the $F \times N$ power spectrogram $|\mathbf{S}_j|^2 = [|s_{j,fn}|^2]_{fn}$ of source j as a product of two nonnegative matrices $\mathbf{W}_j$ and $\mathbf{H}_j$, such that

$$|\mathbf{S}_j|^2 \approx \mathbf{W}_j\mathbf{H}_j. \tag{4}$$

Given the observed mixture STFTs $\mathbf{X} = \{x_{i,fn}\}_{ifn}$, we are interested in joint estimating the source spectrogram factors $\{\mathbf{W}_j, \mathbf{H}_j\}_j$ and the mixing system $\{\mathbf{A}_f\}_f$, as illustrated in Fig. 1. Our problem splits into two subtasks: 1) defining suitable estimation criteria, and 2) designing algorithms optimizing these criteria.

We adopt a statistical setting in which each source STFT is modeled as a sum of latent Gaussian components, a model introduced by Benaroya *et al.* [9] in a supervised single-channel audio source separation context. A connection between full maximum-likelihood (ML) estimation of the variance parameters in this model and NMF using the Itakura–Saito (IS) divergence was pointed out in [10]. Given this source model, hereafter referred to as *NMF model*, we introduce two estimation criteria together with corresponding inference methods.

- The first method consists of maximizing the exact joint log-likelihood of the multichannel data using an expectation-maximization (EM) algorithm [11]. This method fully exploits the redundancy between the channels, in a statistically optimal way. It draws parallels with several model-based multichannel source separation methods [12]–[18], as described throughout the paper.
- The second method consists of maximizing the sum of individual log-likelihoods of all channels using a multiplicative update (MU) algorithm inspired from NMF methodology. This approach relates to the above-mentioned NTF techniques [6], [7]. However, in contrast to standard NTF which inherently assumes instantaneous mixing, our approach addresses a more general convolutive structure and does not require the posterior binding of the elementary components into J sources.

The general multichannel NMF framework we describe yields a data-driven object-based representation of multichannel data that may benefit many tasks in audio, such as transcription or object-based coding. In this article we will more specifically focus on the convolutive blind source separation (BSS) problem, and as such we also address means of reconstructing source signal estimates from the set of estimated parameters. Our decompositions are conservative in the sense that the spatial source estimates sum up to the original mix. The mixing parameters may also be changed without degrading audio quality, so that music remastering is one potential application of our work. Remixes of well-known songs retrieved from commercial CD recordings are proposed in the results section.

Many convolutive BSS methods have been designed under model (3). Typically, an instantaneous independent component analysis (ICA) algorithm is applied to data $\{\mathbf{x}_{fn}\}_{n=1,\ldots,N}$ in each frequency subband f, yielding a set of J source subband estimates per frequency bin. This approach is usually referred to as frequency-domain ICA (FD-ICA) [19]. The source labels remain however unknown because of the ICA standard permutation indeterminacy, leading to the well-known FD-ICA permutation alignment problem, which cannot be solved without using additional *a priori* knowledge about the sources and/or about the mixing filters. For example in [20] the sources in different frequency bins are grouped *a posteriori* relying on their temporal correlation, thus using prior knowledge about the sources, and in [21], [22] the sources and the filters are estimated assuming a particular structure of convolutive filters, i.e., using prior knowledge about the filters. The permutation ambiguity arises from the individual processing of each subband, which implicitly assumes mutual independence of one source's subbands. This is not the case in our work where our source model implies a coupling of the frequency bands, and joint estimation of the source

parameters and mixing coefficients frees us from the permutation alignment problem.

Our EM-based method is related to some multichannel source separation techniques employing Gaussian mixture models (GMMs) as source models. Univariate independent and identically distributed (i.i.d.) GMMs have been used to model source samples in the time domain for separation of instantaneous [12], [13] and convolutive [12] mixtures. However, such time-domain GMMs are not of the most relevance for audio as they do not model temporal correlations in the signal. In [14], Attias proposes to model the sources in the STFT domain using multivariate GMMs, hence taking into account temporal correlations in the audio signal, assumed stationary in each window frame. The author develops a source separation method for convolutive mixtures, supervised in the sense that the source models are pre-trained in advance. A similar approach with log-spectral domain GMMs is developed by Weiss *et al.* in [15]. Arberet *et al.* [16] propose a multivariate GMM-based separation method for instantaneous mixing that involves a computationally efficient strategy for learning the source GMMs separately, using intermediate source estimates obtained by some BSS method. As compared to these works, we use a different source model (the NMF model), which might be considered more suitable than the GMM for musical signals. Indeed, the NMF is well suited to polyphony as it basically takes the source to be a sum of elementary components with characteristic spectral signatures. In contrast, the GMM takes the source as a single component with many states, each representative of a characteristic spectral signature, but not mixed *per se*. To put it in an other way, in the NMF model a summation occurs in the STFT domain (or equivalently, in the time domain), while in the GMM the summation occurs on the distribution of the frames. Moreover, as discussed later, the computational complexity of inference in our model grows linearly with the number of components while the complexity of standard inference in GMMs grows combinatorially.

The remaining of this paper is organized as follows. NMF source model and noise model are introduced in Section II. Section III is devoted to the definition of our two estimation criteria, with corresponding optimization algorithms. Section IV presents results of our methods to stereo source separation in various settings, including blind and supervised separation of music and speech sources in synthetic instantaneous and convolutive mixtures, as well as in professionally produced music recordings. Conclusions are drawn in Section V. Preliminary aspects of this work are presented in [23]. We here considerably extend on the simulations part as well as on the theoretical developments related to our algorithms.

II. Models

A. Sources

Let $K \geq J$ and $\{\mathcal{K}_j\}_{j=1}^{J}$ be a nontrivial partition of $\mathcal{K} = \{1, \ldots, K\}$. Following [9], [10], we assume the complex random variable $s_{j,fn}$ to be a sum of $\#\mathcal{K}_j$ latent components, such that

$$s_{j,fn} = \sum_{k \in \mathcal{K}_j} c_{k,fn} \quad \text{with} \quad c_{k,fn} \sim \mathcal{N}_c(0, w_{fk}h_{kn}) \tag{5}$$

where $w_{fk}, h_{kn} \in \mathbb{R}^+$ and $\mathcal{N}_c(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the *proper* complex Gaussian distribution [24] with probability density function (pdf)

$$N_c(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{|\pi\boldsymbol{\Sigma}|} \exp\left[-(\mathbf{x}-\boldsymbol{\mu})^H \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]. \tag{6}$$

In the rest of the paper, the quantities $s_{j,fn}$ and $c_{k,fn}$ are, respectively, referred to as "source" and "component". The components are assumed *mutually* independent and *individually* independent across frequency f and frame n. It follows that

$$s_{j,fn} \sim \mathcal{N}_c\left(0, \sum_{k \in \mathcal{K}_j} w_{fk}h_{kn}\right). \tag{7}$$

Denoting $\mathbf{S}_j$ the $F \times N$ STFT matrix $[s_{j,fn}]_{fn}$ of source j and introducing the matrices $\mathbf{W}_j = [w_{fk}]_{f,k \in \mathcal{K}_j}$ and $\mathbf{H}_j = [h_{kn}]_{k \in \mathcal{K}_j, n}$, respectively, of dimensions $F \times \#\mathcal{K}_j$ and $\#\mathcal{K}_j \times N$, it is easily shown [10] that the minus log-likelihood of the parameters describing source j writes

$$-\log p(\mathbf{S}_j | \mathbf{W}_j, \mathbf{H}_j) \stackrel{c}{=} \sum_{fn} d_{IS}\left(|s_{j,fn}|^2 | [\mathbf{W}_j\mathbf{H}_j]_{fn}\right)$$

where "$\stackrel{c}{=}$" denotes equality up to a constant and

$$d_{IS}(x|y) = \frac{x}{y} - \log\frac{x}{y} - 1 \tag{8}$$

is the IS divergence. In other words, ML estimation of $\mathbf{W}_j$ and $\mathbf{H}_j$ given source STFT $\mathbf{S}_j$ is equivalent to NMF of the power spectrogram $|\mathbf{S}_j|^2$ into $\mathbf{W}_j\mathbf{H}_j$, where the IS divergence is used. MU and EM algorithms for IS-NMF are, respectively, described in [25], [26] and in [10]; in essence, this paper describes a generalization of these algorithms to a multichannel multisource scenario. In the following, we will use the notation $\mathbf{P}_j = \mathbf{W}_j\mathbf{H}_j$, i.e., $p_{j,fn} = \mathbb{E}\{|s_{j,fn}|^2\}$.

Our source model is related to the GMM used for example in [14], [16] in the same source separation context, with the difference that one source frame is here modeled as a sum of $\#\mathcal{K}_j$ elementary components while in the GMM one source frame is modeled as a process which can take one of many states, each characterized by a covariance matrix. The computational complexity of inference in our model with our algorithms described next grows linearly with the total number of components while the derivation of the equivalent EM algorithm for GMM leads to an algorithm that has combinatorial complexity with the number of states [12], [13], [15]. It is possible to achieve linear complexity in the GMM case also, but at the price of approximate inference [14], [16]. Note that all considered algorithms, either for the NMF model or GMM, only ensure convergence to a stationary point of the objective function, and, as a consequence, the final result depends strongly on the parameters initialization. We wish to emphasize that we here take a fully data-driven approach in the sense that no parameter is pre-trained.

B. Noise

In the most general case, we may assume noisy data and the following algorithms can easily accommodate estimation of noise statistics under Gaussian independent assumptions and given covariance structures such as $\boldsymbol{\Sigma}_{\mathbf{b},fn} = \boldsymbol{\Sigma}_{\mathbf{b},f}$ or $\boldsymbol{\Sigma}_{\mathbf{b},n}$. In

this paper, we consider for simplicity stationary and spatially uncorrelated noise such that

$$b_{i,fn} \sim \mathcal{N}_c\left(0, \sigma^2_{i,f}\right) \tag{9}$$

and $\mathbf{\Sigma}_{\mathbf{b},f} = \mathrm{diag}([\sigma^2_{i,f}]_i)$. The musical data we consider in Section IV-A is not noisy in the usual sense, but the noise component can account for model discrepancy and/or quantization noise. Moreover, this noise component is required in the EM algorithm to prevent from potential numerical instabilities (see Section III-A1 below) and slow convergence (see Section III-A6 below). In Section IV-D, we will consider several scenarios: when the variances are equal and fixed to a small value $\tilde{\sigma}^2$, when the variances are estimated from data, and most importantly when annealing is performed via the noise variance, so as to speed up convergence as well as favor global solutions.

C. Convolutive Mixing Model Revisited

With (5), the mixing model (3) can be recast as

$$\mathbf{x}_{fn} = \mathring{\mathbf{A}}_f \mathbf{c}_{fn} + \mathbf{b}_{fn} \tag{10}$$

where $\mathbf{c}_{fn} = [c_{1,fn}, \ldots, c_{K,fn}]^T \in \mathbb{C}^{K\times 1}$ and $\mathring{\mathbf{A}}_f$ is the so called "augmented mixing matrix" of dimension $I \times K$, with elements defined by $\mathring{a}_{ik,f} = a_{ij,f}$ if and only if $k \in \mathcal{K}_j$. Thus, for every frequency bin f, our model is basically a linear mixing model with I channels and K elementary Gaussian sources $c_{k,fn}$, with structured mixing coefficients (i.e., subsets of elementary sources are mixed identically). Subsequently, we will note $\mathring{\mathbf{\Sigma}}_{\mathbf{c},fn} = \mathrm{diag}([w_{fk}h_{kn}]_k)$ the covariance of $\mathbf{c}_{fn}$.

