Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation

Factorisation en matrices à coefficients positifs de données multicanal convolutives pour la séparation de sources audio

---

Alexey Ozerov
Cédric Févotte

2009D001

Janvier 2009

Département Traitement du Signal et des Images
Groupe AAO : Audio, Acoustique et Ondes

Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation

Factorisation en matrices à coefficients positifs de données multicanal convolutives pour la séparation de sources audio

Alexey Ozerov[1] and Cédric Févotte[2]

[1] Institut TELECOM; TELECOM ParisTech; CNRS LTCI
[2] CNRS LTCI; TELECOM ParisTech
37-39, rue Dareau, 75014 Paris, France.
{alexey.ozerov,cedric.fevotte}@telecom-paristech.fr

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).

Keywords: Multichannel audio, nonnegative matrix factorization, nonnegative tensor factorization, expectation-maximization algorithm, underdetermined convolutive blind source separation.

Résumé

Nous considérons le problème de l'estimation de représentations objet adaptatives de données audio multicanal, supposées générées par un mélange éventuellement sous-déterminé et convolutif de signaux sources. Le domaine de modélisation est le domaine de la Transformée de Fourier Court-Terme (TFCT), dans lequel la convolution peut être approchée par des mélanges linéaires instantanés dans chaque sous-bande fréquentielle. La TFCT de chaque source est modélisée par un modèle de type factorisation en matrices non-négatives avec la divergence d'Itakura-Saito, qui sous-tend une modélisation statistique de type gaussienne composite. Nous proposons deux méthodes d'estimation des paramètres de mélange et des sources. La première méthode consiste à maximiser la vraisemblance conjointe exacte des données avec un algorithme espérance-maximisation (EM). La deuxième méthode consiste à maximiser la somme des vraisemblances individuelles de chaque canal avec un algorithme de mises à jour multiplicatives. Nos algorithmes de décomposition sont appliqués au problème de la séparation de mélanges audio stéréo, dans différentes configurations : séparation aveugle et supervisée, séparation de sources musicales et de parole, mélanges synthétiques instantanés et convolutifs, ainsi que séparation d'enregistrements musicaux professionnels produits en studio. Notre méthode EM donne des résultats compétitifs par rapport à l'état de l'art, comme l'illustrent ses performances sur deux tâches de la campagne internationale d'évaluation de séparation de sources SiSEC 2008 (Signal Separation Evaluation Campaign).

Mots-clés: Audio multicanal, factorisation en matrices à coefficients positifs, factorisation en tenseurs à coefficients positifs, algorithme espérance-maximisation, séparation aveugle de sources en mélanges convolutifs et sous-déterminés.

Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation

Alexey Ozerov and Cédric Févotte

***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*—Multichannel audio, nonnegative matrix factorization, nonnegative tensor factorization, expectation-maximization algorithm, underdetermined convolutive blind source separation.**

I. Introduction

NONNEGATIVE matrix factorization (NMF) is a linear regression 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 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 *source* signals $\tilde{s}_j(t)$ $(i = 1, \ldots, J)$ such that

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

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 equation (1) can be approximated in the Short-Time Fourier Transform (STFT) domain as

$$x_{i,fn} = \sum\nolimits_{j=1}^{J} a_{ij,f}\, s_{j,fn} + b_{i,fn}, \qquad (2)$$

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

$$\mathbf{x}_{fn} = \mathbf{A}_f\, \mathbf{s}_{fn} + \mathbf{b}_{fn}, \qquad (3)$$

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 $\mathbf{A}_f = [a_{ij,f}]_{ij} \in \mathbb{C}^{I \times J}$.

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 $\mathbf{W}_j$ et $\mathbf{H}_j$, such that

$$|\mathbf{S}_j|^2 \approx \mathbf{W}_j \mathbf{H}_j. \qquad (4)$$

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 on Fig. 1. Our problem splits into two subtasks: (i) defining suitable estimation criteria, and (ii) designing algorithms optimizing these criteria.

A. Ozerov is with Institut TELECOM, TELECOM ParisTech, CNRS LTCI, 37-39, rue Dareau, 75014 Paris, France (e-mail: alexey.ozerov@telecom-paristech.fr).

C. Févotte is with CNRS LTCI, TELECOM ParisTech, 37-39, rue Dareau, 75014 Paris, France (e-mail: cedric.fevotte@telecom-paristech.fr).

This work was supported in part by the French ANR project SARAH (*StAndardisation du Remastering Audio Haute-Définition*).

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 exploit 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. Remix of well-known songs retrieved from commercial CD recordings are proposed in the results section.

Many convolutive blind source separation (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., prior knowledge about filters is used. 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 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 may be considered not suitable for audio because they do not model temporal correlations across signal samples. In [14], Attias proposes to model the sources in the STFT domain using multivariate GMMs, hence taking into account temporal correlations in audio signals (assumed stationary in each window frame). He 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, involving a computationally efficient strategy for learning GMMs independently from 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 for musical signals than the GMM. Moreover, the computational complexity of inference in our model grows linearly with the number of components while the complexity of exact inference in GMMs grows combinatorially. Also, our method is fully adaptative (blind) and is applicable to the general case of convolutive noisy mixtures and covers both the (over)determined ($I \geq J$) and under-determined ($I < J$) cases.

