A VARIATIONAL EM ALGORITHM FOR THE SEPARATION OF MOVING SOUND SOURCES

Dionyssos Kounades-Bastian[1], Laurent Girin[1,2], Xavier Alameda-Pineda[3], Sharon Gannot[4], Radu Horaud[1]

[1] INRIA Grenoble Rhône-Alpes, France
[2] Univ. Grenoble Alpes, GIPSA-lab, France
[3] University of Trento, Dept. Information Ing. Comp. Sc., Italy
[4] Bar-Ilan University, Faculty of Engineering, Israel

ABSTRACT

This paper addresses the problem of separation of moving sound sources. We propose a probabilistic framework based on the complex Gaussian model combined with non-negative matrix factorization. The properties associated with moving sources are modeled using time-varying mixing filters described by a stochastic temporal process. We present a variational expectation-maximization (VEM) algorithm that employs a Kalman smoother to estimate the mixing filters. The sound sources are separated by means of Wiener filters, built from the estimators provided by the proposed VEM algorithm. Preliminary experiments with simulated data show that, while for static sources we obtain results comparable with the baseline method [1], in the case of moving source our method outperforms a piece-wise version of the baseline method.

Index Terms— Audio-source separation, time-varying mixing filters, moving sources, Kalman smoother, variational EM.

1. INTRODUCTION

Audio-source separation methods aim at recovering J unobserved source signals, $\mathbf{s} = [s_1, \ldots, s_J]^\top$ ($^\top$ denotes the transpose operator), from I observed mixed signals $\mathbf{x} = [x_1, \ldots, x_I]^\top$. A large body of literature deals with various source-separation configuration problems and their associated methods [2]. In this paper we consider the difficult case of convolutive mixtures of moving audio sources, i.e., the source-to-microphone channels are modeled with time-varying linear filters, thus taking into account possible motions of the sources and/or of the sensors (as may be the case in, e.g., human-robot interaction scenarios). Moreover the mixtures can be possibly underdetermined, i.e. we may have $I < J$.

To address this difficult problem, we focus on probabilistic methods based on complex-valued Gaussian models of source signals in the time-frequency (TF) domain, as initially proposed in [3] and further considered in, e.g., [1, 4, 5, 6, 7, 8, 9]. More specifically, we inspire from [1], where the complex Gaussian source model is combined with a non-negative matrix factorization (NMF) model [10, 11] for the source power spectral density (PSD) matrix. The model parameters (mixing filters and NMF coefficients) are estimated using the EM algorithm and the sources are separated by Wiener filters built from the estimated parameters. In [1] only time-invariant mixing filters are addressed. In this paper we propose a probabilistic model, and an associated estimation algorithm, based on the complex-Gaussian and NMF models and able to separate sound sources convolved with time-varying filters. Modeling convolutive mixtures with time-varying filters was already proposed in, e.g., [9, 12]. However, up to our knowledge, this is the first attempt to incorporate a latent-continuous model for the time-varying mixing filters in the TF-domain complex-Gaussian framework. Unlike [9], where the mixing system is parameterized with the angle of arrival, ruled by a discrete temporal model, our mixing model uses a more general propagation regime which is expected to be more suitable to reverberant environments. Moreover, [9] relies on binary masking for separating the sources, which is known to introduce speech distortion, whereas we use the more general Wiener filtering.

The paper is organized as follows. Section 2 describes the source model and introduces the proposed mixing model. In Section 3, we present a variational EM (VEM) algorithm for both the estimation of model parameters and inference of latent variables, in batch mode. A first series of experiments is reported in Section 4. Conclusions and future work are depicted in Section 5.

2. SOUND MIXTURES WITH TIME-VARYING FILTERS

2.1. The Source Model

Assuming that we work in the TF domain, as a result of applying the short-time Fourier transform (STFT) to the time-domain signals, the following notations are introduced: $f \in [1, F]$ denotes the frequency bin index, $\ell \in [1, L]$ denotes the time frame index, $\{\mathcal{K}_j\}_j$ denotes a non-trivial partition of $\{1, \ldots, K\}$, $K \geq J$ (K_j denotes the cardinal of $\mathcal{K}_j$). Following [1], each source $s_{j,f\ell}$ at TF bin (f, ℓ) is modeled as the sum of K_j latent components $c_{k,f\ell}$, $k \in \mathcal{K}_j$, namely:

$$s_{j,f\ell} = \sum_{k \in \mathcal{K}_j} c_{k,f\ell} \Leftrightarrow \mathbf{s}_{f\ell} = \mathbf{G}\mathbf{c}_{f\ell}, \tag{1}$$