III. Methods

A. Maximization of Exact Likelihood With EM

1) Criterion: Let $\boldsymbol{\theta} = \{\mathbf{A}, \mathbf{W}, \mathbf{H}, \mathbf{\Sigma}_\mathbf{b}\}$ be the set of all parameters, where $\mathbf{A}$ is the $I \times J \times F$ tensor with entries $a_{ij,f}$, $\mathbf{W}$ is the $F \times K$ matrix with entries w_{fk}, $\mathbf{H}$ is the $K \times N$ matrix with entries h_{kn}, and $\mathbf{\Sigma}_\mathbf{b}$ are the noise covariance parameters. Under previous assumptions, data vector $\mathbf{x}_{fn}$ has a zero-mean proper Gaussian distribution with covariance

$$\mathbf{\Sigma}_{\mathbf{x},fn}(\boldsymbol{\theta}) = \mathbf{A}_f \mathbf{\Sigma}_{\mathbf{s},fn} \mathbf{A}_f^H + \mathbf{\Sigma}_{\mathbf{b},f} \tag{11}$$

where $\mathbf{\Sigma}_{\mathbf{s},fn} = \mathrm{diag}([p_{j,fn}]_j)$ is the covariance of $\mathbf{s}_{fn}$. ML estimation is consequently shown to amount to minimization of

$$C_1(\boldsymbol{\theta}) = \sum_{fn} \mathrm{trace}\left(\mathbf{x}_{fn}\mathbf{x}_{fn}^H \mathbf{\Sigma}^{-1}_{\mathbf{x},fn}\right) + \log\det \mathbf{\Sigma}_{\mathbf{x},fn}. \tag{12}$$

The noise covariance term $\mathbf{\Sigma}_{\mathbf{b},f}$ appears necessary so as to prevent from ill-conditioned inverses that occur if 1) $\mathrm{rank}(\mathbf{A}_f) < I$, and in particular if $I > J$, i.e., in the overdetermined case, or if 2) $\mathbf{\Sigma}_{\mathbf{s},fn}$ has more than $(J - I)$ null diagonal coefficients in the underdetermined case $(I < J)$. Case 2) might happen in regions of the time–frequency plane where sources are inactive.

For fixed f and n, the BSS problem described by (3) and (12), and the following EM algorithm, is reminiscent of works by Cardoso *et al.*, see, e.g., [27] for the square noise-free case, [17] for other cases and [18] for use in an audio setting. In these papers, a grid of the representation domain is chosen, in each cell of which the source statistics are assumed constant. This is not required in our case where we instead solve F parallel linear instantaneous mixtures tied across frequency by the source model.[1]

2) Indeterminacies: Criterion (12) suffers from obvious scale, phase and permutation indeterminacies.[2] Regarding scale and phase, let $\hat{\boldsymbol{\theta}} = \{\{\mathbf{A}_f\}_f, \{\mathbf{W}_j, \mathbf{H}_j\}_j\}$ be a minimizer of (12) and let $\{\mathbf{D}_f\}_f$ and $\{\mathbf{\Lambda}_j\}_j$ be sets of respectively *complex* and *nonnegative* diagonal matrices. Then, the set

$$\tilde{\boldsymbol{\theta}} = \left\{\left\{\mathbf{A}_f\mathbf{D}_f^{-1}\right\}_f, \left\{\mathrm{diag}\left(\left[|d_{jj,f}|^2\right]_f\right)\mathbf{W}_j\mathbf{\Lambda}_j^{-1}\right\}_j, \{\mathbf{\Lambda}_j\mathbf{H}_j\}_j\right\}$$

leads to $\mathbf{\Sigma}_{\mathbf{x},fn}(\hat{\boldsymbol{\theta}}) = \mathbf{\Sigma}_{\mathbf{x},fn}(\tilde{\boldsymbol{\theta}})$, hence same likelihood value. Similarly, permuted diagonal matrices would also leave the criterion unchanged. In practice, we remove the scale and phase ambiguity by imposing $\sum_i |a_{ij,f}|^2 = 1$ and $a_{1j,f} \in \mathbb{R}^+$ (and scaling the rows of $\mathbf{W}_j$ accordingly) and then by imposing $\sum_f w_{fk} = 1$ (and scaling the rows of $\mathbf{H}_j$ accordingly). With these conventions, the columns of $\mathbf{A}_f$ convey normalized mixing proportions between the channels, the columns of $\mathbf{W}$ convey normalized frequency shapes and all time-dependent amplitude information is relegated into $\mathbf{H}$.

3) Algorithm: We derive an EM algorithm based on *complete data* $\{\mathbf{X}, \mathbf{C}\}$, where $\mathbf{C}$ is the $K \times F \times N$ STFT tensor with coefficients $c_{k,fn}$. The complete data pdfs $\{p(\mathbf{X}, \mathbf{C}|\boldsymbol{\theta})\}_{\boldsymbol{\theta}}$ form an *exponential family* (see, e.g., [11] or [29, Appendix]) and the set $\{\mathbf{R}_{\mathbf{xx},f}, \mathbf{R}_{\mathbf{xs},f}, \mathbf{R}_{\mathbf{ss},f}, \{u_{k,fn}\}_{kn}\}_f$ defined by

$$\mathbf{R}_{\mathbf{xx},f} = \frac{1}{N}\sum_n \mathbf{x}_{fn}\mathbf{x}_{fn}^H, \quad \mathbf{R}_{\mathbf{xs},f} = \frac{1}{N}\sum_n \mathbf{x}_{fn}\mathbf{s}_{fn}^H \tag{13}$$

$$\mathbf{R}_{\mathbf{ss},f} = \frac{1}{N}\sum_n \mathbf{s}_{fn}\mathbf{s}_{fn}^H, \quad u_{k,fn} = |c_{k,fn}|^2 \tag{14}$$

is shown to be a *natural (sufficient) statistics* [29] for this family. Thus, one iteration of EM consists of computing the expectation of the natural statistics conditionally on the current parameter estimates (E step) and of reestimating the parameters using the updated natural statistics, which amounts to maximizing the conditional expectation of the complete data log-likelihood $Q(\boldsymbol{\theta}|\boldsymbol{\theta}') = \int[\log p(\mathbf{X}, \mathbf{C}|\boldsymbol{\theta})]p(\mathbf{C}|\mathbf{X}, \boldsymbol{\theta}')d\mathbf{C}$ (M step). The resulting updates are given in Algorithm 1, with more details given in Appendix A.

Algorithm 1 EM algorithm (one iteration)

- **E step**. Conditional expectations of natural statistics:

$$\widehat{\mathbf{R}}_{\mathbf{xx},f} = \mathbf{R}_{\mathbf{xx},f} = \frac{1}{N}\sum_n \mathbf{x}_{fn}\mathbf{x}_{fn}^H, \tag{15}$$

$$\widehat{\mathbf{R}}_{\mathbf{xs},f} = \frac{1}{N}\sum_n \mathbf{x}_{fn}\hat{\mathbf{s}}_{fn}^H, \tag{16}$$

$$\widehat{\mathbf{R}}_{\mathbf{ss},f} = \frac{1}{N}\sum_n \hat{\mathbf{s}}_{fn}\hat{\mathbf{s}}_{fn}^H + \mathbf{\Sigma}_{\mathbf{s},fn} - \mathbf{G}_{\mathbf{s},fn}\mathbf{A}_f\mathbf{\Sigma}_{\mathbf{s},fn} \tag{17}$$

$$\hat{u}_{k,fn} = \left[\hat{\mathbf{c}}_{fn}\hat{\mathbf{c}}_{fn}^H + \mathring{\mathbf{\Sigma}}_{\mathbf{c},fn} - \mathbf{G}_{\mathbf{c},fn}\mathring{\mathbf{A}}_f\mathring{\mathbf{\Sigma}}_{\mathbf{c},fn}\right]_{k,k} \tag{18}$$

[1]In [17] and [27], the ML criterion can be recast as a measure of fit between observed and parameterized covariances, where the measure of deviation writes $D(\mathbf{\Sigma}_1|\mathbf{\Sigma}_2) = \mathrm{trace}(\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{-1}) - \log\det\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{-1} - I$ and $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are positive definite matrices of size $I \times I$ (note that the IS divergence is obtained in the special case $I = 1$). The measure is simply the KL divergence between the pdfs of two zero-mean Gaussians with covariances $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$. Such a formulation cannot be used in our case because $\mathbf{\Sigma}_1 = \mathbf{x}_{fn}\mathbf{x}_{fn}^H$ is not invertible for $I > 1$.

[2]There might also be other less obvious indeterminacies, such as those inherent to NMF (see, e.g., [28]), but this study is here left aside.

where
$$\hat{\mathbf{s}}_{fn} = \mathbf{G}_{\mathbf{s},fn}\mathbf{x}_{fn}, \quad \mathbf{G}_{\mathbf{s},fn} = \mathbf{\Sigma}_{\mathbf{s},fn}\mathbf{A}_f^H\mathbf{\Sigma}_{\mathbf{x},fn}^{-1} \tag{19}$$
$$\hat{\mathbf{c}}_{fn} = \mathbf{G}_{\mathbf{c},fn}\mathbf{x}_{fn}, \quad \mathbf{G}_{\mathbf{c},fn} = \mathbf{\Sigma}_{\mathbf{c},fn}\mathring{\mathbf{A}}_f^H\mathbf{\Sigma}_{\mathbf{x},fn}^{-1} \tag{20}$$
$$\mathbf{\Sigma}_{\mathbf{x},fn} = \mathbf{A}_f\mathbf{\Sigma}_{\mathbf{s},fn}\mathbf{A}_f^H + \mathbf{\Sigma}_{\mathbf{b},f} \tag{21}$$
$$\mathbf{\Sigma}_{\mathbf{s},fn} = \mathrm{diag}\left(\left[\sum_{k\in\mathcal{K}_j} w_{fk}h_{kn}\right]_j\right) \tag{22}$$
$$\mathbf{\Sigma}_{\mathbf{c},fn} = \mathrm{diag}\left([w_{fk}h_{kn}]_k\right) \tag{23}$$

and $\mathring{\mathbf{A}}_f$ is defined in Section II-C.

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

$$\mathbf{A}_f = \widehat{\mathbf{R}}_{\mathbf{xs},f}\widehat{\mathbf{R}}_{\mathbf{ss},f}^{-1}, \tag{24}$$
$$\mathbf{\Sigma}_{\mathbf{b},f} = \mathrm{diag}\Big(\widehat{\mathbf{R}}_{\mathbf{xx},f} - \mathbf{A}_f\widehat{\mathbf{R}}_{\mathbf{xs},f}^H - \widehat{\mathbf{R}}_{\mathbf{xs},f}\mathbf{A}_f^H + \mathbf{A}_f\widehat{\mathbf{R}}_{\mathbf{ss},f}\mathbf{A}_f^H\Big) \tag{25}$$
$$w_{fk} = \frac{1}{N}\sum_n \frac{\hat{u}_{k,fn}}{h_{kn}}, \quad h_{kn} = \frac{1}{F}\sum_f \frac{\hat{u}_{k,fn}}{w_{fk}}. \tag{26}$$

- Normalize $\mathbf{A}$, $\mathbf{W}$ and $\mathbf{H}$ according to Section III-A2.