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 the results of application 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 the 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 non-trivial 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

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

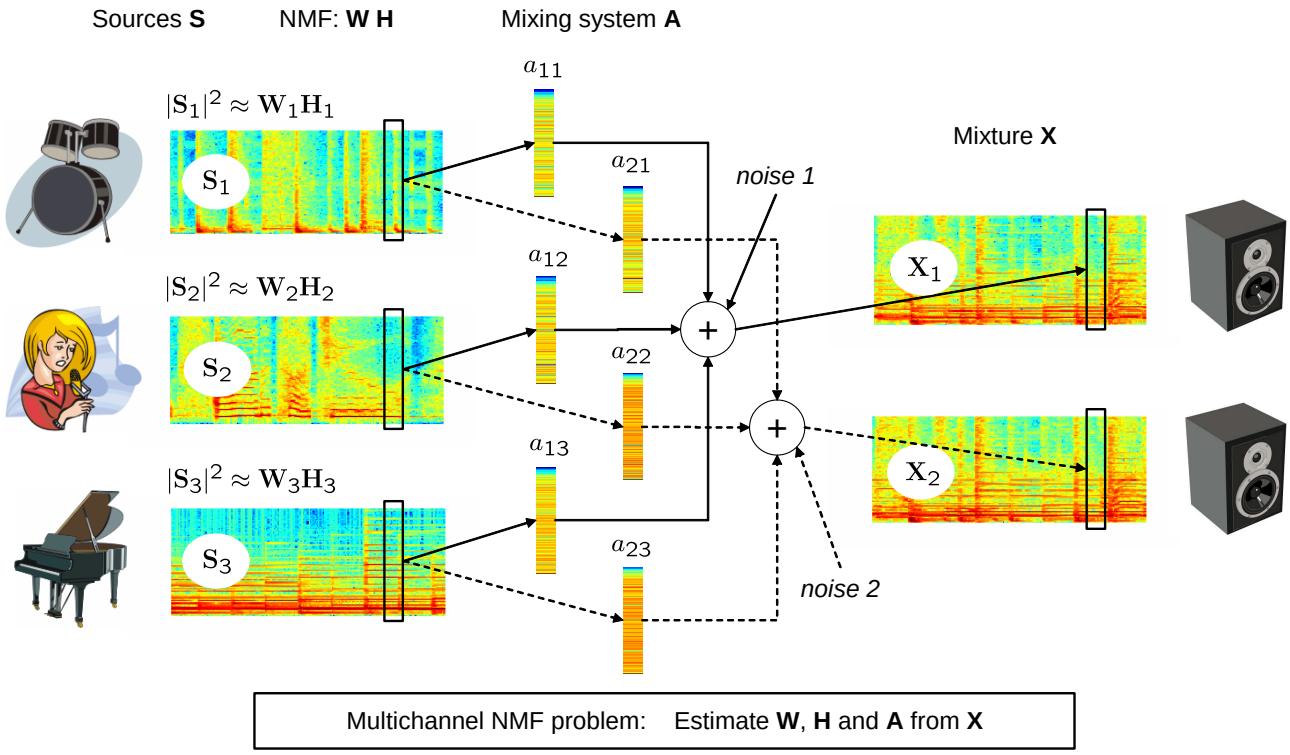

Fig. 1. Representation of convolutive mixing system and formulation of Multichannel NMF problem.

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

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

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

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

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 log-likelihood of the parameters describing source j writes

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

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 [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 implied by our model grows linearly with the number of components while the complexity of exact inference in the GMM grows combinatorially with the number of states. EM algorithms proposed in [14] and [16] for GMM have linear complexity as well, but at the price of approximate inference. 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 only 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 $\boldsymbol{\Sigma}_{\mathbf{b},f} = \mathrm{diag}([\sigma^2_{i,f}]_i)$. The musical data we consider in Section IV-A are not noisy in the common 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 slow convergence and potential

numerical instabilities, as discussed later. 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

The mixing model (3) can be recast as

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

where $\mathbf{c}_{fn} = [c_{1,fn}, \ldots, c_{K,fn}]^T \in \mathbb{C}^{K \times 1}$ and $\vec{\mathbf{A}}_f$ is the "extended mixing matrix" of dimension $I \times K$, with elements defined by $\vec{a}_{ik,f} = a_{ij,f}$ 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 arrive from same directions). Subsequently, we will note $\mathbf{\Sigma}_{\mathbf{c},fn} = \mathrm{diag}\left([w_{fk} h_{kn}]_k\right)$ 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 variance parameters. Under previous assumptions, data $\mathbf{x}_{fn}$ 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

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

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

For a fixed f, the BSS problem described by Eq. (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 scale, phase and permutation indeterminacies. 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

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

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).

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 Appendix of [28]) 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

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

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

is shown to be a *natural (sufficient) statistics* [28] 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 re-estimating the parameters using the updated natural statistics, which amounts to maximizing the conditional expectation of the complete data 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.

4) Implementation issues: The computation of the source Wiener gain $\mathbf{G}_{\mathbf{s},fn}$ given by Eq. (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 [29]

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

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 $\mathbf{G}_{\mathbf{c},fn}$, Eq. (20), if

[1] In [17], [27] the ML criterion can be recast as a measure of fit between observed and parameterized covariances, where the measure of deviation writes $D(\mathbf{\Sigma}_1|\mathbf{\Sigma}_2) = \mathrm{trace}\left(\mathbf{\Sigma}_1 \, \mathbf{\Sigma}_2^{-1}\right) - \log\det \mathbf{\Sigma}_1 \, \mathbf{\Sigma}_2^{-1} - I$ 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 $I = 1$). This is in fact the Kullback-Leibler divergence between 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 $I > 1$.