where $\mathbf{G} \in \mathbb{N}^{J \times K}$ is a binary selection matrix with elements $\mathbf{G}_{jk} = 1$ if $k \in \mathcal{K}_j$ and $\mathbf{G}_{jk} = 0$ otherwise, $\mathbf{s}_{f\ell} = [s_{1,f\ell}, \ldots, s_{J,f\ell}]^\top$ and $\mathbf{c}_{f\ell} = [c_{1,f\ell}, \ldots, c_{K,f\ell}]^\top$. Each component $c_{k,f\ell}$ is assumed to follow a zero-mean proper complex Gaussian distribution [13] of variance $w_{fk}h_{k\ell}$, where $w_{fk}, h_{k\ell} \in \mathbb{R}^+$, i.e., $c_{k,f\ell} \sim \mathcal{N}_c(0, w_{fk}h_{k\ell})$. The components are assumed to be mutually independent and individually independent across frequency and time, so that we have:

$$s_{j,f\ell} \sim \mathcal{N}_c(0, \sum_{k \in \mathcal{K}_j} w_{fk}h_{k\ell}). \tag{2}$$

Support from EU-FP7 ERC Advanced Grant VHIA (#340113) and STREP EARS (#609645) is greatly acknowledged.

This corresponds to model the source PSD matrix with the NMF model [10, 11], i.e., $E[|\mathbf{s}_j|^2] = \{E[|s_{j,f\ell}|^2]\}_{f\ell} = \mathbf{W}_j\mathbf{H}_j$, with non-negative matrices $\mathbf{W}_j = \{w_{fk}\}_{f,k\in\mathcal{K}_j}$ of size $F \times K_j$ and $\mathbf{H}_j = \{h_{k\ell}\}_{k\in\mathcal{K}_j,\ell}$ of size $K_j \times L$. The columns of $\mathbf{W}_j$ are generally referred to as *spectral pattern vectors*, and the rows of $\mathbf{H}_j$ are referred to as *temporal activation vectors*. Such an NMF model is widely used in audio analysis, source separation, or speech enhancement, e.g., [3, 14, 15, 16], since it appropriately models a large range of sounds by providing harmonic and non-harmonic patterns. Additionally, it alleviates the well-known source permutation problem across frequencies met in many TF-domain source separation algorithms [1].

2.2. The Mixing Model

In numerous source separation methods, including [1], the multichannel mixed signal is modeled as a convolutive noisy mixture of the source signals. Relying on the so-called narrow-band assumption (the filters are shorter than the TF analysis window), the mixed signal $\mathbf{x}_{f\ell} = [x_{1,f\ell}, \dots, x_{I,f\ell}]^\top$ in the TF domain writes [17, 18]: $\mathbf{x}_{f\ell} = \mathbf{A}_f\mathbf{s}_{f\ell} + \mathbf{b}_{f\ell}$, where $\mathbf{b}_{f\ell} = [b_{1,f\ell}, \dots, b_{I,f\ell}]^\top$ is a zero-mean complex-Gaussian residual noise with $\mathbf{b}_{f\ell} \sim \mathcal{N}_c(\mathbf{0}, \mathrm{v}_f\mathbf{I}_I)$ and $\mathbf{A}_f = [\mathbf{a}_{1,f}, \dots, \mathbf{a}_{J,f}]$ is the frequency-dependent mixing matrix of size $I \times J$ ($\mathbf{a}_{j,f}$ is the mixing vector for source j).

Traditionally, the mixing matrix depends on the frequency f but not on the time frame ℓ, meaning that the filters are assumed to be time-invariant. Here we propose to extend this framework to time-varying filters and hence the mixture model becomes:

$$\mathbf{x}_{f\ell} = \mathbf{A}_{f\ell}\mathbf{s}_{f\ell} + \mathbf{b}_{f\ell}, \tag{3}$$

with $\mathbf{A}_{f\ell}$ being both frequency and frame (time) dependent.