4) Implementation Issues: The computation of the source Wiener gain $\mathbf{G}_{\mathbf{s},fn}$ given by (19) requires the inversion of the $I \times I$ matrix $\mathbf{\Sigma}_{\mathbf{x},fn}$ at every time–frequency (TF) point. When $I > J$ (overdetermined case) it may be preferable for sake of computational efficiency to use the following alternative formulation of $\mathbf{G}_{\mathbf{s},fn}$, obtained using Woodbury matrix identity [30]

$$\mathbf{G}_{\mathbf{s},fn} = \mathbf{\Xi}_{\mathbf{s},fn}^{-1}\mathbf{A}_f^H\mathbf{\Sigma}_{\mathbf{b},f}^{-1} \tag{27}$$

with

$$\mathbf{\Xi}_{\mathbf{s},fn} = \mathbf{A}_f^H\mathbf{\Sigma}_{\mathbf{b},f}^{-1}\mathbf{A}_f + \mathbf{\Sigma}_{\mathbf{s},fn}^{-1}. \tag{28}$$

This second formulation requires the inversion of the $J \times J$ matrix $\mathbf{\Xi}_{\mathbf{s},fn}$ instead of the inversion of the $I \times I$ matrix $\mathbf{\Sigma}_{\mathbf{x},fn}$. The same idea applies to the computation of $\mathbf{G}_{\mathbf{c},fn}$, (20), if $I > K$. Thus, this second formulation may become interesting in practice only if $I > J$ and $I > K$, i.e., if $I > K$ (recall that $K \geq J$). As we only consider undetermined mixtures in the experimental part of this article $(I < J)$, we turn to the original formulation given by (19). As we more precisely consider stereo mixtures, we only need inverting 2×2 matrices per TF point and our MATLAB code was efficiently vectorized so as to manipulate time–frequency matrices directly, thanks to Cramer's explicit matrix inversion formula. Note also that we only need to compute the diagonal elements of the $K \times K$ matrix in (18). Hence, the computational complexity of one EM algorithm iteration grows linearly (and not quadratically) with the number of components.

5) Linear Instantaneous Case: Linear instantaneous mixing is a special case of interest, that concerns for example "pan pot" mixing. Here, the mixing matrix is real-valued and shared between all the frequency subbands, i.e., $\mathbf{A}_f = \mathbf{A}_{\mathrm{inst}} \in \mathbb{R}^{I\times J}$. In that case, (24) needs only be replaced by

$$\mathbf{A}_{\mathrm{inst}} = \Re\left\{\sum_f \widehat{\mathbf{R}}_{\mathbf{xs},f}\right\}\left[\Re\left\{\sum_f \widehat{\mathbf{R}}_{\mathbf{ss},f}\right\}\right]^{-1}. \tag{29}$$

6) Simulated Annealing: If one computes $\mathbf{A}_f$ through (24), (16), (17), (19), and (21), assuming $\mathbf{\Sigma}_{\mathbf{b},f} = 0$, one has $\mathbf{A}_f = \mathbf{A}_f$ as result. Thus, by continuity, when the covariance matrix $\mathbf{\Sigma}_{\mathbf{b},f}$ tends to zero, the resulting update rule for $\mathbf{A}_f$ tends to $\mathbf{A}_f \leftarrow \mathbf{A}_f$. Hence, the convergence of $\mathbf{A}_f$ becomes very slow for small values of $\sigma_{i,f}^2$. To overcome this difficulty and also favor global convergence, we have tested in the experimental section several simulated annealing strategies. In our framework, simulated annealing consists in setting the noise variances $\sigma_{i,f}^2$ to a common iteration-dependent value $\sigma_{i,f}^2(iter)$, initialized with an arbitrary large value $\hat{\sigma}_{i,f}^2$ and gradually decreased through iterations to a small value $\tilde{\sigma}_{i,f}^2$. Besides improving convergence speed, this scheme should also favor convergence to global solutions, as typical of annealing algorithms: the cost function is rendered flatter in the first iterations due to the (assumed) presence of high noise, smoothing out local minima, and is gradually brought back to its exact shape in the subsequent iterations.

7) Reconstruction of the Sources: Minimum mean square error (MMSE) estimates $\hat{\mathbf{s}}_{fn} = \mathbb{E}[\mathbf{s}_{fn}|\mathbf{x}_{fn};\boldsymbol{\theta}]$ of the source STFTs are directly retrieved using Wiener filter of (19). Time-domain sources may then be obtained through inverse STFT using an adequate overlap-add procedure with dual synthesis window (see e.g., [31]).

By conservativity of Wiener reconstruction the spatial images of the estimated sources and of the estimated noise sum up to the original mix in STFT domain, i.e., $\hat{\mathbf{A}}_f$, $\hat{\mathbf{s}}_{fn}$, and $\hat{\mathbf{b}}_{fn} = \mathbf{\Sigma}_{\mathbf{b},f}\mathbf{\Sigma}_{\mathbf{x},fn}^{-1}\mathbf{x}_{fn}$ satisfy (3). Thanks to linearity of the inverse-STFT, the reconstruction is conservative in the time domain as well.

B. Maximization of Individual Likelihoods With MU Rules

1) Criterion: We now consider a different approach consisting of maximizing the sum of individual channel log-likelihoods $\sum_i \log p(\mathbf{X}_i|\boldsymbol{\theta})$, hence discarding mutual information between the channels. This is equivalent to setting the off-diagonal terms of $\mathbf{x}_{fn}\mathbf{x}_{fn}^H$ and $\mathbf{\Sigma}_{\mathbf{x},fn}$ to zero in criterion (12), leading to minimization of cost

$$C_2(\boldsymbol{\theta}) = \sum_{ifn} d_{IS}\left(|x_{i,fn}|^2|\hat{v}_{i,fn}\right) \tag{30}$$

where $\hat{v}_{i,fn}$ is the structure defined by

$$\hat{v}_{i,fn} = \sum_j q_{ij,f}\underbrace{\sum_{k\in\mathcal{K}_j} w_{fk}h_{kn}}_{p_{j,fn}} \quad \left(+\sigma_{i,f}^2\right) \tag{31}$$

and $q_{ij,f} = |a_{ij,f}|^2$. For a fixed channel i, $\hat{v}_{i,fn}$ is basically the sum of the source variances modulated by the mixing weights. A noise variance term $\sigma_{i,f}^2$ might be considered, either fixed or to be estimated, but we will simply set it to zero as we will not here encounter the issues described in Section III-A6 about convergence of EM in noise-free observations.

Criterion (30) may also be read as the ML criterion corresponding to the model where the contributions of each component (and thus, of each source) to each channel would be different and independent realizations of the same Gaussian process, as opposed to the same realization. In other words, this

assumption amounts to changing our observation and source models given by (2) and (5) to

$$x_{i,fn} = \sum_{j=1}^{I} a_{ij,f} s^{(i)}_{j,fn} + b_{i,fn} \quad (32)$$

$$s^{(i)}_{j,fn} = \sum_{k \in \mathcal{K}_j} c^{(i)}_{k,fn} \quad \text{with} \quad c^{(i)}_{k,fn} \sim \mathcal{N}_c(0, w_{fk} h_{kn}) \quad (33)$$

and thus changing (7) to

$$s^{(i)}_{j,fn} \sim \mathcal{N}_c \left(0, \sum_{k \in \mathcal{K}_j} w_{fk} h_{kn} \right) \quad (34)$$

where $c^{(i)}_{k,fn}$ (resp. $s^{(i)}_{j,fn}$) denotes the contribution of component k (resp. source j) to channel i, and these contributions are assumed independent over channels (i.e., over i).

Our approach differs from the NTF approach of [6], [7] where the following PARAFAC structure [32] is considered

$$\hat{v}^{NTF}_{i,fn} = \sum_{k} q^{NTF}_{ik} w_{fk} h_{kn}. \quad (35)$$

It is only a sum of $I \times F \times N$ rank-1 tensors and amounts to assuming that $\hat{\mathbf{V}}^{NTF}_i = [\hat{v}^{NTF}_{i,fn}]_{fn}$ is a linear combination of $F \times N$ time–frequency patterns $\mathbf{w}_k h_k$, where $\mathbf{w}_k$ is column k of $\mathbf{W}$ and h_k is row k of $\mathbf{H}$. It intrinsically implies a linear instantaneous mixture and requires a postprocessing binding step in order to group the K elementary patterns into J sources, based on clustering of the ratios $\{q^{NTF}_{1k}/q^{NTF}_{2k}\}_k$ (in the stereo case). To ease comparison, our model can be rewritten as

$$\hat{v}_{i,fn} = \sum_{k} \overset{\circ}{q}_{ik,f} w_{fk} h_{kn} \quad (36)$$

subject to the constraint $\overset{\circ}{q}_{ik,f} = q_{ij,f}$ if and only if $k \in \mathcal{K}_j$ (with the notation introduced in Section II-C, we have also $\overset{\circ}{q}_{ik,f} = |\overset{\circ}{a}_{ik,f}|^2$). Hence, our model has the following merits with respect to (w.r.t.) the PARAFAC-NTF model: 1) it accounts for convolutive mixing by considering frequency-dependent mixing proportions ($\overset{\circ}{q}_{ik,f}$ instead of q^{NTF}_{ik}) and 2) the constraint that the K mixing proportions $\{\overset{\circ}{q}_{ik,f}\}_k$ can only take J possible values implies that the clustering of the components is taken care of within the decomposition as opposed to after the decomposition.

We have here chosen to use the IS divergence as a measure of fit in (30) because it connects with the optimal inference setting of Section III-A and because it was shown a relevant cost for factorization of audio power spectrograms [10], but other costs could be considered, such as the standard Euclidean distance and the generalized Kullback–Leibler (KL) divergence, which are the costs considered in [6] and [7].

2) Indeterminacies: Criterion (30) suffers from same scale, phase and permutations ambiguities as criterion (12), with the exception that ambiguity on the phase of $a_{ij,f}$ is now total as this parameter only appears through its squared-modulus. In the following, the scales are fixed as in Section III-A2.

3) Algorithm: We describe for the minimization of $C_2(\boldsymbol{\theta})$ an iterative MU algorithm inspired from NMF methodology [1], [33], [34]. Continual descent of the criterion under this algorithm was observed in practice. The algorithm simply consists of updating each scalar parameter θ_l by multiplying its value at previous iteration by the ratio of the negative and positive parts of the derivative of the criterion w.r.t. this parameter, namely

$$\theta_l \leftarrow \theta_l \frac{[\nabla_{\theta_l} C_2(\boldsymbol{\theta})]_-}{[\nabla_{\theta_l} C_2(\boldsymbol{\theta})]_+} \quad (37)$$

where $\nabla_{\theta_l} C_2(\boldsymbol{\theta}) = [\nabla_{\theta_l} C_2(\boldsymbol{\theta})]_+ - [\nabla_{\theta_l} C_2(\boldsymbol{\theta})]_-$ and the summands are both nonnegative [10]. Not any cost function gradient may be separated in two such summands, but this is the case for the Euclidean, KL and IS costs, and more generally the β-divergence of which they are specific cases [10], [26]. This scheme automatically ensures the non-negativity of the parameter updates, provided initialization with a nonnegative value.