Algorithm 1 EM algorithm (one iteration)

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

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

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

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

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

  where

$$\hat{\mathbf{s}}_{fn} = \mathbf{G}_{\mathbf{s},fn}\mathbf{x}_{fn}, \qquad \mathbf{G}_{\mathbf{s},fn} = \mathbf{\Sigma}_{\mathbf{s},fn}\mathbf{A}_f^H\mathbf{\Sigma}_{\mathbf{x},fn}^{-1}, \tag{19}$$

$$\hat{\mathbf{c}}_{fn} = \mathbf{G}_{\mathbf{c},fn}\mathbf{x}_{fn}, \qquad \mathbf{G}_{\mathbf{c},fn} = \mathbf{\Sigma}_{\mathbf{c},fn}\vec{\mathbf{A}}_f^H\mathbf{\Sigma}_{\mathbf{x},fn}^{-1}, \tag{20}$$

$$\mathbf{\Sigma}_{\mathbf{x},fn} = \mathbf{A}_f\mathbf{\Sigma}_{\mathbf{s},fn}\mathbf{A}_f^H + \mathbf{\Sigma}_{\mathbf{b},f} \tag{21}$$

$$\mathbf{\Sigma}_{\mathbf{s},fn} = \mathrm{diag}\left(\left[\sum\nolimits_{k\in\mathcal{K}_j} w_{fk}h_{kn}\right]_j\right) \tag{22}$$

$$\mathbf{\Sigma}_{\mathbf{c},fn} = \mathrm{diag}\left(\left[w_{fk}h_{kn}\right]_k\right) \tag{23}$$

  and $\vec{\mathbf{A}}_f$ is defined in Sec. II-C.
- **M step**. Update the parameters:

$$\mathbf{A}_f = \widehat{\mathbf{R}}_{\mathbf{xs},f}\widehat{\mathbf{R}}_{\mathbf{ss},f}^{-1}, \tag{24}$$

$$\mathbf{\Sigma}_{\mathbf{b},f} = \mathrm{diag}\left(\widehat{\mathbf{R}}_{\mathbf{xx},f} - \mathbf{A}_f\widehat{\mathbf{R}}_{\mathbf{xs},f}^H - \widehat{\mathbf{R}}_{\mathbf{xs},f}\mathbf{A}_f^H + \mathbf{A}_f\widehat{\mathbf{R}}_{\mathbf{ss},f}\mathbf{A}_f^H\right), \tag{25}$$

$$w_{fk} = \frac{1}{N}\sum_n \frac{\hat{u}_{k,fn}}{h_{kn}}, \qquad h_{kn} = \frac{1}{F}\sum_f \frac{\hat{u}_{k,fn}}{w_{fk}}. \tag{26}$$

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

$I > K$. 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 Eq. (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 Eq. (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, Eq. (24) must be replaced by:

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

6) Simulated annealing: If one computes $\mathbf{A}_f$ through equations (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: Wiener reconstructions of the source STFTs are directly retrieved from Eq. (19). Time-domain sources may then be obtained through inverse STFT using an adequate overlap-add procedure with dual synthesis window. 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 Eq. (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

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

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

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

with $q_{ij,f} = |a_{ij,f}|^2$. 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.

Our approach differs from the NTF approach of [6], [7] where the following PARAFAC structure [30] is considered

$$\hat{v}_{i,fn}^{NTF} = \sum\nolimits_k q_{ik}^{NTF}\,w_{fk}\,h_{kn}. \tag{32}$$

It is only a sum of $I \times F \times N$ rank-1 tensors and amounts to assuming that $\hat{\mathbf{V}}_i^{NTF} = [\hat{v}_{i,fn}^{NTF}]_{fn}$ 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 post-processing binding step in order to group the K elementary patterns into J sources, based on clustering of the ratios $\{q_{1k}^{NTF}/q_{2k}^{NTF}\}_k$ (in the stereo case). To ease comparison, our model can be rewritten as

$$\hat{v}_{i,fn} = \sum\nolimits_k \vec{q}_{ik,f}\, w_{fk}\, h_{kn} \tag{33}$$

subject to the constraint $\vec{q}_{ik,f} = q_{ij,f}$ iif $k \in \mathcal{K}_j$ (with the notation introduced in Section II-C, we have also $\vec{q}_{ik,f} = |\vec{a}_{ik,f}|^2$). Hence, our model has the following merits w.r.t. the PARAFAC-NTF model: (i) it accounts for convolutive mixing by considering frequency-dependent mixing proportions ($\vec{q}_{ik,f}$ instead of q_{ik}^{NTF}) and (ii) the constraint that the K mixing proportions $\{\vec{q}_{ik,f}\}_k$ 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 Eq. (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 spectrogram [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], [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], [31], [32]. 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

$$\theta_l \leftarrow \theta_l \, \frac{[\nabla_{\theta_l} C_2(\boldsymbol{\theta})]_-}{[\nabla_{\theta_l} C_2(\boldsymbol{\theta})]_+}, \tag{34}$$

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 nonnegativity 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.

