A Variational EM Algorithm for the Separation of Time-Varying Convolutive Audio Mixtures

Dionyssos Kounades-Bastian, Laurent Girin, Xavier Alameda-Pineda, *Member, IEEE*, Sharon Gannot, *Senior Member, IEEE*, and Radu Horaud

***Abstract*—This paper addresses the problem of separating audio sources from time-varying convolutive mixtures. We propose a probabilistic framework based on the local complex-Gaussian model combined with non-negative matrix factorization. The time-varying mixing filters are modeled by a continuous temporal stochastic process. We present a variational expectation–maximization (VEM) algorithm that employs a Kalman smoother to estimate the time-varying mixing matrix, and that jointly estimate the source parameters. The sound sources are then separated by Wiener filters constructed with the estimators provided by the VEM algorithm. Extensive experiments on simulated data show that the proposed method outperforms a blockwise version of a state-of-the-art baseline method.**

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

I. Introduction

SOURCE separation aims at recovering unobserved source signals from observed mixtures [1]. Audio source separation (ASS) is mainly concerned with mixtures of speech, music, ambient noise, etc. For acoustic signals in natural environments, the mixing process is generally considered as *convolutive*, i.e., the acoustic channel between each source and each microphone is modeled by a linear filter that represents the multiple source-to-microphone paths due to reverberations. Source separation is a major component of machine audition systems, since it is used as a preprocessing step for many higher-level processes such as speech recognition, human-computer or human-robot interaction.

The vast majority of works on ASS from convolutive mixtures deals with *time-invariant* mixing filters, which means that the position of sources and microphones is assumed to be fixed. In other words, the source-to-microphone acoustic paths are assumed to remain the same over the duration of the recordings. In this work we consider the more realistic case of *time-varying* convolutive mixtures corresponding to source-to-microphone channels that can change over time. This should be able to take into account possible source or microphone motions. For example, in many Human-robot interaction scenarios, there is a strong need to consider mixed speech signals emitted by *moving* speakers, and/or recorded by a *moving* robot, and perturbed by reverberations. More generally, changes in the environment such as door/window opening/closing or curtain pulling must also be accounted for. Note that in this paper, the mixtures under consideration can be *underdetermined*, i.e., there may be less microphones than sources, which is a difficult ASS problem in its own right [1].

A. Related Work

The ASS literature that deals with time-invariant mixing filters is much larger than the literature dealing with time-varying filters. Therefore, we briefly discuss the former before reviewing the latter. State-of-the-art time-invariant ASS methods generally start with a time-frequency (TF) decomposition of the temporal signals, e.g., by applying the short-time Fourier transform (STFT). In the TF domain, the time-invariant convolutive filters are converted to multiplicative coefficients independent at each frequency bin [2]. These methods can then be classified into three (non-exclusive) categories [3]. Firstly, separation methods based on independent component analysis (ICA) consist in estimating the demixing filters that maximize the independency of separated sources [1], [4]. Unfortunately, ICA-based methods are subject to the well-known scale ambiguity and source permutation problems across frequency bins. In addition, these methods cannot be applied to underdetermined mixtures. Secondly, methods based on sparse component analysis and binary masking rely on the assumption that only one source is active at each TF point [5], [6]. Thirdly, more recent methods are based on complex-valued local Gaussian models (LGMs) for the sources [7], and the model proposed here is a member of this family of methods.

The LGM was initially proposed for single-microphone speech enhancement [8], then extended to single-channel ASS [9], [10] and multi-channel ASS [11]–[14]. The method proposed in [12] provides a rigorous framework for ASS from underdetermined convolutive mixtures: An LGM source model is combined with a nonnegative matrix factorization (NMF) model [15], [16] applied to the source PSD matrix [17], which

Manuscript received October 13, 2015; revised March 10, 2016 and April 11, 2016; accepted April 11, 2016. Date of publication April 14, 2016; date of current version May 27, 2016. The work of D. Kounades-Bastian, L. Girin, and R. Horaud was supported by the European FP7 STREP project EARS 609465 and by the European Research Council through the ERC Advanced Grant VHIA 340113. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yunxin Zhao.

D. Kounades-Bastian and R. Horaud are with the INRIA Grenoble Rhône-Alpes, Montbonnot-Saint-Martin 38334, France (e-mail: dionyssos.kounades-bastian@inria.fr; radu.horaud@inria.fr).

L. Girin is with the INRIA Grenoble Rhône-Alpes, Montbonnot-Saint-Martin 38334, France, and also with GIPSA-lab, Université Grenoble Alpes, Grenoble 38402, France (e-mail: laurent.girin@gipsa-lab.grenoble-inp.fr).