The resulting parameter updates are described in Algorithm 2, where "." indicates element-wise matrix operations, $\mathbf{1}_{N \times 1}$ is a N-vector of ones, $\mathbf{q}_{ij}$ is the $F \times 1$ vector $[q_{ij,f}]_f$ and $\mathbf{V}_i$ (resp. $\hat{\mathbf{V}}_i$) is the $F \times N$ matrix $[|x_{i,fn}|^2]_{fn}$ (resp. $[\hat{v}_{i,fn}]_{fn}$). Some details about the derivation of the algorithm are given in Appendix B.

Algorithm 2 MU rules (one iteration)

- Update $\mathbf{Q}$

$$\mathbf{q}_{ij} \leftarrow \mathbf{q}_{ij} . \frac{\left[\hat{\mathbf{V}}_i^{-2} . \mathbf{V}_i . (\mathbf{W}_j \mathbf{H}_j)\right] \mathbf{1}_{N \times 1}}{\left[\hat{\mathbf{V}}_i^{-1} . (\mathbf{W}_j \mathbf{H}_j)\right] \mathbf{1}_{N \times 1}}. \quad (38)$$

- Update $\mathbf{W}$

$$\mathbf{W}_j \leftarrow \mathbf{W}_j . \frac{\sum_{i=1}^{I} \mathrm{diag}(\mathbf{q}_{ij}) \left(\hat{\mathbf{V}}_i^{-2} . \mathbf{V}_i\right) \mathbf{H}_j^T}{\sum_{i=1}^{I} \mathrm{diag}(\mathbf{q}_{ij}) \hat{\mathbf{V}}_i^{-1} \mathbf{H}_j^T}. \quad (39)$$

- Update $\mathbf{H}$

$$\mathbf{H}_j \leftarrow \mathbf{H}_j . \frac{\sum_{i=1}^{I} \left(\mathrm{diag}(\mathbf{q}_{ij}) \mathbf{W}_j\right)^T \left(\hat{\mathbf{V}}_i^{-2} . \mathbf{V}_i\right)}{\sum_{i=1}^{I} \left(\mathrm{diag}(\mathbf{q}_{ij}) \mathbf{W}_j\right)^T \hat{\mathbf{V}}_i^{-1}}. \quad (40)$$

- Normalize $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$ according to Section III-B2.

4) Linear Instantaneous Case: In the linear instantaneous case, when $q_{ij,f} = q_{ij}$, we obtain the following update rule for the mixing matrix coefficients:

$$q_{ij} \leftarrow q_{ij} . \frac{\mathrm{sum}\left[\hat{\mathbf{V}}_i^{-2} . \mathbf{V}_i . (\mathbf{W}_j \mathbf{H}_j)\right]}{\mathrm{sum}\left[\hat{\mathbf{V}}_i^{-1} . (\mathbf{W}_j \mathbf{H}_j)\right]} \quad (41)$$

where $\mathrm{sum}[\mathbf{M}]$ is the sum of all coefficients in $\mathbf{M}$. Then, $\mathrm{diag}(\mathbf{q}_{ij})$ needs only be replaced by q_{ij} in (39) and (40). The overall algorithm yields a specific case of PARAFAC-NTF which directly assigns the elementary components to J directions of arrival (DOA). This scheme however requires to fix in advance the partition $\{\mathcal{K}_j\}_{j=1}^{J}$ of $\mathcal{K} = \{1, \ldots, K\}$, i.e., assign a given number of components per DOA. In the specific linear instantaneous case, multiplicative updates for the whole matrices $\mathbf{Q}$, $\mathbf{W}$, $\mathbf{H}$ can be exhibited (instead of individual updates for q_{ij}, $\mathbf{W}_j$, $\mathbf{H}_j$), but are not given here for conciseness. They are similar in form to [33], [34] and lead to a faster MATLAB implementation.

5) Reconstruction of the Source Images: Criterion (30) being equivalent to the ML criterion under the model defined by (32) and (33), the MMSE estimate $\hat{s}^{(i)\mathrm{im}}_{j,fn} = \mathbb{E}[s^{(i)\mathrm{im}}_{j,fn} | \mathbf{x}_{fn}; \boldsymbol{\theta}]$ of the

image $s_{j,fn}^{(i)\mathrm{im}} \stackrel{\mathrm{def}}{=} a_{ij,f} s_{j,fn}^{(i)}$ of source j in channel i is computed through

$$\hat{s}_{j,fn}^{(i)\mathrm{im}} = \frac{q_{ij,f} p_{i,fn}}{\hat{v}_{i,fn}} x_{i,fn} \tag{42}$$

i.e., by Wiener filtering of each channel. A noise component (if any) can similarly be reconstructed as $\hat{b}_{i,fn} = (\sigma_{i,f}^2 / \hat{v}_{i,fn}) x_{i,fn}$. Overall the decomposition is conservative, i.e., $\sum_j \hat{s}_{j,fn}^{(i)\mathrm{im}} + \hat{b}_{i,fn} = x_{i,fn}$.

IV. Experiments

In this section, we first describe the test data and evaluation criteria, and then proceed with experiments. All the audio datasets and separation results are available from our demo web page [35]. MATLAB implementations of the proposed algorithms are also available from the authors' web pages.

A. Datasets

Four audio datasets have been considered and are described below.

- **Dataset A** consists of two synthetic stereo mixtures, one instantaneous the other convolutive, of $J = 3$ musical sources (drums, lead vocals and piano) created using 17-s excerpts of original separated tracks from the song "Sunrise" by S. Hurley, available under a Creative Commons License at [36] and downsampled to 16 kHz. The mixing parameters (instantaneous mixing matrix and the convolutive filters) were taken from the 2008 Signal Separation Evaluation Campaign (SiSEC'08) "under-determined speech and music mixtures" task development datasets [37], and are described below.
- **Dataset B** consists of synthetic (instantaneous and convolutive) and live-recorded (convolutive) stereo mixtures of speech and music sources, corresponding to the test data for the 2007 Stereo Audio Source Separation Evaluation Campaign (SASSEC'07) [38]. It also coincides with development dataset `dev2` of SiSEC'08 "under-determined speech and music mixtures" task. All the mixtures are 10 s long and sampled at 16 kHz. The instantaneous mixing is characterized by static positive gains. The synthetic convolutive filters were generated with the Roomsim toolbox [39]. They simulate a pair of omnidirectional microphones placed 1 m apart in a room of dimensions $4.45 \times 3.55 \times 2.5$ m with reverberation time 130 ms, which correspond to the setting employed for the live-recorded mixtures. The distances between the sources and the center of the microphone pair vary between 80 cm and 1.20 m. For all mixtures the source directions of arrival vary between -60° and $+60^\circ$ with a minimal spacing of 15° (for more details see [37]).
- **Dataset C** consists of SiSEC'08 test and development datasets for task "professionally produced music recordings". The test dataset consists of two excerpts (of about 22 s long) from two different professionally produced stereo songs, namely "Que pena tanto faz" by Tamy and "Roads" by Bearlin. The development dataset consists of two other excerpts (of about 12 s long) from the same songs, with all original stereo tracks provided separately. All recordings are sampled at 44 kHz (CD quality).
- **Dataset D** consists of three excerpts of length between 25 and 50 s taken from three professionally produced stereo recordings of well-known pop and reggae songs, and downsampled to 22 kHz.

TABLE I
STFT Window Lengths Used in Different Experiments

experiment section	dataset	window length		sampling freq. (Hz)
		samples	milliseconds	
IV-D, IV-E	A	1024	64	16000
IV-F	B - inst.	1024	64	16000
	B - conv.	2048	128	16000
IV-G	C	2048	46	44100
IV-H	D	2048	93	22050

B. Source Separation Evaluation Criteria

In order to evaluate our multichannel NMF algorithms in terms of audio source separation we use the signal-to-distortion ratio (SDR) numerical criterion defined in [38], which essentially compares the reconstructed source images with the original ones. The quality of the mixing system estimates was assessed with the mixing error ratio (MER) described at [37], which is an SNR-like criterion expressed in decibels. MATLAB routines for computing these criteria were obtained from the SiSEC'08 web page [37]. These evaluation criteria can only be computed when the original source spatial images (and mixing systems) are available. When not (i.e., for datasets C and D), separation performance is assessed perceptually and informally by listening to the separated source images, available online at [35].

C. Algorithm Parameters

1) STFT Parameters: In all the experiments below we used STFTs with half-overlapping sine windows, using the STFT computation tools for MATLAB available from [37]. The choice of the STFT window size is rather important, and is a matter of compromise between 1) good frequency resolution and validity of the convolutive mixing approximation of (2) and 2) validity of the assumption of source local stationarity. We have tried various window sizes (powers of 2) for every experiment, and the most satisfactory window sizes are reported in Table I.

2) Model Order: In our case the model order parameters consist of the total number of components K and the allocation of the components among the J sources, i.e., the partition $\{\mathcal{K}_1, \ldots, \mathcal{K}_J\}$. The value of J may be set by hand to the number of instrumental sources in the recording, although, as we shall discussed later, the existence of non-point sources or the existence of sources mixed similarly might render the choice of J trickier. The choice of the number of components per source may raise more questions. As a first guess one may choose a high value, so that the model can account for all of the diversity of the source; basically, one may think of one component per note or elementary sound object. This leads to increased flexibility in the model, but, at the same time, can lead to data overfitting (in case of few data), and favors the existence of local minima,

thus rendering optimization more difficult, as well as more intensive. Interestingly, it has been noted in [10] that, given a limited number of components, IS-NMF is also able to learn higher level structures in the musical signal. One or a few components can capture a large part of one source or a subset of sources, so that a coherent sound decomposition can be achieved to some extent. A similar behavior was logically observed in our multichannel scenario, with even more success as the spatial information helps to discriminate between the sources. Hence, satisfying source separation results could be obtained with small values of K.

In the experiments of Sections IV-D and IV-E we set $\#\mathcal{K}_j = 4$; however, this has minor importance there as the aim of these experiments is merely to investigate the algorithms behavior, and not to obtain optimal source separation performance. In the experiments of Sections IV-F and IV-G, $\#\mathcal{K}_j$ is chosen by hand through trials so as to obtain most satisfying results. In the experiment of Section IV-H the total number of components is arbitrary set to either $K = 15$ or 20, depending on the recording, and the numbers of components per source $\#\mathcal{K}_j$ are chosen automatically by the initialization procedure, see below.

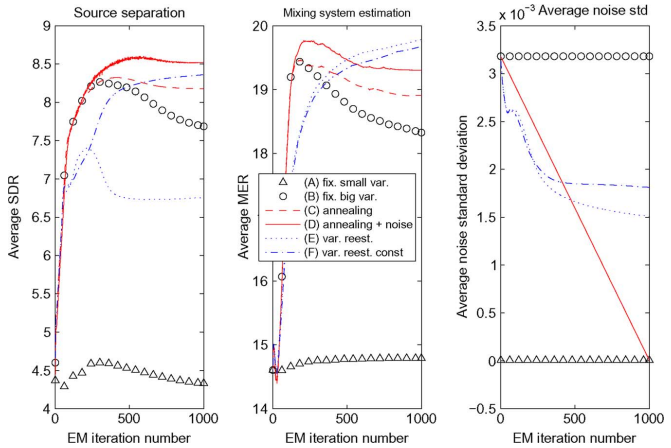

Fig. 2. EM algorithm results on convolutive mixture of dataset A, using various noise variance update schemes. (Left) Average source separation SDR. (Middle) average mixing system identification MER. (Right) average noise standard deviation. (A) Triangles: small fixed noise variance. (B) Circles: larger fixed noise variance. (C) Dashed line: annealing. (D)Solid line: annealing with noise injection. (E) Dotted line: diagonal noise covariance reestimation. (F) Dash-dotted line: isotropic noise variance reestimation.