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

$$q_{ij} \leftarrow q_{ij} . \frac{\text{sum}\left[\hat{\mathbf{V}}_i^{-2}.\mathbf{V}_i.(\mathbf{W}_j\mathbf{H}_j)\right]}{\text{sum}\left[\hat{\mathbf{V}}_i^{-1}.(\mathbf{W}_j\mathbf{H}_j)\right]} \tag{38}$$

where sum$[\mathbf{M}]$ is the sum of all coefficients in $\mathbf{M}$. Then, $\text{diag}(\mathbf{q}_{ij})$ needs only be replaced by q_{ij} in Eq. (36) and (37). 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.

5) Reconstruction of the source images: While the joint-likelihood EM optimization setting provides a mean of reconstructing the source STFTs $\mathbf{s}_{fn}$ in a principled way using Wiener filtering, it is not obvious how this should be done in the present setting where only the sum of individual likelihoods is maximized. The most natural way is to reconstruct an image $s_{ij,fn}^{im}$ of source j in channel i through

$$\hat{s}_{ij,fn}^{im} = \frac{q_{ij,f} p_{i,fn}}{\hat{v}_{i,fn}}\, x_{i,fn}, \tag{39}$$

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., $\sum_j \hat{s}_{ij,fn}^{im} + \hat{b}_{i,fn} = x_{i,fn}$. We have also tried other reconstruction schemes consisting of forming an estimate of $\mathbf{A}_f$ from its squared absolute values $\mathbf{Q}_f$ (e.g., $a_{ij,f} = \sqrt{q_{ij,f}}$) and then applying Wiener estimation (19), but they proved less satisfying.

Algorithm 2 MU rules (one iteration)

- Update $\mathbf{Q}$

$$\mathbf{q}_{ij} \leftarrow \mathbf{q}_{ij} . \frac{\left[\hat{\mathbf{V}}_i^{-2}.\mathbf{V}_i.(\mathbf{W}_j\mathbf{H}_j)\right]\mathbf{1}_{N\times 1}}{\left[\hat{\mathbf{V}}_i^{-1}.(\mathbf{W}_j\mathbf{H}_j)\right]\mathbf{1}_{N\times 1}} \tag{35}$$

- Update $\mathbf{W}$

$$\mathbf{W}_j \leftarrow \mathbf{W}_j . \frac{\sum_{i=1}^I \text{diag}(\mathbf{q}_{ij})(\hat{\mathbf{V}}_i^{-2}.\mathbf{V}_i)\mathbf{H}_j^T}{\sum_{i=1}^I \text{diag}(\mathbf{q}_{ij})\hat{\mathbf{V}}_i^{-1}\mathbf{H}_j^T} \tag{36}$$

- Update $\mathbf{H}$

$$\mathbf{H}_j \leftarrow \mathbf{H}_j . \frac{\sum_{i=1}^I (\text{diag}(\mathbf{q}_{ij})\mathbf{W}_j)^T(\hat{\mathbf{V}}_i^{-2}.\mathbf{V}_i)}{\sum_{i=1}^I (\text{diag}(\mathbf{q}_{ij})\mathbf{W}_j)^T\hat{\mathbf{V}}_i^{-1}} \tag{37}$$

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

IV. Experiments

In this section we first describe the test data and evaluation criteria, and then we proceed with experiments. All the audio datasets and separation results are available from our demo webpage [33].

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 seconds-excerpts of original separated tracks from the song "Sunrise" by S. Hurley, available under a Creative Commons License at [34] 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) "underdetermined speech and music mixtures" task development datasets [35], 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) [36]. It also coincides with development dataset `dev2` of SiSEC'08 "underdetermined speech and music mixtures" task. All the mixtures are 10 seconds-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 [37]. They simulate a pair of omnidirectional microphones placed 1 m apart in a room of dimensions 4.45 x 3.55 x 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 degrees with a minimal spacing of 15 degrees (for more details see [35]).
- **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 seconds-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 seconds-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 seconds taken from three professionally produced stereo recordings of well-known pop and reggae songs, and downsampled to 22 kHz.

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) of reconstructed source images described in [36], which is a global measure unifying the Image to Spatial distortion Ratio (ISR), the Source to Interference Ratio (SIR) and the Sources to Artifacts Ratio (SAR) [36]. To assess the quality of the mixing system estimates we used the Mixing Error Ratio (MER) described at [35], which is an SNR-like criterion expressed in decibels. MATLAB routines for computing these criteria were obtained from the SiSEC'08 webpage [35].

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 & D), separation performance can only be assessed perceptually by listening to the separated source images, available online at [33].

C. STFT parameters

In all the experiments below we used STFTs with half-overlapping sine windows, using the STFT computation tools for MATLAB available from [35]. The choice of the STFT window size is rather important, and is a matter a compromise between (i) good frequency resolution and validity of the convolutive mixing approximation (2) and (ii) validity of the assumption of source local stationnarity. We have tried various window sizes (from powers of 2 in samples) for every experiment, with size yielding best separation results given in Table I.

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

TABLE I
STFT WINDOW LENGTHS USED IN DIFFERENT EXPERIMENTS.