Importantly, the straightforward extension of [1] to time-varying filters is unfeasible. Indeed, instead of estimating the JFI complex parameters of all $\mathbf{A}_f$, one would need to estimate $LJFI$ complex parameters of all $\mathbf{A}_{f\ell}$ (with only LFI observations). In order to circumvent this issue, we propose to model the mixing matrices as latent variables, and parameterize their temporal relationship instead (with far less parameters). More precisely, we model the temporal evolution as a random walk with evolution (covariance) matrix $\mathbf{\Sigma}^a_f \in \mathbb{C}^{IJ\times IJ}$ and prior mean $\boldsymbol{\mu}^a_f \in \mathbb{C}^{IJ}$:

$$\mathbf{a}_{:,f\ell}|\mathbf{a}_{:,f\ell-1} \sim \mathcal{N}_c(\mathbf{a}_{:,f\ell-1}, \mathbf{\Sigma}^a_f), \tag{4}$$

$$\mathbf{a}_{:,f\ell=1} \sim \mathcal{N}_c(\boldsymbol{\mu}^a_f, \mathbf{\Sigma}^a_f), \tag{5}$$

where $\mathbf{a}_{:,f\ell} = [\mathbf{a}^\top_{1,f\ell}, \dots, \mathbf{a}^\top_{J,f\ell}]^\top$ is the column-wise vectorization of $\mathbf{A}_{f\ell}$. The graphical model of the proposed probabilistic model for audio source separation of time-varying convolutive mixtures is represented in Fig. 1.

3. VEM FOR SOURCE SEPARATION

3.1. Principle

Expectation-maximisation is a standard procedure to find maximum likelihood (ML) estimates in hidden variable problems. Alternating between evaluating the posterior distribution of the hidden variables (E-step) and maximizing the expected complete-data log-likelihood (M-step), EM provides (locally) optimal parameters in the ML sense for a given set of observations. In this work the set of hidden variables $\mathcal{H} = \{\mathbf{a}_{:,f\ell}, \mathbf{c}_{f\ell}\}^{F,L}_{f,\ell=1}$ consists of the mixing filters and the source components. Hence the mixing filters

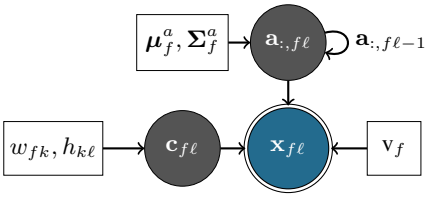

Figure 1: Graphical model for time-varying convolutive mixtures with NMF source model.

are modeled as stochastic processes instead of being defined by deterministic parameters as proposed in [19]. The set of parameters $\theta = \{\boldsymbol{\mu}^a_f, \mathbf{\Sigma}^a_f, w_{fk}, h_{k\ell}, \mathrm{v}_f\}^{F,L,K}_{f,\ell,k=1}$ consists of the parameters modeling the evolution of the mixing filters, the NMF coefficients and the sensor noise variance.

Given that the posterior distribution of the hidden variables, $q(\mathcal{H}) = p(\mathcal{H}|\{\mathbf{x}_{f\ell}\}^{F,L}_{f,\ell=1}; \theta)$, cannot be expressed in closed-form, we opt for variational inference, where $q(\mathcal{H})$, in our case, is assumed to factorise as:

$$q(\mathcal{H}) \approx \prod_{f=1}^{F} q(\mathbf{a}_{:,f1:L}) \prod_{f,\ell=1}^{F,L} q(\mathbf{c}_{f\ell}). \tag{6}$$

It is known [20], ch.10, that given a factorisation of $q(\mathcal{H})$, over a partition of the latent variables, the optimal posterior distribution of a subset $\mathcal{H}_0 \subseteq \mathcal{H}$ can be computed with:

$$q(\mathcal{H}_0) \propto \exp \mathbb{E}_{q(\mathcal{H}/\mathcal{H}_0)}\big[\log p(\mathcal{H}, \{\mathbf{x}_{f\ell}\}^{F,L}_{f,\ell=1}; \theta)\big], \tag{7}$$

with $\mathcal{H}/\mathcal{H}_0$ denoting $\mathcal{H}$ deprived of $\mathcal{H}_0$. Subsequently, $q(\mathcal{H})$ can be inferred in an alternating manner for each $\mathcal{H}_0 \in \mathcal{H}$. Below we present the proposed VEM algorithm that alternates between inference of $\mathbf{a}_{f\ell}, \mathbf{c}_{f\ell}$ and update of θ. A detailed derivation of the algorithm is beyond the scope of this short paper. It is important to note that the proposed algorithm is designed to work in batch mode, i.e., iterations are applied "globally" on a complete sequence of observed data (L frames). An online processing scheme is left for future work.