D. Dealing With the Noise Part in the EM Algorithm

In this section, we experiment strategies for updating the noise parameters in the EM algorithm. We here arbitrarily use the convolutive mixture of dataset A and set the total number of components to $K = 12$, equally distributed between $J = 3$ sources. Our EM algorithm being sensitive to parameters initialization, we used the following *perturbed oracle* initializations so as to ensure "good" initialization: factors $\mathbf{W}$ and $\mathbf{H}$ as computed from the original sources using IS-NMF [10] and original mixing system $\mathbf{A}$, all perturbed with high level additive noise. We have tested the following noise update schemes.

- (**A**): $\mathbf{\Sigma}_{\mathbf{b},f} = \tilde{\sigma}^2\mathbf{I}_I$, with fixed $\tilde{\sigma}^2$ set to 16-bit PCM quantization noise variance.
- (**B**): $\mathbf{\Sigma}_{\mathbf{b},f} = \hat{\sigma}_f^2\mathbf{I}_I$, with fixed $\hat{\sigma}_f^2$ set to the average channel empirical variance in every frequency band divided by 100, i.e., $100\hat{\sigma}_f^2 = \sum_{in} |x_{i,fn}|^2/IN$.
- (**C**): $\mathbf{\Sigma}_{\mathbf{b},f} = \sigma_f^2\mathbf{I}_I$ with standard deviation σ_f decreasing linearly through iterations from $\hat{\sigma}_f$ to $\tilde{\sigma}$. This is what we refer to as simulated annealing.
- (**D**): Same strategy as (C), but with adding a random noise with covariance $\mathbf{\Sigma}_{\mathbf{b},f}$ to $\mathbf{X}$ at every EM iteration. We refer to this as annealing with noise injection.
- (**E**): $\mathbf{\Sigma}_{\mathbf{b},f} = \mathrm{diag}([\sigma_{i,f}^2]_i)$ is reestimated with update (25).
- (**F**): Noise covariance is reestimated like in scheme E, but under the more constrained structure $\mathbf{\Sigma}_{\mathbf{b},f} = \sigma_f^2\mathbf{I}_I$ (isotropic noise in each subband). In that case, operator $\mathrm{diag}(\cdot)$ in (25) needs to be replaced with $\mathrm{trace}(\cdot)\mathbf{I}_I/I$.

The algorithm was run for 1000 iterations in each case and the results are presented in Fig. 2, which displays the average SDR and MER along iterations, as well as the noise standard deviations $\sigma_{i,f}$, averaged over all channels i and frequencies f. As explained in Section III-A6, we observe that with a small fixed noise variance (scheme A), the mixing parameters stagnate. With a fixed larger noise variance (scheme B) convergence starts well but then performance drops due to artificially high noise variance. Simulated annealing (scheme C) overcomes this problem, and artificial noise injection (scheme D) even improves the results (both in terms of source separation and mixing system estimation). Noise variance reestimation allows to obtain performances almost similar to annealing, but only in the case when the variance is constrained to be the same in both channels (scheme F). However, we observed that faster convergence is obtained in general using annealing with noise injection (scheme D) for similar results.

Finally, it should be noted that for the schemes with annealing (C and D) both the average SDR and MER start decreasing from about 400 iterations (for SDR) and 200 iterations (for MER). We believe this is because the final noise variance $\tilde{\sigma}^2$ (set to 16-bit PCM quantization noise variance) might be too small to account for discrepancy in the convolutive mixing equation STFT-approximation (2). Indeed, with scheme F (constrained reestimated variance) the average noise standard deviation seem to be converging to a value in the range of 0.002 (see right plot of Fig. 2), which is much larger than $\tilde{\sigma}$. Thus, if computation time is not an issue, scheme F can be considered the most advantageous because this is the only scheme to systematically increase both the average SDR and MER at every iteration and it allows to adjust a suitable noise level adaptively. However, as we want to keep the number of iterations low (e.g., 300–500) for sake of short computation time, we will resort to scheme D in the following experiments.

E. Convergence and Separation Performance

In this experiment we wish to check consistency of optimization of the proposed criteria with respect to source separation performance improvement, in the least as measured by the SDR. We used both mixtures of dataset A (instantaneous and convolutive) and ran 1000 iterations of both algorithms (EM and MU) from ten different perturbed oracle initializations, obtained as in previous section. Again we used $K = 12$ components, equally split into $J = 3$ sources. Figs. 3 and 4 report results for the instantaneous and convolutive mixtures, respectively. Plots on

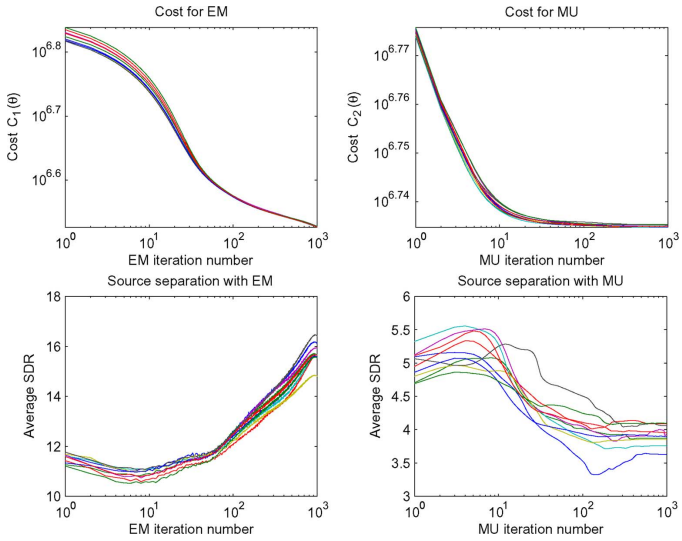

Fig. 3. Ten runs of EM and MU from ten perturbed oracle initializations using instantaneous mixture of dataset A. (Top) cost functions. (Bottom) average SDRs.

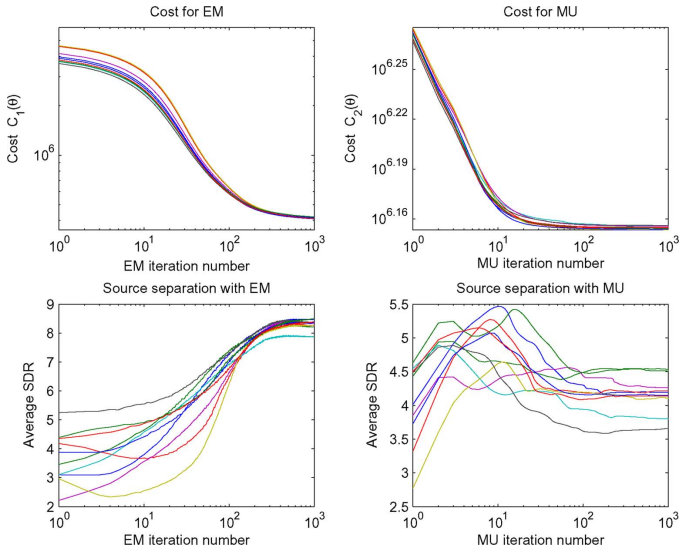

Fig. 4. Ten runs of EM and MU from ten perturbed oracle initializations using convolutive mixture of dataset A. (Top) cost functions. (Bottom) average SDRs.

top row display in log-scale the cost functions $C_1(\boldsymbol{\theta})$ and $C_2(\boldsymbol{\theta})$ w.r.t. iterations for all ten runs. Note that cost $C_1(\boldsymbol{\theta})$ is not positive in general, see (12), so that we have added a common large constant value to all curves so as to ensure positivity, and to be able plotting cost value in the logarithmic scale. Plots on bottom row display the average SDRs.

The results show that maximization of the joint likelihood with the EM algorithm leads to consistent improvement of source separation performance in term of SDR, in the sense that final average SDR values are higher than values at initialization. This is not the case with MU, which results in nearly every case in worsening the SDR values obtained from oracle initialization. This is undoubtedly a consequence of discarding mutual information between the channels.

As for computational loads, our MATLAB implementation of EM (resp. MU) algorithm takes about 80 min (resp. 20 min) per 1000 iterations, for this particular experiment with 17-s stereo mixture (sampled at 16 kHz), $J = 3$ sources, and $K = 12$ components.

F. Blind Separation of Under-Determined Speech and Music Mixtures

In this section, we compare our algorithms with the methods that achieved competitive results at the SASSEC'07 evaluation campaign for the tasks of underdetermined mixtures of respectively speech and music signals, in both instantaneous and convolutive cases. We used the same data and evaluation criteria as in the campaign. More precisely, our algorithms are compared in the instantaneous case to the method of Vincent [40], based on source STFT reconstruction using a minimum l_0 norm constraint given a mixing matrix estimate obtained with the method of Arberet *et al.* [41]. In the convolutive case, our algorithms are compared to the method of Sawada, based on frequency-dependent complex-valued mixing matrices estimation [42], and *a posteriori* grouping relying on temporal correlations between sources in different frequency bins [20]. We used the outputs of these methods to initialize our own algorithms. In the linear instantaneous case, we were given MATLAB implementations of [40] and [41]. In the convolutive case, we simply downloaded the source image estimates from the SASSEC'07 web page [43]. In both cases we built initializations of $\mathbf{W}$ and $\mathbf{H}$ based on NMF of the source spectrogram estimates.[3]

We have found satisfactory separation results through trials using $\#\mathcal{K}_j = 4$ components for musical sources and $\#\mathcal{K}_j = 10$ components for speech sources. More components seem to be needed for speech so as to account for its higher variability (e.g., vibrato). The EM and MU algorithms were run for 500 iterations, final source separation SDR results together with reference methods results are displayed in Table II.[4] The EM method yields a significant separation improvement for all linear instantaneous mixtures. Improvement is also obtained in the convolutive case for most source estimates, but is less significant in terms of SDRs. However, and maybe most importantly, we believe our source estimates to be generally more pleasant to listen to. Indeed, one drawback of sparsity-based, nonlinear source reconstruction is musical noise, originating from unnatural, isolated time-frequency atoms scattered over the time–frequency plane. In contrast, our Wiener source estimates, obtained as a linear combination of data in each TF cell, appear to be less prone to such artifacts as can be listened to at demo web page [35]. We have entered our EM algorithm to the "under-determined speech and music mixtures" task of SiSEC'08 for instantaneous mixtures, and our results can be compared to other

[3]However, in that case we used KL-NMF instead of IS-NMF, not to fit the lower-energy residual artifacts and interferences, to which IS-NMF might be overly sensitive as a consequence of its scale-invariance. This seemed to lead to better initializations indeed.

[4]The reference algorithms performances in Table II do not always coincide with those given on the SASSEC'07 web page [43]. In the instantaneous case, this is because we have not used the exact same implementation of the l_0 minimization algorithm [40] that was used for SASSEC. In the convolutive case, this is because we have removed the dc component from all speech signals (including reference, source image estimates, and mixtures) using high-pass filtering, in order to avoid numerical instabilities.