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 is very sensitive to parameters initialization, and to be sure that we have a "good initialization", we provide it with *perturbed oracle initializations*: 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)**: $\mathbf{\Sigma}_{\mathbf{b},f} = \mathrm{diag}([\sigma^2_{i,f}]_i)$ is reestimated with update Eq. (25).
- **(F)**: Noise covariance is reestimated like in scheme E, but under the more constrained structure $\mathbf{\Sigma}_{\mathbf{b},f} = \sigma^2_f \mathbf{I}_I$ (isotropic noise in each subband). In that case, operator $\mathrm{diag}(\cdot)$ in Eq. (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 Figure 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 stagnates. 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, and we will thus use this scheme in the rest of experiments.

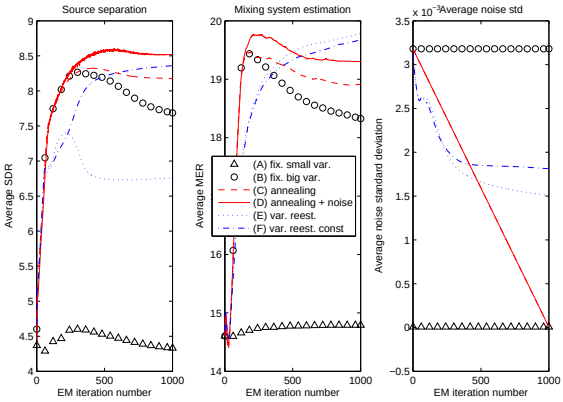

Fig. 2. EM algorithm results on convolutive mixture of dataset A, using various noise variance update schemes. (Left) Average source separation SDR, (Middle) Average mixing system identification MER, (Right) Average noise standard deviation. (A) triangles: small fixed noise variance, (B) circles: larger 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.

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 numerical criteria defined in [36]. We used both mixtures of dataset A (instantaneous and convolutive) and ran 1000 iterations of both algorithms (EM and MU) from 10 different perturbed oracle initializations, obtained as in previous section. Again we used $K = 12$ components, equally split into $J = 3$ sources. Figures 3 and 4 reports results for the instantaneous and convolutive mixtures, respectively. Plots on top row display in log-scale the cost functions $C_1(\boldsymbol{\theta})$ and $C_2(\boldsymbol{\theta})$ w.r.t. iterations for all 10 runs. Note that cost $C_1(\boldsymbol{\theta})$ is not positive in general, see Eq. (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 seconds stereo mixture (sampled at 16 kHz), $J = 3$ sources, and $K = 12$ components.

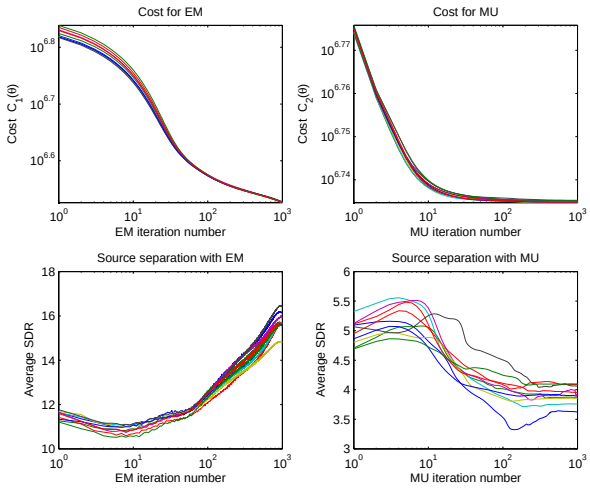

Fig. 3. 10 runs of EM and MU from 10 perturbed oracle initializations using instantaneous mixture of dataset A. (Top) Cost functions, (Bottom) Average SDRs.

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 mixtures. We used exact same data and evaluation criteria. More precisely, our algorithms are compared in the instantaneous case to the method of Vincent [38], 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.* [39]. In the convolutive case, our algorithms are compared to the method of Sawada,

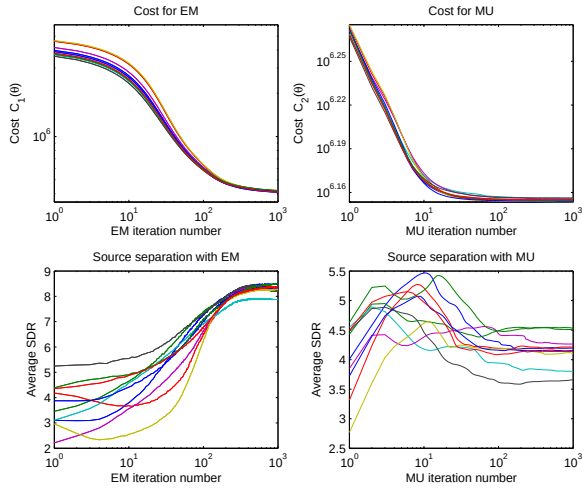

Fig. 4. 10 runs of EM and MU from 10 perturbed oracle initializations using convolutive mixture of dataset A. (Top) Cost functions, (Bottom) Average SDRs.

based on frequency-dependent complex-valued mixing matrices estimation [40], 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 [38] and [39]. In the convolutive case, we simply downloaded the source image estimates from the SASSEC'07 webpage [41]. In both cases we built initialization of $\mathbf{W}$ and $\mathbf{H}$ based on NMF of the source spectrogram estimates.[2]

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 are 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.[3] 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 webpage [33]. We have also participated with our EM algorithm in "under-determined speech and music mixtures" task of SiSEC'08 for instantaneous mixtures, and our results can be compared to other methods at [4] and [5].