3.2. E-A Step

Using (7) the a posteriori distribution of $\mathbf{a}_{:,f\ell}$ writes:

$$q(\mathbf{a}_{:,f1:L}) \propto p(\mathbf{a}_{:,f1:L}) \prod_{\ell=1}^{L} \exp \mathbb{E}_{q(\mathbf{c}_{f\ell})}\big[\log p(\mathbf{x}_{f\ell}|\mathbf{a}_{:,f\ell}, \mathbf{c}_{f\ell})\big]. \tag{8}$$

The exponential term reduces to a complex-Gaussian distribution: $\exp \mathbb{E}_{q(\mathbf{c}_{f\ell})}\big[\log p(\mathbf{x}_{f\ell}|\mathbf{a}_{:,f\ell}, \mathbf{c}_{f\ell})\big] \propto \mathcal{N}_c(\boldsymbol{\mu}^{\iota a}_{f\ell}; \mathbf{a}_{:,f\ell}, \mathbf{\Sigma}^{\iota a}_{f\ell})$, with $\boldsymbol{\mu}^{\iota a}_{f\ell} \in \mathbb{C}^{IJ}$ and $\mathbf{\Sigma}^{\iota a}_{f\ell} \in \mathbb{C}^{IJ\times IJ}$ defined as

$$\boldsymbol{\mu}^{\iota a}_{f\ell} = \mathrm{vec}\Big(\mathbf{x}_{f\ell}\hat{\mathbf{s}}^{\mathrm{H}}_{f\ell}(\mathbf{Q}^{\eta s}_{f\ell})^{-1}\Big),\ \mathbf{\Sigma}^{\iota a}_{f\ell} = \big(\mathbf{Q}^{\eta s}_{f\ell}\big)^{-\top} \otimes \big(\mathrm{v}_f\mathbf{I}_I\big). \tag{9}$$

In the above equations, $\otimes$ denotes the Kronecker matrix product, vec(.) is the column-wise vectorization operator, and $\hat{\mathbf{s}}_{f\ell} \in \mathbb{C}^J$ and $\mathbf{Q}^{\eta s}_{f\ell} \in \mathbb{C}^{J\times J}$ are the posterior mean vector and posterior PSD matrix of $\mathbf{s}_{f\ell}$:

$$\hat{\mathbf{s}}_{f\ell} = \mathbf{G}\mathbb{E}_{q(\mathbf{c}_{f\ell})}[\mathbf{c}_{f\ell}] = \mathbf{G}\boldsymbol{\mu}^{\eta c}_{f\ell}, \tag{10}$$

$$\mathbf{Q}^{\eta s}_{f\ell} = \mathbf{G}\mathbb{E}_{q(\mathbf{c}_{f\ell})}\big[\mathbf{c}_{f\ell}\mathbf{c}^{\mathrm{H}}_{f\ell}\big]\mathbf{G}^\top = \mathbf{G}\mathbf{\Sigma}^{\eta c}_{f\ell}\mathbf{G}^\top + \hat{\mathbf{s}}_{f\ell}\hat{\mathbf{s}}^{\mathrm{H}}_{f\ell}, \tag{11}$$

which in turn are expressed in terms of $\boldsymbol{\mu}_{f\ell}^{\eta c} \in \mathbb{C}^K$ and of $\boldsymbol{\Sigma}_{f\ell}^{\eta c} \in \mathbb{C}^{K\times K}$, the a posteriori statistics of $\mathbf{c}_{f\ell}$ (see Section 3.3).

Since both terms in (8) are Gaussian, the a posteriori distribution is a linear dynamical system (LDS) along the frames ℓ. Therefore, the marginal posterior distribution for each frame is computed using the Kalman smoother recursions (see [20], Ch. 13). For the sake of clarity, we will denote the marginal and pair-wise joint (two succesive frames) a posteriori probability distributions by:

$$q\big(\mathbf{a}_{:,f\ell}\big) = \mathcal{N}_c\left(\mathbf{a}_{:,f\ell}; \boldsymbol{\mu}_{f\ell}^{\eta a}, \boldsymbol{\Sigma}_{f\ell}^{\eta a}\right) \text{ and} \tag{12}$$

$$q\big(\mathbf{a}_{:,f\ell}, \mathbf{a}_{:,f\ell-1}\big) = \mathcal{N}_c\left(\left[\mathbf{a}_{:,f\ell}^\top, \mathbf{a}_{:,f\ell-1}^\top\right]^\top; \boldsymbol{\mu}_{f\ell}^{\xi a}, \boldsymbol{\Sigma}_{f\ell}^{\xi a}\right), \tag{13}$$

where $\boldsymbol{\mu}_{f\ell}^{\eta a} \in \mathbb{C}^{IJ}, \boldsymbol{\Sigma}_{f\ell}^{\eta a} \in \mathbb{C}^{IJ\times IJ}$ and $\boldsymbol{\mu}_{f\ell}^{\xi a} \in \mathbb{C}^{2IJ}, \boldsymbol{\Sigma}_{f\ell}^{\xi a} \in \mathbb{C}^{2IJ\times 2IJ}$ are the mean and the covariance of the marginal and of the pair-wise joint a posteriori distributions respectively.