TABLE II
SOURCE SEPARATION RESULTS FOR SASSEC DATA IN TERMS OF SDR (dB)

Linear instantaneous mixtures

	female4				male4				nodrums			wdrums			average
	s1	s2	s3	s4	s1	s2	s3	s4	s1	s2	s3	s1	s2	s3	
l_0 min.	12.6	6.1	4.7	7.3	15.6	2.7	5.3	6.9	21.2	1.7	15.8	-0.5	3.1	28.4	9.6
EM	14.2	7.8	5.9	8.6	16.8	3.5	8.2	9.6	27.1	7.6	21.4	0.9	4.6	29.8	12.3
MU	3.9	0.9	0.1	2.2	8.6	-0.7	2.8	2.9	8.8	-6.4	3.3	10.0	2.9	19.3	4.4
Synthetic convolutive mixtures (1m)															
Sawada	5.2	5.3	3.2	2.6	4.5	0.6	4.9	2.3	3.0	1.0	-1.6	4.4	-12.7	0.6	1.3
EM	7.7	6.4	4.1	3.2	6.2	0.4	5.5	2.7	4.1	1.0	-1.8	3.9	-12.4	1.3	1.9
MU	5.2	3.3	2.7	1.4	3.4	-0.9	3.0	1.7	2.8	1.0	-2.0	5.9	-10.9	1.9	1.1
Live-recorded convolutive mixtures (1m)															
Sawada	4.1	3.8	6.0	3.3	3.0	1.6	4.8	2.4	4.1	5.1	-3.8	4.1	4.5	6.0	3.5
EM	5.3	3.6	7.2	4.3	3.5	2.1	5.6	3.1	4.5	7.3	-4.5	4.9	5.5	8.0	4.3
MU	1.6	-0.2	4.3	1.8	1.1	0.0	2.8	2.1	3.9	3.6	-4.9	4.1	4.5	7.5	2.4

methods in [44], and online at [45]. Note that among the ten algorithms participating in this task our algorithm outperformed all the other competiting methods by at least 1 dB for all separation measures (SDR, ISR, SIR, and SAR), see [44, Table 2].

G. Supervised Separation of Professionally Produced Music Recordings

We here apply our algorithms to the separation of the professionally produced music recordings of dataset B. This is a supervised setting in the sense that training data is available to learn the source spectral patterns $\mathbf{W}$ and filters. The following procedure is used.

- Learn mixing parameters $\{a_{ij,f}^{tr}\}_{if}$, spectral patterns $\mathbf{W}_j^{tr}$, and activation coefficients $\mathbf{H}_j^{tr}$ from available training signal images of source j (using 200 iterations of EM/MU); discard $\mathbf{H}_j^{tr}$.
- Clamp $\mathbf{A}$ and $\mathbf{W}$ to their trained values $\mathbf{A}^{tr}$ and $\mathbf{W}^{tr}$ and reestimate activation coefficients $\mathbf{H}$ from test data $\mathbf{X}$ (using 200 iterations of EM/MU).
- Reconstruct source image estimates from $\mathbf{A}^{tr}$, $\mathbf{W}^{tr}$ and $\mathbf{H}$.

Except for the training of mixing coefficient, the procedure is similar in spirit to supervised single-channel separation schemes proposed, e.g., in [9] and [46].

One important issue with professionally produced modern music mixtures is that they do not always comply with the mixing assumptions of (3). This might be due to nonlinear sound effects (e.g., dynamic range compression), to reverberation times longer than the analysis window length, and maybe most importantly to when the *point source* assumption does not hold anymore, i.e., when the channels of a stereo instrumental track cannot be represented as a convolution of the *same* source signal. The latter situation might happen when a sufficiently voluminous musical instrument (e.g., piano, drums, acoustic guitar) is recorded with several microphones placed close to the instrument. As such, the guitar track of the "Que pena tanto faz" song from dataset C is a non-point source image. Such tracks may be modeled as a sum of several point sources, with different mixing filters.

For the "Que pena tanto faz" song, the vocal part is modeled as an instantaneously mixed point source image with $\#\mathcal{K}_1 = 8$ components while the guitar part is modeled as a sum of three convolutively mixed point source images, each modeled with $\#\mathcal{K}_2 = \#\mathcal{K}_3 = \#\mathcal{K}_4 = 3$ components. For the "Roads" song, the bass and vocals parts are each modeled as instantaneously mixed point source images with six components, the piano part is modeled as a convolutive point source image with six components and finally, the residual background music (sum of remaining tracks) is modeled as a sum of three convolutive point source images with four components. The audio results, available at [35], tend to show better performance of the EM approach, especially on the "Roads" song. Our results can be compared to those of the other methods that entered the "professionally produced music recordings" task of SiSEC'08 in [44], and online at [47].

H. Blind Separation of Professionally Produced Music Recordings

In the last experiment, we have tested the EM and MU algorithms for the separation of professionally produced music recordings (commercial CD excerpts) in a fully unsupervised (blind) setting. We used the following parameter initialization procedure, inspired from [48], which yielded satisfactory results.

- Stack left and right mixture STFTs so as to create a $2F \times N$ complex-valued matrix $\mathbf{X}_{2\text{ch}} = [\mathbf{X}_L^T \; \mathbf{X}_R^T]^T$.
- Produce a K-components IS-NMF decomposition of $|\mathbf{X}_{2\text{ch}}|^2 \approx \mathbf{W}_{2\text{ch}}\mathbf{H}_{2\text{ch}}$.
- Initialize $\mathbf{W}$ as the average of $\mathbf{W}_L$ and $\mathbf{W}_R$, where $\mathbf{W}_{2\text{ch}} = [\mathbf{W}_L^T \; \mathbf{W}_R^T]^T$. Initialize $\mathbf{H} = \mathbf{H}_{2\text{ch}}$.
- Reconstruct K components $\widehat{\mathbf{C}}_{2\text{ch},k} = [\widehat{\mathbf{C}}_{L,k}^T \; \widehat{\mathbf{C}}_{R,k}^T]^T$ from $\mathbf{X}_{2\text{ch}}$, $\mathbf{W}_{2\text{ch}}$, and $\mathbf{H}_{2\text{ch}}$, using single-channel Wiener filtering (see, e.g., [10]). Produce K ad-hoc left and right component-dependent mixing filters estimates by averaging $\widehat{\mathbf{C}}_{L,k}/\Phi$ and $\widehat{\mathbf{C}}_{R,k}/\Phi$ over frames, with $\Phi = \arg(\widehat{\mathbf{C}}_{L,k})$, and normalizing according to Section III-A2. Cluster the resulting filter estimates with the K-means algorithm, whose output can be used to

define the partition $\{\mathcal{K}_j\}_{j=1}^{J}$ (using cluster indices) and a mixing system estimate $\hat{\mathbf{A}}$ (using cluster centroids).

Depending on the recording we set the number of sources J to 3 or 4 and used a total of $K = 15$ to 20 components. The EM and MU algorithms were run for 300 iterations in every case. On these specific examples the superiority of the EM method w.r.t. the MU method is not as clear as with previous datasets. A likely reason is the existence of nonpoint sources breaking the validity of mixing assumptions (2). In such precise cases, choosing not to exploit inter-channel dependencies might be better, because our model of these dependencies is now wrong. Looking for suitable probabilistic models of nonpoint sources is a new and interesting research direction.

In some cases the source image estimates contain several musical instruments and some musical instruments are spread over several source images. Besides poor initialization, this can be explained by 1) sources mixed similarly (e.g, same directions of arrival), and thus impossible to separate in our fully blind setting, 2) nonpoint sources, not well represented by our model and thus split into different source image estimates.

One way to possibly refine separation results is to reconstruct individual stereo component images (i.e., obtained via Wiener filtering (20) in case of EM method, or via (42) by replacing $p_{i,fn}$ with $w_{fk}h_{kn}$ in case of MU method), and manually group them through listening, either to separate sources mixed similarly, or to reconstruct multidirectional sound sources that better match our understanding/perception of a single source.

Finally, to show the potential of our source separation approach for music remixing, we have created some remixes using the blindly separated source images and/or the manually regrouped ones. The remixes were created in Audacity [49] by simply re-panning the source image estimates between left and right channels and by changing their gains. The audio results can be listened to at [35].

V. Conclusion

We have presented a general probabilistic framework for the representation of multichannel audio, under possibly underdetermined and noisy convolutive mixing assumptions. We have introduced two inference methods: an EM algorithm for the maximization of the channels joint log-likelihood and a MU algorithm for the maximization of the sum of individual channel log-likelihoods. The complexity of these algorithms grows linearly with the number of model components, and make them thus suitable to real-world audio mixtures with any number of sources. The corresponding CPU computational loads are in the order of a few hours for a song, which may be considered reasonable for applications such as remixing, where real-time is not an issue.

We have applied our decomposition algorithms to stereo source separation in various settings, covering blind and supervised separation, music and speech sources, synthetic instantaneous and convolutive mixtures, as well as professionally produced music recordings.

The EM algorithm was shown to outperform state-of-the-art methods, given appropriate initializations. Both our methods have indeed been found sensitive to parameter initialization, but we have come up with two satisfying initialization schemes. The first one, described in Section IV-F, consists in using the output of a different separation algorithm. We show that our EM algorithm improves the separation results in almost all cases. The second scheme, described in Section IV-H, consists in a single-channel NMF decomposition followed by K-means filters clustering. Our experiments tend to show that the NMF model is more suitable to music than speech: music sources can be represented by a small number of components to attain good separation performance, and informal listening indicates better separation of music signals.

Given that the mixed signals follow the mixing and point source assumptions inherent to (2), the EM method gives better separation results than the MU method, because between-channel dependencies are optimally exploited. However, the performance of the EM method may significantly drop when these assumptions are not verified. In contrast, we have observed that the MU method, which relies on a weaker model of between-channel dependencies, yields more even results overall and higher robustness to model discrepancies (that may for example occur in professionally produced recordings).

Let us now mention some further research directions. Algorithms faster than EM (both in terms of convergence rate and CPU time per iteration) would be desirable for optimization of the joint likelihood (12). As such, we envisage turning to Newton gradient optimization, as inspired from [50]. Mixed strategies could also be considered, consisting of employing EM in the first few iterations to get a sharp decrease of the likelihood before switching to faster gradient search once in the neighborhood of a solution.

Bayesian extensions of our algorithm are readily available, using for example priors favoring sparse activation coefficients h_k, or even sparse filters $q_{ij,f}$ like in [51]. Minor changes are required in the MU rules so as to yield algorithms for maximum *a posteriori* (MAP) estimation. More complex priors structure can also be envisaged within the EM method, such as Markov chains favoring smoothness of the activation coefficients $\mathbf{H}$ [10].

An important perspective is automatic order selection. In our case, that concerns the total number of components K, the number of sources J and the partition $\{\mathcal{K}_j\}_j$. Regarding the total number of components K, ideas from *automatic relevance determination* can be explored, see [52] in a NMF setting. Then the problem of partitioning can be viewed as a clustering problem with unknown number of clusters J, which is a typical machine learning problem.