[2]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.

[3]The reference algorithms performances in Table II do not always coincide with those given on the SASSEC'07 webpage [41]. In the instantaneous case this is because we have not used the exact same implementation of the l_0 minimization algorithm [38] 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.

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 $\{a^{tr}_{ij,f}\}_{if}$, spectral patterns $\mathbf{W}^{tr}_j$ and activation coefficients $\mathbf{H}^{tr}_j$ from available training signal images of source j (using 200 iterations of EM/MU); discard $\mathbf{H}^{tr}_j$,
- 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], [42].

One important issue with professionally produced modern music mixtures is that they do not always comply with the mixing assumptions of Eq. (3). This might be due to non-linear sound effects (e.g., dynamic range compression), reverberation times longer that 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 3 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 6 components, the piano part is modeled as a convolutive point source image with 6 components and finally, the residual background music (sum of remaining tracks) is modeled as a sum of 3 convolutive point-source images with 4 components. The audio results, available at [33], again illustrate the better performance of the EM approach. Our results can be compared to other methods

[4]http://sassec.gforge.inria.fr/SiSEC_underdetermined/test_eval.html

[5]http://sassec.gforge.inria.fr/SiSEC_underdetermined/dev2_eval.html

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

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

that entered the "professionally produced music recordings" task of SiSEC'08 at [6].

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 [43], which yielded satisfactory results:

- Stack left and right mixture STFTs so as to create a $2F \times N$ complex-valued matrix $\mathbf{X}_{2\mathrm{ch}} = [\mathbf{X}_L^T \; \mathbf{X}_R^T]^T$.
- Produce a K-components IS-NMF decomposition of $|\mathbf{X}_{2\mathrm{ch}}|^2 \approx \mathbf{W}_{2\mathrm{ch}}\mathbf{H}_{2\mathrm{ch}}$.
- Initialize $\mathbf{W}$ as the average of $\mathbf{W}_L$ and $\mathbf{W}_R$, where $\mathbf{W}_{2\mathrm{ch}} = [\mathbf{W}_L^T \; \mathbf{W}_R^T]^T$. Initialize $\mathbf{H} = \mathbf{H}_{2\mathrm{ch}}$.
- Reconstruct K components $\widehat{\mathbf{C}}_{2\mathrm{ch},k} = [\widehat{\mathbf{C}}_{L,k}^T \; \widehat{\mathbf{C}}_{R,k}^T]^T$ from $\mathbf{X}_{2\mathrm{ch}}$, $\mathbf{W}_{2\mathrm{ch}}$ and $\mathbf{H}_{2\mathrm{ch}}$, using Wiener filtering. 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 $\widehat{\mathbf{A}}$ (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. Interestingly, on these specific examples the superiority of the EM method w.r.t. the MU method is not as clear as with previous datasets. One reason could be the existence of non-point sources breaking the validity of mixing assumptions (3). 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 non-point 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 (i) sources mixed in the same directions, and thus impossible to separate in our fully blind setting, (ii) non-point 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 Eq. (39) 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 originating from same (or close) directions, 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 [44] 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 [33].

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 likelihood and a MU algorithm for the maximization of the sum of individual channel 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.

[6] http://sassec.gforge.inria.fr/SiSEC_professional/

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. As expected, the EM method gives better results in terms of separation performance than the MU method in most cases, confirming the importance of keeping between-channel dependencies in the optimization criterion.

The EM algorithm was also shown to outperform state-of-the-art methods, given appropriate initializations. 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 rather than speech: music sources need only be represented by a small number of components to attain good separation performance, and informal listening indicates better separation of music signals.

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 [45]. 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 [46]. Minor changes are required in the MU rules so as to yield convergent algorithms for MAP estimation. More complex priors structure can also be envisaged within the EM method, such as Gamma Markov chains favoring smoothness [10].

We have found the number of components $\#\mathcal{K}_j$ per source j difficult to choose. Underestimating it may lead to poor results, while overestimating it increases the degrees of freedom in the model, favoring the existence of local minima in the criteria and thus rendering initialization difficult. The number of sources J may be itself be difficult to choose, for example when dealing with non point-source as discussed in Section IV-G. Exploring ideas from *automatic relevance determination* (see [47] in a NMF setting), an interesting line of research will consist of fixing a total number of components K (a budget) and design algorithms that let data self-assign a relevant number of components to a self-determined number of clusters (DOAs).

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 would 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 non-point sources, non-linear sound effects, such as dynamic range compression, and long reverberation times.