3.3. E-C Step

Using (7) the posterior distribution of the source components writes:

$$q(\mathbf{c}_{f\ell}) \propto \prod_{k=1}^{K} p(c_{k,f\ell}) \exp \mathbb{E}_{q(\mathbf{a}_{:,f\ell})}\big[\log p(\mathbf{x}_{f\ell}|\mathbf{a}_{:,f\ell}, \mathbf{c}_{f\ell})\big], \tag{14}$$

which can be shown to be a multivariate complex-Gaussian:

$$q(\mathbf{c}_{f\ell}) = \mathcal{N}_c(\mathbf{c}_{f\ell}; \boldsymbol{\mu}_{f\ell}^{\eta c}, \boldsymbol{\Sigma}_{f\ell}^{\eta c}), \tag{15}$$

with posterior covariance $\boldsymbol{\Sigma}_{f\ell}^{\eta c} \in \mathbb{C}^{K\times K}$ and mean $\boldsymbol{\mu}_{f\ell}^{\eta c} \in \mathbb{C}^K$:

$$\boldsymbol{\Sigma}_{f\ell}^{\eta c} = \left[\mathrm{diag}_K(w_{fk}h_{k\ell})^{-1} + \frac{1}{\mathrm{v}_f}\mathbf{G}^\top \mathbf{U}_{f\ell}^\top \mathbf{G}\right]^{-1}, \tag{16}$$

$$\boldsymbol{\mu}_{f\ell}^{\eta c} = \boldsymbol{\Sigma}_{f\ell}^{\eta c}\mathbf{G}^\top \left(\boldsymbol{\mu}_{f\ell}^{\eta A}\right)^{\mathrm{H}} \frac{\mathbf{x}_{f\ell}}{\mathrm{v}_f}. \tag{17}$$

In the above equations, $\mathrm{diag}_K(d_k)$ denotes the $K \times K$ diagonal matrix with entries from vector $[d_k]_{k=1}^K$, $\mathbf{U}_{f\ell} \in \mathbb{C}^{J\times J}$ is a matrix whose jr^{th} entry is $[\mathbf{U}_{f\ell}]_{jr} = \mathrm{tr}\{\mathbf{Q}_{jr,f\ell}^{\eta a}\}$, where $\mathbf{Q}_{jr,f\ell}^{\eta a}$ is the jr^{th} $I \times I$ block of the filters posterior second order moment: $\mathbf{Q}_{f\ell}^{\eta a} = \mathbb{E}_{q(\mathbf{a}_{:,f\ell})}\left[\mathbf{a}_{:,f\ell}\mathbf{a}_{:,f\ell}^{\mathrm{H}}\right] = \boldsymbol{\Sigma}_{f\ell}^{\eta a} + \boldsymbol{\mu}_{f\ell}^{\eta a}(\boldsymbol{\mu}_{f\ell}^{\eta a})^{\mathrm{H}}$ and $\boldsymbol{\mu}_{f\ell}^{\eta A} \in \mathbb{C}^{I\times J}$ is the matrixification (inverse of column-wise vectorisation) of $\boldsymbol{\mu}_{f\ell}^{\eta a}$.

3.4. M Step

After computing the a posteriori statistics for $\mathbf{a}$ and $\mathbf{c}$, the expected complete-data log-likelihood, $\mathbb{E}_{q(\mathcal{H})} \log p\big(\mathcal{H}, \{\mathbf{x}_{f\ell}\}_{f,\ell=1}^{F,L}; \theta\big)$, is maximised with respect to the parameters. We obtain closed-form expressions for v_f, $\boldsymbol{\mu}_f^a$, and $\boldsymbol{\Sigma}_f^a$, namely:

$$\mathrm{v}_f = \frac{1}{LI}\sum_{\ell=1}^{L}\left(\mathbf{x}_{f\ell}^{\mathrm{H}}\mathbf{x}_{f\ell} + \mathrm{tr}\left\{\mathbf{U}_{f\ell}^\top \mathbf{Q}_{f\ell}^{\eta s}\right\} - 2\mathfrak{Re}\left\{\mathbf{x}_{f\ell}^{\mathrm{H}}\boldsymbol{\mu}_{f\ell}^{\eta A}\hat{\mathbf{s}}_{f\ell}\right\}\right),$$

$$\boldsymbol{\mu}_f^a = \boldsymbol{\mu}_{f\ell=1}^{\eta a},$$

$$\boldsymbol{\Sigma}_f^a = \frac{1}{L}\left(\mathbf{Q}_{11,f}^{\xi a} - \mathbf{Q}_{21,f}^{\xi a} - \mathbf{Q}_{12,f}^{\xi a} + \mathbf{Q}_{22,f}^{\xi a} + \boldsymbol{\Sigma}_{f\ell=1}^{\eta a}\right), \tag{18}$$

where $\mathbf{Q}_f^{\xi a}$ is defined as $\mathbf{Q}_f^{\xi a} = \sum_{\ell=1}^{L-1}\left(\boldsymbol{\Sigma}_{f\ell}^{\xi a} + \boldsymbol{\mu}_{f\ell}^{\xi a}\boldsymbol{\mu}_{f\ell}^{\xi a\mathrm{H}}\right)$.