While we have assessed the validity of our model in terms of source separation, our decompositions more generally provide a data-driven object-based representation of multichannel audio that could be relevant to other problems such as audio transcription, indexing and object-based coding. As such, it will be interesting to investigate the semantics revealed by the learnt spectral patterns $\mathbf{W}$ and activation coefficients $\mathbf{H}$.

Finally, as discussed in Section IV-H, new models should be considered for professionally produced music recordings, dealing with nonpoint sources, nonlinear sound effects, such as dynamic range compression, and long reverberation times.

APPENDIX A

APPENDIX A
EM ALGORITHM DERIVATION OUTLINE

The complete data minus log-likelihood can be written as

$$\begin{aligned}
&-\log p(\mathbf{X},\mathbf{C}|\boldsymbol{\theta}) \\
&\quad = -\log p(\mathbf{X}|\mathbf{C},\boldsymbol{\theta}) - \log p(\mathbf{C}|\boldsymbol{\theta}) \\
&\quad \stackrel{c}{=} \sum_{fn}\left[\log|\boldsymbol{\Sigma}_{\mathbf{b},f}| + (\mathbf{x}_{fn}-\mathbf{A}_f\mathbf{s}_{fn})^H\boldsymbol{\Sigma}_{\mathbf{b},f}^{-1}(\mathbf{x}_{fn}-\mathbf{A}_f\mathbf{s}_{fn})\right] \\
&\qquad + \sum_k\sum_{fn}\left[\log(h_{k,n}w_{k,f}) + \frac{|c_{k,fn}|^2}{h_{k,n}w_{k,f}}\right] \\
&\quad = \sum_{fn}\left[\log|\boldsymbol{\Sigma}_{\mathbf{b},f}| + \sum_k\log(h_{k,n}w_{k,f}) + \sum_k\frac{|c_{k,fn}|^2}{h_{k,n}w_{k,f}}\right] \\
&\qquad + N\sum_f \operatorname{trace}\Big[\boldsymbol{\Sigma}_{\mathbf{b},f}^{-1}\mathbf{R}_{\mathbf{xx},f} - \boldsymbol{\Sigma}_{\mathbf{b},f}^{-1}\mathbf{A}_f\mathbf{R}_{\mathbf{xs},f}^H \\
&\qquad\qquad - \boldsymbol{\Sigma}_{\mathbf{b},f}^{-1}\mathbf{R}_{\mathbf{xs},f}\mathbf{A}_f^H + \boldsymbol{\Sigma}_{\mathbf{b},f}^{-1}\mathbf{A}_f\mathbf{R}_{\mathbf{ss},f}\mathbf{A}_f^H\Big]
\end{aligned} \tag{43}$$

with $\mathbf{R}_{\mathbf{xx},f}$, $\mathbf{R}_{\mathbf{xs},f}$, $\mathbf{R}_{\mathbf{ss},f}$, and $u_{k,fn}$ defined by (13) and (14). Thus, we have shown that the complete data log-likelihood can be represented in the following form:

$$\log p(\mathbf{X},\mathbf{C}|\boldsymbol{\theta}) = \langle\eta(\boldsymbol{\theta}),\mathbf{T}(\mathbf{X},\mathbf{C})\rangle + \nu(\boldsymbol{\theta}) \tag{44}$$

where $\mathbf{T}(\mathbf{X},\mathbf{C})$ is a vector of all scalar elements of $\mathbf{t}(\mathbf{X},\mathbf{C}) \triangleq \{\mathbf{R}_{\mathbf{xx},f},\mathbf{R}_{\mathbf{xs},f},\mathbf{R}_{\mathbf{ss},f},\{u_{k,fn}\}_{kn}\}_f$, and $\eta(\boldsymbol{\theta})$ and $\nu(\boldsymbol{\theta})$ are some vector and scalar functions of parameters. That means that the complete data pdfs $\{p(\mathbf{X},\mathbf{C}|\boldsymbol{\theta})\}_{\boldsymbol{\theta}}$ form an *exponential family* (see, e.g., [11], [29]) and complete data statistics $\mathbf{t}(\mathbf{X},\mathbf{C})$ is a *natural (sufficient) statistics* [11], [29] for this family. To derive an EM algorithm in this special case one needs to 1) solve complete data ML criterion (thanks to (44) this solution can be always expressed as a function of natural statistics $\mathbf{t}(\mathbf{X},\mathbf{C})$), and 2) replace in this solution $\mathbf{t}(\mathbf{X},\mathbf{C})$ by its conditional expectation $\hat{\mathbf{t}}(\mathbf{X},\boldsymbol{\theta}') \triangleq \int\mathbf{t}(\mathbf{X},\mathbf{C})p(\mathbf{C}|\mathbf{X},\boldsymbol{\theta}')d\mathbf{C}$ using model $\boldsymbol{\theta}'$ estimated at the previous step of EM.

To solve the complete data ML criterion, we first compute the derivatives of $\log p(\mathbf{X},\mathbf{C}|\boldsymbol{\theta})$ (43) w.r.t. model parameters $\boldsymbol{\theta}$ (see [53] for issues regarding derivation w.r.t. complex-valued parameters), set them to zero and solve the corresponding equations (subject to the constraint that $\boldsymbol{\Sigma}_{\mathbf{b},f}$ is diagonal), and we have:[5]

$$\mathbf{A}_f = \mathbf{R}_{\mathbf{xs},f}\mathbf{R}_{\mathbf{ss},f}^{-1} \tag{45}$$

$$\begin{aligned}
\boldsymbol{\Sigma}_{\mathbf{b},f} = \operatorname{diag}\big(&\mathbf{R}_{\mathbf{xx},f} - \mathbf{A}_f\mathbf{R}_{\mathbf{xs},f}^H \\
&- \mathbf{R}_{\mathbf{xs},f}\mathbf{A}_f^H + \mathbf{A}_f\mathbf{R}_{\mathbf{ss},f}\mathbf{A}_f^H\big)
\end{aligned} \tag{46}$$

$$w_{fk} = \frac{1}{N}\sum_n\frac{u_{k,fn}}{h_{kn}}, \quad h_{kn} = \frac{1}{F}\sum_f\frac{u_{k,fn}}{w_{fk}}. \tag{47}$$

[5]Bayesian MAP estimation can be carried out instead of ML by simply adding a prior term $-\log p(\boldsymbol{\theta})$ to the right part of (43) and solving the corresponding complete data MAP criterion.

Our EM algorithm is strictly speaking only a *Generalized* EM algorithm [54] because it only ensures $Q(\boldsymbol{\theta}^{m+1}|\boldsymbol{\theta}^m) \geq Q(\boldsymbol{\theta}^m|\boldsymbol{\theta}^m)$. Indeed, in (47) $\mathbf{W}$ is still a function of $\mathbf{H}$, and reversely, $\mathbf{H}$ is a function of $\mathbf{W}$.

To finish derivation of our EM algorithm we need to compute conditional expectation of the natural statistics $\mathbf{t}(\mathbf{X},\mathbf{C})$. It can be shown that given $\mathbf{x}_{fn}$ the source vector $\mathbf{s}_{fn}$ is a proper Gaussian random vector, i.e.,

$$p(\mathbf{s}_{fn}|\mathbf{x}_{fn};\boldsymbol{\theta}) = N_c\left(\mathbf{s}_{fn};\hat{\mathbf{s}}_{fn},\boldsymbol{\Sigma}_{\mathbf{s},fn}^{\text{post}}\right) \tag{48}$$

with mean vector $\hat{\mathbf{s}}_{fn}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{s},fn}^{\text{post}}$ as follows:

$$\begin{aligned}
\hat{\mathbf{s}}_{fn} &= \boldsymbol{\Sigma}_{\mathbf{s},f}\mathbf{A}_f^H\left(\mathbf{A}_f\boldsymbol{\Sigma}_{\mathbf{s},f}\mathbf{A}_f^H + \boldsymbol{\Sigma}_{\mathbf{b},f}\right)^{-1}\mathbf{x}_{fn}, \\
\boldsymbol{\Sigma}_{\mathbf{s},fn}^{\text{post}} &= \boldsymbol{\Sigma}_{\mathbf{s},f} - \boldsymbol{\Sigma}_{\mathbf{s},f}\mathbf{A}_f^H\left(\mathbf{A}_f\boldsymbol{\Sigma}_{\mathbf{s},f}\mathbf{A}_f^H + \boldsymbol{\Sigma}_{\mathbf{b},f}\right)^{-1}\mathbf{A}_f\boldsymbol{\Sigma}_{\mathbf{s},f}.
\end{aligned}$$

Computing conditional expectations of $\mathbf{R}_{\mathbf{xs},f}$ and $\mathbf{R}_{\mathbf{ss},f}$ using (48) leads to (16) and (17) of EM Algorithm 1. Very similar derivations can be done to compute the conditional expectations of $u_{k,fn}$. To that matter, one only needs to compute the posterior distribution of $\mathbf{c}_{fn}$ instead of $\mathbf{s}_{fn}$, using mixing equation (10) instead of mixing equation (3).

APPENDIX B
MU ALGORITHM DERIVATION OUTLINE

Let θ be a scalar parameter of the set $\{\mathbf{Q},\mathbf{W},\mathbf{H}\}$. The derivative of cost $C_2(\boldsymbol{\theta})$, given by (30), w.r.t. θ simply writes

$$\nabla_\theta D(\mathbf{V}|\hat{\mathbf{V}}) = \sum_{ifn}(\nabla_\theta\hat{v}_{i,fn})d'_{IS}(v_{i,fn}|\hat{v}_{i,fn}) \tag{49}$$

where $d'_{IS}(x|y)$ is the derivative of $d_{IS}(x|y)$ w.r.t. y given by

$$d'_{IS}(x|y) = \frac{1}{y} - \frac{x}{y^2}. \tag{50}$$

Using (49), we obtain the following derivatives:

$$\begin{aligned}
\nabla_{q_{ij,f}}D(\mathbf{V}|\hat{\mathbf{V}}) &= \sum_{n=1}^N p_{j,fn}d'(v_{i,fn}|\hat{v}_{i,fn}) \\
\nabla_{w_{jfk}}D(\mathbf{V}|\hat{\mathbf{V}}) &= \sum_{i=1}^I\sum_{n=1}^N q_{ij,f}h_{j,kn}d'(v_{i,fn}|\hat{v}_{i,fn}) \\
\nabla_{h_{jkn}}D(\mathbf{V}|\hat{\mathbf{V}}) &= \sum_{i=1}^I\sum_{f=1}^F q_{ij,f}w_{j,fk}d'(v_{i,fn}|\hat{v}_{i,fn})
\end{aligned}$$

which can be written in the following matrix forms:

$$\begin{aligned}
\nabla_{\mathbf{q}_{ij}}D(\mathbf{V}|\hat{\mathbf{V}}) &= \left(\hat{\mathbf{V}}_i^{\cdot-1}\mathbf{P}_j - \hat{\mathbf{V}}_i^{\cdot-2}.\mathbf{V}_i.\mathbf{P}_j\right)\mathbf{1}_{N\times 1} \\
\nabla_{\mathbf{W}_j}D(\mathbf{V}|\hat{\mathbf{V}}) &= \sum_{i=1}^I\operatorname{diag}(\mathbf{q}_{ij})\left(\hat{\mathbf{V}}_i^{\cdot-2}.(\hat{\mathbf{V}}_i - \mathbf{V}_i)\right)\mathbf{H}_j^T \\
\nabla_{\mathbf{H}_j}D(\mathbf{V}|\hat{\mathbf{V}}) &= \sum_{i=1}^I(\operatorname{diag}(\mathbf{q}_{ij})\mathbf{W}_j)^T\left(\hat{\mathbf{V}}_i^{\cdot-2}.(\hat{\mathbf{V}}_i - \mathbf{V}_i)\right).
\end{aligned}$$

Hence, the update rules given in Algorithm 2, following the multiplicative update strategy described in Section III-B3.