X. Alameda-Pineda is with the University of Trento, Trento 38122, Italy (e-mail: xavier.alamedapineda@unitn.it).

S. Gannot is with the Faculty of Engineering, Bar Ilan University, Ramat Gan 5290002, Israel (e-mail: Sharon.Gannot@biu.ac.il).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org.

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

Digital Object Identifier 10.1109/TASLP.2016.2554286

2329-9290 © 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

is reminiscent of pioneering works such as [9]. This allows one to drastically reduce the number of model parameters and to alleviate the source permutation problem. However, in [12] the mixing filters do not vary over time: they are considered as model parameters and, together with the NMF coefficients, they are estimated via an expectation–maximization (EM) algorithm. Then, the sound sources are separated with Wiener filters constructed from the learned parameters. A similar LGM-based approach is adopted in [18], though the speech signal PSD is here modeled as a time-varying auto-regressive (AR) model. Here also, all model parameters are estimated by maximizing the likelihood of the observed signals and solved by EM iterations.

In comparison to the time-invariant methods that we just mentioned, the literature dealing with time-varying acoustic mixtures is scarce. Early attempts addressing the separation of time-varying mixtures basically consisted in block-wise adaptations of time-invariant methods: An STFT frame sequence is split into blocks, and a time-invariant ASS algorithm is applied to each block. Hence, block-wise adaptations assume time-invariant filters within blocks. The separation parameters are updated from one block to the next and the separation result over a block can be used to initialize the separation of the next block. Frame-wise algorithms can be considered as particular cases of block-wise algorithms, with single-frame blocks, and hybrid methods may combine block-wise and frame-wise processing. Notice that, depending on the implementation, some of these methods may run online.

Interestingly, most of the block-wise approaches use ICA, either in the temporal domain [19] (limited to anechoic setups), [20]–[23] or in the Fourier domain [24], [25] (limited to instantaneous mixtures), [26]. In addition to being limited to overdetermined mixtures, block-wise ICA methods need to account for the source permutation problem, not only across frequency bins, as usual, but across successive blocks as well. Examples of block-wise adaptation of binary-masking or LGM-based methods are more scarce. As for binary masking, a block-wise adaptation of [27] is proposed in [28]. This method performs source separation by clustering the observation vectors in the source image space. As for LGM, [29] describes an online block- and frame-wise adaptation of the general LGM framework proposed in [14]. One important problem, common to all block-wise approaches, is the difficulty to choose the block size. Indeed, the block size must assume a good trade-off between local channel stationarity (short blocks) and sufficient data to infer relevant statistics (long blocks). The latter constraint can drastically limit the dynamics of either the sources or the sensors [28]. Other parameters such as the step-size of the iterative update equations may also be difficult to set [29]. In general, systematic convergence towards a good separation solution using a limited amount of signal statistics remains an open issue.

Dynamic scenarios were also addressed differently in [30], where a beamforming method for extracting multiple moving sources is proposed. This method is applicable only to overdetermined mixture. Also, iterative and sequential approaches for speech enhancement in reverberant environment were proposed in [31]. The proposed methods utilize the EM framework to jointly estimate the desired speech signal and the required (deterministic) parameters, namely the speech AR coefficients, and the speech and noise mixing filters taps. For on-line implementation, a recursive version of the M-step was developed and the Kalman smoother, used in the batch mode, is substituted by the Kalman filter. However, only the case of a 2×2 mixture was addressed.

Separating underdetermined time-varying convolutive mixtures using binary masking within a probabilistic LGM framework was proposed in [32]. The mixing filters are considered as latent variables that follow a Gaussian distribution with mean vector depending on the direction of arrival (DOA) of the corresponding source. The DOA is modeled as a discrete latent variable taking values from a finite set of angles and following a discrete hidden Markov model (HMM). A variational expectation–maximization (VEM) algorithm is derived to perform the inference, including forward-backward equations to estimate the DOA sequence. This approach provides interesting results but it suffers from several limitations. First, the separation quality is poor, proper to binary masking approaches. Second, the accuracy is limited, which is inherent to the use of a discrete temporal model to represent a continuous variable, namely the source DOAs. Moreover, constraining the mixing filter to a DOA-dependent model can be problematic in highly reverberant environments. Finally, it must be noted that no specific source variance model is exploited, and that the filter and DOA models are assumed to solve the source permutation problem (both in frequency and time).