Regarding the NMF parameters, their optimisation can be done independently for each component. However, the joint optimization of w_{fk} and $h_{k\ell}$ is a non-convex problem without exact solution. Alternate maximisation is a classical solution to solve for the NMF parameters, since the updates are in closed-form, for instance:

$$w_{fk} \leftarrow \frac{1}{L}\sum_{\ell=1}^{L}\frac{\mathbf{Q}_{kk,f\ell}^{\eta c}}{h_{k\ell}}, \quad h_{k\ell} \leftarrow \frac{1}{F}\sum_{f=1}^{F}\frac{\mathbf{Q}_{kk,f\ell}^{\eta c}}{w_{fk}}, \tag{19}$$

where $\mathbf{Q}_{f\ell}^{\eta c} = \boldsymbol{\Sigma}_{f\ell}^{\eta c} + \boldsymbol{\mu}_{f\ell}^{\eta c}(\boldsymbol{\mu}_{f\ell}^{\eta c})^{\mathrm{H}}$ is the a posteriori second order moment of the sources' components.

Algorithm 1 Separation of J moving sound sources

input $\{\mathbf{x}_{f\ell}\}_{f,\ell=1}^{F,L}$, binary matrix $\mathbf{G}$ and initial parameters θ.
Initialise posterior statistics $\boldsymbol{\Sigma}_{f\ell}^{\eta A}, \boldsymbol{\mu}_{f\ell}^{\eta A}$.
repeat
  E-C step: Compute $\boldsymbol{\Sigma}_{f\ell}^{\eta c}$ using (16) and $\boldsymbol{\mu}_{f\ell}^{\eta c}$ using (17), estimate source STFTs $\hat{\mathbf{s}}_{f\ell}$ from (10).
  E-A step: Compute $\boldsymbol{\Sigma}_{f\ell}^{\iota a}$ and $\boldsymbol{\mu}_{f\ell}^{\iota a}$ with (9) and compute $\{\boldsymbol{\Sigma}_{f\ell}^{\eta a}, \boldsymbol{\mu}_{f\ell}^{\eta a}, \boldsymbol{\Sigma}_{f\ell}^{\xi a}, \boldsymbol{\mu}_{f\ell}^{\xi a}\}$ using the Kalman recursions.
  M step: Update the set of parameters θ with (18) and (19).
until Convergence
return The estimated source images $\boldsymbol{\mu}_{j,f\ell}^{\eta A}\hat{\mathbf{s}}_{j,f\ell}, j \in [1, J]$.

3.5. Estimation of Source Images

As is usual in source separation problems, the present framework suffers from the well-known scaling indeterminacy, i.e. source signals and filters can be estimated only up to some arbitrary compensating multiplicative factors [2]. In [1], this problem is addressed with a normalization of the (stationary) estimated mixing filters. In the present study, we have tested several normalization procedures which were not found to improve the separation performances. Therefore, we rather fix the scale indeterminacy problem by providing the estimates of the source images as outputs, i.e. the estimates of the source signals as recorded at the microphones [6, 21]. These are given by: $\mathbb{E}_{q(\mathcal{H})}[\mathbf{a}_{j,f\ell}\mathbf{s}_{j,f\ell}] = \boldsymbol{\mu}_{j,f\ell}^{\eta A}\hat{\mathbf{s}}_{j,f\ell}$ ($\boldsymbol{\mu}_{j,f\ell}^{\eta A}$ is the j-th column of $\boldsymbol{\mu}_{f\ell}^{\eta A}$; time-domain output signals are obtained by inverse STFT). The complete variational EM separating J moving sound sources is depicted in Algorithm 1.