ACKNOWLEDGMENT

The authors would like to thank S. Arberet for kindly sharing his implementation of DEMIX algorithm [41], all the organizers of SiSEC'08 for well-prepared evaluation campaign, as well as the anonymous reviewers for their valuable comments.

REFERENCES

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects with nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[2] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2003, pp. 177–180.

[3] N. Bertin, R. Badeau, and G. Richard, "Blind signal decompositions for automatic transcription of polyphonic music: NMF and K-SVD on the benchmark," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP'07)*, Honolulu, HI, 2007, pp. 65–68.

[4] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

[5] P. Smaragdis, "Convolutive speech bases and their application to speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 1–12, Jan. 2007.

[6] R. M. Parry and I. A. Essa, "Estimating the spatial position of spectral components in audio," in *Proc. 6th Int. Conf. Ind. Compon. Anal. Blind Signal Separation (ICA'06)*, Charleston, SC, Mar. 2006, pp. 666–673.

[7] D. FitzGerald, M. Cranitch, and E. Coyle, "Non-negative tensor factorisation for sound source separation," in *Proc. Irish Signals Syst. Conf.*, Dublin, Ireland, Sep. 2005, pp. 8–12.

[8] L. Parra and C. Spence, "Convolutive blind source separation of non-stationary sources," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 320–327, May 2000.

[9] L. Benaroya, R. Gribonval, and F. Bimbot, "Non negative sparse representation for Wiener based source separation with a single sensor," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'03)*, Hong Kong, 2003, pp. 613–616.

[10] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Comput.*, vol. 21, no. 3, pp. 793–830, Mar. 2009.

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

[12] E. Moulines, J.-F. Cardoso, and E. Gassiat, "Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'97)*, Apr. 1997, pp. 3617–3620.

[13] H. Attias, "Independent factor analysis," *Neural Comput.*, vol. 11, pp. 803–851, 1999.

[14] H. Attias, "New EM algorithms for source separation and deconvolution," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'03)*, 2003, pp. 297–300.

[15] R. J. Weiss, M. I. Mandel, and D. P. W. Ellis, "Source separation based on binaural cues and source model constraints," in *Proc. Interspeech'08*, 2008, pp. 419–422.

[16] 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. Blind Source Separation (ICA'09)*, 2009, pp. 751–758.

[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. 11th Eur. Signal Process. Conf. (EUSIPCO'02)*, 2002, pp. 561–564.

[18] C. Févotte and J.-F. Cardoso, "Maximum likelihood approach for blind audio source separation using time-frequency Gaussian models," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA'05)*, Mohonk, NY, Oct. 2005, pp. 78–81.

[19] P. Smaragdis, "Efficient blind separation of convolved sound mixtures," in *IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA'97)*, New Paltz, NY, Oct. 1997, 4 pp..

[20] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS," in *IEEE Int. Symp. Circuits Syst. (ISCAS'07)*, May 27–30, 2007, pp. 3247–3250.

[21] M. I. Mandel, D. P. W. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," in *Adv. Neural Inf. Process. Syst. (NIPS 19)*, 2007, pp. 953–960.

[22] 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'07)*, Oct. 2007, pp. 147–150.

[23] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures. With application to blind audio source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'09)*, Taipei, Taiwan, Apr. 2009, pp. 3137–3140.

[24] F. D. Neeser and J. L. Massey, "Proper complex random processes with applications to information theory," *IEEE Trans. Inf. Theory*, vol. 39, no. 4, pp. 1293–1302, Jul. 1993.

[25] S. A. Abdallah and M. D. Plumbley, "Polyphonic transcription by non-negative sparse coding of power spectra," in *Proc. 5th Int. Symp. Music Inf. Retrieval (ISMIR'04)*, Oct. 2004, pp. 318–325.

[26] A. Cichocki, R. Zdunek, and S. Amari, "Csiszar's divergences for non-negative matrix factorization: Family of new algorithms," in *Proc. 6th Int. Conf. Ind. Compon. Anal. Blind Signal Separation (ICA'06)*, Charleston, SC, 2006, pp. 32–39.

[27] D.-T. Pham and J.-F. Cardoso, "Blind separation of instantaneous mixtures of non stationary sources," *IEEE Trans. Signal Process.*, vol. 49, no. 9, pp. 1837–1848, Sep. 2001.

[28] H. Laurberg, M. G. Christensen, M. D. Plumbley, L. K. Hansen, and S. H. Jensen, "Theorems on positive data: On the uniqueness of NMF," *Comput. Intell. Neurosci.*, vol. 2008, pp. 1–9, 2008.

[29] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1564–1578, Jul. 2007.

[30] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[31] M. Goodwin, "The STFT, sinusoidal models, and speech modification," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. New York: Springer, 2008, ch. 12, pp. 229–258.

[32] R. Bro, "PARAFAC. Tutorial and applications," *Chemometrics Intell. Lab. Syst.*, vol. 38, no. 2, pp. 149–171, Oct. 1997.

[33] M. Welling and M. Weber, "Positive tensor factorization," *Pattern Recognition Lett.*, vol. 22, no. 12, pp. 1255–1261, 2001.

[34] A. Shashua and T. Hazan, "Non-negative tensor factorization with applications to statistics and computer vision," in *Proc. 22nd Int. Conf. Mach. Learn.*, Bonn, Germany, 2005, pp. 792–799, ACM.

[35] Example Web Page [Online]. Available: http://www.irisa.fr/metiss/ozerov/demos.html#ieee_taslp09

[36] S. Hurley, Call for Remixes: Shannon Hurley [Online]. Available: http://www.ccmixter.org/shannon-hurley

[37] in *Signal Separation Evaluation Campaign (SiSEC 2008)*, 2008 [Online]. Available: http://www.sisec.wiki.irisa.fr

[38] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First stereo audio source separation evaluation campaign: Data, algorithms and results," in *Proc. Int. Conf. Ind. Compon. Anal. Blind Source Separation (ICA'07)*, 2007, pp. 552–559, Springer.

[39] D. Campbell, Roomsim Toolbox [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/5184

[40] E. Vincent, "Complex nonconvex lp norm minimization for underdetermined source separation," in *Proc. Int. Conf. Ind. Compon. Anal. Blind Source Separation (ICA'07)*, 2007, pp. 430–437.

[41] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a stereophonic linear instantaneous mixture," in *Proc. Int. Conf. Ind. Compon. Anal. Blind Source Separation (ICA'06)*, 2006, pp. 536–543.

[42] P. D. O'Grady and P. A. Pearlmutter, "Soft-LOST: EM on a mixture of oriented lines," in *Proc. Int. Conf. Ind. Compon. Anal. Blind Source Separation (ICA)*, 2004, pp. 428–435.

[43] in *Stereo Audio Source Separation Evaluation Campaign (SASSEC 2007)*, 2007 [Online]. Available: http://www.sassec.gforge.inria.fr/

[44] E. Vincent, S. Araki, and P. Bofilld, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," in *Proc. Int. Conf. Ind. Compon. Anal. Signal Separation (ICA'09)*, 2009, pp. 734–741 [Online]. Available: http://www.sassec.gforge.inria.fr/SiSEC_ICA09.pdf

[45] in *SiSEC 2008 Under-Determined Speech and Music Mixtures Task Results*, 2008 [Online]. Available: http://www.sassec.gforge.inria.fr/SiSEC_underdetermined/

[46] P. Smaragdis, B. Raj, and M. V. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proc. 7th Int. Conf. Ind. Compon. Anal. Signal Separation (ICA'07)*, London, U.K., Sep. 2007, pp. 414–421.

[47] in *SiSEC 2008 Professionally Produced Music Recordings Task Results*, 2008 [Online]. Available: http://www.sassec.gforge.inria.fr/SiSEC_professional/

[48] S. Winter, H. Sawada, S. Araki, and S. Makino, "Hierarchical clustering applied to overcomplete BSS for convolutive mixtures," in *Proc. ISCA Tutorial Research Workshop Statistical and Perceptual Audio Process. (SAPA 2004)*, Oct. 2004, pp. 652–660.

[49] "Audacity: The Free, Cross-Platform Sound Editor," [Online]. Available: http://www.audacity.sourceforge.net/

[50] J.-F. Cardoso and M. Martin, "A flexible component model for precision ICA," in *Proc. 7th Int. Conf. Ind. Compon. Anal. Signal Separation (ICA'07)*, London, U.K., Sep. 2007, pp. 1–8.

[51] Y. Lin and D. D. Lee, "Bayesian regularization and nonnegative deconvolution for room impulse response estimation," *IEEE Trans. Signal Process.*, vol. 54, no. 3, pp. 839–847, Mar. 2006.

[52] V. Y. F. Tan and C. Févotte, "Automatic relevance determination in nonnegative matrix factorization," in *Proc. Workshop Signal Process. Adaptative Sparse Structured Representations (SPARS'05)*, Saint-Malo, France, Apr. 2009.

[53] A. van den Bos, "Complex gradient and Hessian," *IEE Proc. Vision, Image, Signal Process.*, vol. 141, pp. 380–382, 1994.

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

Alexey Ozerov received the M.Sc. degree in mathematics from the Saint-Petersburg State University, Saint-Petersburg, Russia, in 1999, the M.Sc. degree in applied mathematics from the University of Bordeaux 1, Bordeaux, France, in 2003, and the Ph.D. degree in signal processing from the University of Rennes 1, Rennes, France, in 2006.

He worked towards the Ph.D. degree from 2003 to 2006 in the labs of France Telecom R&D and in collaboration with the IRISA institute. Earlier, From 1999 to 2002, he worked at Terayon Communication Systems as a R&D Software Engineer, first in Saint-Petersburg and then in Prague, Czech Republic. He was for one year (2007) in the Sound and Image Processing Lab at KTH (Royal Institute of Technology), Stockholm, Sweden, and for one year and half (2008-2009) with TELECOM ParisTech / CNRS LTCI—Signal and Image Processing (TSI) Department. Currently, he is with the METISS team of IRISA/INRIA—Rennes as a Postdoctoral Researcher. His research interests include audio source separation, source coding, and automatic speech recognition.

Cédric Févotte received the State Engineering degree and the M.Sc. degree in control and computer science from École Centrale de Nantes, Nantes, France, in 2000, and the Ph.D. degree in 2003 from the University of Nantes.

From 2003 to 2006, he was a Research Associate with the Signal Processing Laboratory at the University of Cambridge, Cambridge, U.K., working on Bayesian approaches to audio signal processing tasks such as audio source separation, denoising, and feature extraction. From May 2006 to February 2007, he was a Research Engineer with the start-up company Mist-Technologies (Paris), working on mono/stereo to 5.1 surround sound upmix solutions. In March 2007, he joined CNRS LTCI/Telecom ParisTech, first as a Research Associate and then as a CNRS tenured Research Scientist in November 2007. His research interests generally concern statistical signal processing and unsupervised machine learning with audio applications.