APPENDIX A
EM ALGORITHM DERIVATION OUTLINE

The complete data log-likelihood can be written as:

$$\begin{aligned}
&-\log p(\mathbf{X},\mathbf{C}|\boldsymbol{\theta}) = -\log p(\mathbf{X}|\mathbf{C},\boldsymbol{\theta}) - \log p(\mathbf{C}|\boldsymbol{\theta}) \\
&\stackrel{\text{c}}{=} \sum_{fn}\left[\log|\boldsymbol{\Sigma}_{\mathbf{b},f}| + (\mathbf{x}_{fn}-\mathbf{A}_f\mathbf{s}_{fn})^H\boldsymbol{\Sigma}_{\mathbf{b},f}^{-1}(\mathbf{x}_{fn}-\mathbf{A}_f\mathbf{s}_{fn})\right] \\
&\qquad + \sum_k\sum_{fn}\left[\log(h_{k,n}w_{k,f}) + \frac{|c_{k,fn}|^2}{h_{k,n}w_{k,f}}\right] \\
&= \sum_{fn}\left[\log|\boldsymbol{\Sigma}_{\mathbf{b},f}| + \sum_k\log(h_{k,n}w_{k,f}) + \sum_k\frac{|c_{k,fn}|^2}{h_{k,n}w_{k,f}}\right] \\
&\qquad + N\sum_f \operatorname{trace}\Big[\boldsymbol{\Sigma}_{\mathbf{b},f}^{-1}\mathbf{R}_{\mathbf{xx},f} - \boldsymbol{\Sigma}_{\mathbf{b},f}^{-1}\mathbf{A}_f\mathbf{R}_{\mathbf{xs},f}^H \\
&\qquad\qquad -\boldsymbol{\Sigma}_{\mathbf{b},f}^{-1}\mathbf{R}_{\mathbf{xs},f}\mathbf{A}_f^H + \boldsymbol{\Sigma}_{\mathbf{b},f}^{-1}\mathbf{A}_f\mathbf{R}_{\mathbf{ss},f}\mathbf{A}_f^H\Big], \quad (40)
\end{aligned}$$

with $\mathbf{R}_{\mathbf{xx},f}$, $\mathbf{R}_{\mathbf{xs},f}$, $\mathbf{R}_{\mathbf{ss},f}$ and $u_{k,fn}$ defined by Eqs. (15), (16), (17) and (18). Thus, we have shown that the complete data log-likelihood can be represented in the following form:

$$\log p(\mathbf{X},\mathbf{C}|\boldsymbol{\theta}) = \langle\eta(\boldsymbol{\theta}),\mathbf{T}(\mathbf{X},\mathbf{C})\rangle + \nu(\boldsymbol{\theta}), \quad (41)$$

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], [28]) and complete data statistics $\mathbf{t}(\mathbf{X},\mathbf{C})$ is a *natural (sufficient) statistics* [11], [28] for this family. To derive an EM algorithm in this special case one needs to (i) solve complete data ML criterion (thanks to (41) this solution can be always expressed as a function of natural statistics $\mathbf{t}(\mathbf{X},\mathbf{C})$), and (ii) 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 $\log p(\mathbf{X},\mathbf{C}|\boldsymbol{\theta})$ (Eq. (40)) w.r.t. model parameters $\boldsymbol{\theta}$ (see [48] for issues regarding derivativion 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 [7]:

$$\mathbf{A}_f = \mathbf{R}_{\mathbf{xs},f}\mathbf{R}_{\mathbf{ss},f}^{-1}, \quad (42)$$

$$\begin{aligned}
\boldsymbol{\Sigma}_{\mathbf{b},f} = \operatorname{diag}\big(&\mathbf{R}_{\mathbf{xx},f} - \mathbf{A}_f\mathbf{R}_{\mathbf{xs},f}^H - \mathbf{R}_{\mathbf{xs},f}\mathbf{A}_f^H \\
&+\mathbf{A}_f\mathbf{R}_{\mathbf{ss},f}\mathbf{A}_f^H\big), \quad (43)
\end{aligned}$$

$$w_{fk} = \frac{1}{N}\sum_n\frac{u_{k,fn}}{h_{kn}}, \qquad h_{kn} = \frac{1}{F}\sum_f\frac{u_{k,fn}}{w_{fk}}. \quad (44)$$

[7] 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 (40) and solving the corresponding complete data MAP criterion.

Our EM algorithm is strictly speaking only a *Generalized* EM algorithm [49] because it only ensures $Q(\boldsymbol{\theta}^{m+1}|\boldsymbol{\theta}^m) \geq Q(\boldsymbol{\theta}^m|\boldsymbol{\theta}^m)$. Indeed, in Eq. (44) $\mathbf{W}$ 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.:

$$p(\mathbf{s}_{fn}|\mathbf{x}_{fn}; \boldsymbol{\theta}) = N_c(\mathbf{s}_{fn}; \hat{\mathbf{s}}_{fn}, \boldsymbol{\Sigma}^{\text{post}}_{\mathbf{s},fn}), \tag{45}$$

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

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

Computing conditional expectations of $\mathbf{R}_{\mathbf{xs},f}$ and $\mathbf{R}_{\mathbf{ss},f}$ using (45) leads to equations (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 $C_2(\boldsymbol{\theta})$ (Eq. (30)) w.r.t. θ simply writes

$$\nabla_\theta D(\mathbf{V}|\hat{\mathbf{V}}) = \sum_{ifn} (\nabla_\theta \hat{v}_{i,fn})\; d'_{IS}(v_{i,fn}|\hat{v}_{i,fn}) \tag{46}$$

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

$$d'_{IS}(x|y) = \frac{1}{y} - \frac{x}{y^2}. \tag{47}$$

Using Eq. (46), 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

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

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

ACKNOWLEDGMENTS

The authors would like to thank S. Arberet for kindly sharing his implementation of DEMIX algorithm [39], as well as all the organizers of SiSEC'08 for well-prepared evaluation campaign.

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 *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2003.

[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. International Conference on Acoustics, Speech and Signal Processing (ICASSP'07)*, Honolulu, Hawaii, USA, 2007.