4. EXPERIMENTS

4.1. Simulation Setup

We conducted a series of simulations to assess the performance of the proposed algorithm, and compare it with [1]. Three 16-kHz speech sources were selected from the TIMIT database [22] and shortened to 2s. These speech signals were then convolved with 200-tap head-related impulse responses (HRIR) from the CIPIC database [23] and then summed to provide the mix signal. Therefore, all reported experiments were made with 2×3 underdetermined mixtures, with all source images having roughly the same power. The STFT was applied on the mix signal with a 512-sample sine window, leading to $L = 128$ observation frames. The number of components per source K_j was set to 20. The number of VEM iterations was fixed to 1,000. Note that, for the reported experiments, our unoptimized Matlab implementation requires 30s per iteration on a 2.4GHz 4-core PC (the baseline methods requires about 1s).

We report results for three mixture configurations. In the Mix 1 configuration, all sources are static, at respectively -30°, 0° and

30° azimuth. This configuration is dedicated to test if the proposed algorithm behaves well when the filters are time-invariant. In the Mix 2 configuration, Source 1 is moving from -15° to 15°, Source 2 is moving from 15° to -15°, and Source 3 is static at -80°. Hence, Source 1 and 2 are crossing each other. In Mix 3a and 3b, all sources are moving, Source 1 from 0° to 60°, Source 2 from -60° to 0°, and Source 3 from -15° to 15°. Therefore, sources are not spatially overlapping but Sources 1 and 2 are moving twice as fast as in Mix 2. Mix 3a and 3b have the same mixing filters but different source content.

Standard objective measures of audio source separation were calculated from estimated and ground truth source images, namely: signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR) and signal-to-artifact ratio (SAR) [24].

4.2. Initialisation

Initialization is known to be a crucial step for the performance of EM algorithms. For this reason, in this study we considered using ground truth (GT) parameters, as already considered in some experiments of [6]. This is because we first aim here at testing the behavior of the proposed algorithm within a favorable setting. Evaluating the performance of the algorithm in more realistic conditions, e.g., using corrupted GT parameters or using the output of another source separation technique (since an EM algorithm is essentially an optimization algorithm) is a substantial task that is out of the scope of the present paper. This deserves a highly detailed investigation that will be considered in future works. Therefore, GT mixing filters, $\boldsymbol{\mu}_{f\ell}^{\eta A}$, were calculated as the FFT of the mixing impulse responses. GT NMF parameters, $w_{fk}, h_{k\ell}$, were calculated by applying the KL-NMF algorithm [15] to the original (unmixed) sources' PSD. Finally, v_f is initialized as 1% of the time-average mixture's PSD, $\boldsymbol{\Sigma}_{f\ell}^{\eta a} = 10^{-8}\mathbf{I}$ and $\boldsymbol{\Sigma}_f^a = \mathbf{I}$.

In the dynamic case, we split the sequence of $L = 128$ time frames into $P = 8$ segments of $L/P = 16$ frames. The baseline method [1] was run independently on each segment (it cannot be run independently for each STFT frame since it requires statistics evaluated from several frames). This way, we obtain a piece-wise stationary version of the baseline method that is "adapted" to the time-varying mixing case. The mixing filter corresponding to the center of each segment was selected for the initialization of this segment. For the proposed method, the initial central filter of each segment was replicated at each frame of the segment so as to initialize the complete sequence of L mixing filters. This way both algorithms were initialized with the same amount of information.

4.3. Results

The results of the simulations are presented in Table 1. In the static case (Mix 1), both methods provide similar performance, ensuring very good signal separation. A closer look to the results reveals slightly favorable performance for the baseline method. This behavior is expected, since the baseline method was designed for time-invariant mixing filters. This may also happen because the variational approximation does not solve for the exact model.

In the dynamic case (Mix 2, 3a and 3b), the results obtained by the proposed method show very good separation performance. This is not surprising because ground truth initialization was used. Even still, this demonstrates that the proposed algorithm is able to correctly estimate the mixing filters trajectories (remind that a 128-frame sequence of filters is estimated from only 8 initial frames),

Table 1: Separation results for the proposed method (Prop.) and piece-wise adaptation of the baseline method [1]. Measures are given in dB.

Measure	Source	Method	Mix 1	Mix 2	Mix 3a	Mix 3b
SDR	1	Prop.	16.2	9.9	8.3	8.7
		[1]	17.1	6.0	5.5	8.0
	2	Prop.	16.5	7.2	9.7	10.2
		[1]	17.1	3.5	3.8	4.2
	3	Prop.	26.2	9.5	14.0	12.9
		[1]	28.0	6.3	9.9	10.2
SIR	1	Prop.	23.0	14.0	15.3	14.5
		[1]	23.2	10.3	11.1	14.0
	2	Prop.	21.7	16.2	17.7	17.2
		[1]	22.1	8.6	8.6	6.6
	3	Prop.	31.8	15.2	20.3	20.4
		[1]	33.9	11.6	18.7	19.2
SAR	1	Prop.	17.4	14.5	10.2	11.3
		[1]	18.0	9.1	7.5	10.6
	2	Prop.	17.7	11.3	12.1	12.6
		[1]	18.0	7.1	6.3	7.8
	3	Prop.	28.0	11.4	16.3	16.3
		[1]	30.0	7.9	11.6	13.1