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

[5] P. Smaragdis, "Convolutive speech bases and their application to speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, 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 International Conference on Independent Component Analysis and Blind Signal Separation (ICA'06)*, Charleston SC, USA, Mar. 2006, pp. 666–673.

[7] D. FitzGerald, M. Cranitch, and E. Coyle, "Non-negative tensor factorisation for sound source separation," in *Proc. of the Irish Signals and Systems Conference*, Dublin, Sep. 2005.

[8] L. Parra and C. Spence, "Convolutive blind source separation of non-stationary sources," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, 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 International Conference on Acoustics, Speech and Signal Processing (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 Computation*, vol. 21, no. 3, Mar. 2008, preprint at http://www.tsi.enst.fr/~fevotte/TechRep/techrep08_is-nmf.pdf.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin., "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. 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 International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97)*, April 1997.

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

[14] ——, "New EM algorithms for source separation and deconvolution," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*, 2003.

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

[16] S. Arberet, A. Ozerov, R. Gribonval, and F. Bimbot, "Blind spectral-GMM estimation for underdetermined instantaneous audio source separation," in *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA'09)*, 2009, submitted.

[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. 11^e European Signal Processing Conference (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 on Applications of Signal Processing to Audio and Acoustics (WASPAA'05)*, Mohonk, NY, USA, Oct. 2005.

[19] P. Smaragdis, "Efficient blind separation of convolved sound mixtures," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'97)*, New Paltz, NY, Oct. 1997.

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

[21] M. Mandel, D. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," in *Advances in Neural Information Processing Systems (NIPS 19)*, 2007.

[22] Y. Izumi, N. Ono, and S. Sagayama, "Sparseness-based 2CH BSS using the EM algorithm in reverberant environment," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (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 *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'09)*, 2009, to appear.

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

[25] S. A. Abdallah and M. D. Plumbley, "Polyphonic transcription by nonnegative sparse coding of power spectra," in *Proc. 5th International Symposium Music Information Retrieval (ISMIR'04)*, Oct. 2004, pp. 318–325.

[26] A. Cichocki, R. Zdunek, and S. Amari, "Csiszar's divergences for nonnegative matrix factorization: Family of new algorithms," in *Proc. 6th International Conference on Independent Component Analysis and Blind Signal Separation (ICA'06)*, Charleston SC, USA, 2006, pp. 32–39.

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

[28] 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. on Audio, Speech and Language Proc.*, vol. 15, no. 5, pp. 1564–1578, July 2007.

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

[30] R. Bro, "PARAFAC. Tutorial and applications," *Chemometrics and Intelligent Laboratory Systems*, vol. 38, no. 2, pp. 149–171, Oct. 1997.

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

[32] A. Shashua and T. Hazan, "Non-negative tensor factorization with applications to statistics and computer vision," in *Proc. 22nd International Conference on Machine learning*. Bonn, Germany: ACM, 2005, pp. 792 – 799.

[33] "Example web page." [Online]. Available: http://perso.telecom-paristech.fr/~ozerov/demos.html#taslp09

[34] S. Hurley, "Call for remixes: Shannon Hurley." [Online]. Available: http://ccmixter.org/shannon-hurley

[35] "Signal Separation Evaluation Campaign (SiSEC 2008)," 2008. [Online]. Available: http://sisec.wiki.irisa.fr

[36] 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. on Independent Component Analysis and Blind Source Separation (ICA'07)*. Springer, 2007, pp. 552–559.

[37] D. Campbell, "Roomsim toolbox." [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/5184

[38] E. Vincent, "Complex nonconvex lp norm minimization for underdetermined source separation," in *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA'07)*, 2007.

[39] 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. on Independent Component Analysis and Blind Source Separation (ICA'06)*, 2006, pp. 536–543.

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

[41] "Stereo Audio Source Separation Evaluation Campaign (SASSEC 2007)," 2007. [Online]. Available: http://sassec.gforge.inria.fr/

[42] P. Smaragdis, B. Raj, and M. V. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proc. 7th International Conference on Independent Component Analysis and Signal Separation (ICA'07)*, London, UK, Sep. 2007.

[43] S. Winter, H. Sawada, S. Araki, and S. Makino, "Hierarchical clustering applied to overcomplete BSS for convolutive mixtures," in *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing (SAPA 2004)*, Oct. 2004.

[44] "Audacity : The free, cross-platform sound editor." [Online]. Available: http://audacity.sourceforge.net/

[45] J.-F. Cardoso and M. Martin, "A flexible component model for precision ICA," in *Proc. 7th International Conference on Independent Component Analysis and Signal Separation (ICA'07)*, London, UK, Sep. 2007, pp. 1–8.

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

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

[48] A. van den Bos, "Complex gradient and Hessian," *IEE Proceedings on Vision, Image and Signal Processing*, vol. 141, pp. 380–382, 1994.

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

Dépôt légal : 2009 – 1er trimestre
Imprimé à l'Ecole Nationale Supérieure des Télécommunications – Paris
ISSN 0751-1345 ENST D (Paris) (France 1983-9999)

TELECOM ParisTech

Institut TELECOM - membre de ParisTech

46, rue Barrault - 75634 Paris Cedex 13 - Tél. + 33 (0)1 45 81 77 77 - www.telecom-paristech.frfr

Département TSI