validating the VEM approach. Importantly, in all cases, the proposed method significantly outperforms the piece-wise version of the baseline method. Indeed, the improvement on SDR ranges from 0.7 to 6 dB, the improvement on SIR ranges from 1.2 to 10.6 dB and the improvement on SAR ranges from 0.7 to 5.8 dB. Therefore, the proposed method provides better interference rejection, artifact limitation and overall signal reconstruction capabilities than the baseline method.

Examples of original, mixed and separated signals (corresponding to Mix 3a) are provided as supplementary material at www.gipsa-lab.grenoble-inp.fr/∼laurent.girin/demo/W2015.zip.

5. CONCLUSION AND FUTURE WORK

In this paper we have presented a probabilistic model for sound source separation from (possibly underdetermined) convolutive mixtures with moving sources. This model extends [1] to the challenging case of time-varying mixing filters. We have derived a variational EM for parameters estimation and source (images) extraction. The proposed method has been shown to compete favorably with a piece-wise stationary version of the reference method, on mixtures of three speech signals convolved with time-varying HRIRs. This shows the efficiency of the filter estimation procedure within the VEM, i.e. the Kalman smoother. This work is a proof of concept of considering time-varying filters as latent variables into the complex-Gaussian / NMF framework. Future works will go towards more realistic implementations, considering i) (long) room impulse responses for modeling reverberations, ii) online processing, i.e. causal estimation/update of filters and NMF parameters, using, e.g., Kalman filtering instead of smoothing, iii) normalization, convergence and computational issues, iv) a refined temporal model for filter evolution, v) other models for source PSD, e.g., Gamma priors, and finally vi) more realistic initialization procedures.

6. REFERENCES

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

[2] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation - Independent Component Analysis and Applications*. Academic Press, 2010.

[3] L. Benaroya, L. Donagh, F. Bimbot, and R. Gribonval, "Non negative sparse representation for wiener based source separation with a single sensor," in *IEEE Int. Conf. Acoustics, Speech, and Signal Proc.*, vol. 6, 2003, pp. 613–616.

[4] C. Févotte and J.-F. Cardoso, "Maximum likelihood approach for blind audio source separation using time-frequency gaussian source models," in *IEEE Workshop Applicat. Signal Process. to Audio and Acoust. (WASPAA)*, 2005.

[5] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 14, no. 1, pp. 191–199, 2006.

[6] N. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 18, no. 7, pp. 1830–1840, 2010.

[7] A. Liutkus, B. Badeau, and G. Richard, "Gaussian processes for underdetermined source separation," *IEEE Trans. on Signal Proc.*, vol. 59, no. 7, pp. 3155–3167, 2011.

[8] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 20, no. 4, pp. 1118–1133, 2012.

[9] T. Higuchi, N. Takamune, T. Nakamura, and H. Kameoka, "Underdetermined blind separation and tracking of moving sources based on DOA-HMM," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 3215–3219.

[10] D. Lee and H. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[11] ——, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, vol. 13, pp. 556 – 562, 2001.

[12] S. Markovich Golan, S. Gannot, and I. Cohen, "Subspace tracking of multiple sources and its application to speakers extraction," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Dallas, Texas, USA, 2010, p. 201204.

[13] F. Neeser and J. Massey, "Proper complex random processes with applications to information theory," *IEEE Trans. on Information Theory*, vol. 39, no. 4, p. 12931302, 1993.

[14] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 15, no. 3, pp. 1066–1074, 2007.

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

[16] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 21, no. 10, pp. 2140–2151, 2013.

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

[18] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[19] S. Gannot and M. Moonen, "On the application of the unscented Kalman filter to speech processing," in *IEEE Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Kyoto, Japan, 2003, p. 811.

[20] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[21] N. Sturmel, A. Liutkus, J. Pinel, L. Girin, S. Marchand, G. Richard, R. Badeau, and L. Daudet, "Linear mixing models for active listening of music productions in realistic studio conditions," in *Convention of the Audio Engineering Society (AES)*, Budapest, Hungary, 2012.

[22] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," 1993, linguistic Data Consortium, Philadelphia.

[23] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2001.

[24] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 14, no. 4, pp. 1462–1469, 2006.