B. Contributions

In this paper we adopt the source LGM framework with an NMF PSD model. We consider the very general case of an underlying convolutive mixing process that is allowed to vary over time, and we model this process as a set of, *temporally-linked continuous latent variables*, using a prior model. We propose to parameterize the transfer function of the mixing filters with an unconstrained continuous linear dynamical system (LDS) [33]. We believe that this model can be more effective than the DOA-dependent HMM model of [32] in adverse and reverberant conditions, since the relationship between the transfer function and the source DOA can be quite complex. In addition, [32] relies on binary masking for separating the sources, which is known to introduce speech distortion, whereas we use the more general and more efficient Wiener filtering tied to LGM-based methods.

The proposed method may be viewed as a generalization of [12] to moving sources, moving microphones, or both. However, exact inference of the posterior distribution, as proposed in [12], turns out to be intractable in the more general model that we consider here. Therefore, we propose an approximate solution for the joint estimation of the model parameters and inference of the latent variables. We derive a VEM algorithm in which a *Kalman smoother* is used for the inference of the time-varying mixing filters. In comparison to the methodology described in [29], the proposed model goes beyond block- or frame-wise adaptation because it exploits the information available with

the whole sequence of input mixture frames. To summarize, the proposed method exploits all the available data to estimate the source parameters and mixing process parameters at each frame. As a consequence, it cannot be applied online. Note that an earlier reference to the incorporation of a latent Bayesian continuous model into the underlying filtering, with application to speech processing, can be found in [34]. Two schemes were proposed, namely a dual scheme with two Kalman filters applied sequentially in parallel, and a joint scheme using the approximated unscented Kalman filter. Only very simple filtering schemes were addressed. In the present paper, we provide a more rigorous treatment of the joint signal and parameter estimation problem, using the variational approach.

This paper is an extended version of [35]. A detailed description of the proposed model and of the associated VEM algorithm is now provided. Several mathematical derivations, that were omitted in [35], are now included in order to make the paper self-consistent, easy to understand, and to allow method reproducibility. Moreover, several computational simplifications are proposed, leading to a more efficient implementation. The method is tested over a larger set of signals and configurations, including experiments with blind initialization and real recordings, thus extending the very preliminary results presented in [35]. These results are compared with a block-wise implementation of the baseline method [12]. This may well be viewed as an adaptation of the general framework [29] to convolutive mixtures. Matlab code of the proposed algorithm together with speech test data are provided as supplementary material.[1,2]

The remaining of the paper is organized as follows. Section II describes the source, mixture and channel models. The associated VEM algorithm is described in Section III. Implementation details are discussed in Section IV. The experimental validation is reported in Section V. Conclusions and future works are discussed in Section VI.

II. Audio Mixtures with Time-Varying Filters

A. The Source Model

We work in a TF representation, after applying the STFT to the time-domain mixture signal. Let $f \in [1, F]$ denote the frequency bin index, and $\ell \in [1, L]$ denote the frame index. Consider a mixture of J source signals, with $\mathbf{s}_{f\ell} = [s_{1,f\ell} \dots s_{J,f\ell}]^\top \in \mathbb{C}^J$ denoting the latent vector of source coefficients at TF bin (f, ℓ) ($\mathbf{x}^\top$ and $\mathbf{x}^\mathrm{H}$ respectively denote $\mathbf{x}$ transpose and conjugate-transpose). Let $\{\mathcal{K}_j\}_{j=1}^J$ denote a non-trivial partition of $\{1 \dots K\}$, $K \geq J$ (in practice we may have $K \gg J$), that is known in advance. Following [12], a coefficient $s_{j,f\ell}$ is modeled as the sum of latent components $c_{k,f\ell}$, $k \in \mathcal{K}_j$:

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

where $\mathbf{G} \in \mathbb{N}^{J \times K}$ is a binary selection matrix with entries $G_{jk} = 1$ if $k \in \mathcal{K}_j$ and $G_{jk} = 0$ otherwise, and $\mathbf{c}_{f\ell} = [c_{1,f\ell}, \dots, c_{K,f\ell}]^\top \in \mathbb{C}^K$ is the vector of component coefficients at (f, ℓ). Each component $c_{k,f\ell}$ is assumed to follow a zero-mean proper complex Gaussian distribution with variance $w_{fk}h_{k\ell}$, where $w_{fk}, h_{k\ell} \in \mathbb{R}^+$. The components are assumed to be mutually independent and individually independent across frequency and time. Thus the component vector probability density function (pdf) writes:[3]

$$p(\mathbf{c}_{f\ell}) = \mathcal{N}_c\Big(\mathbf{c}_{f\ell}; \mathbf{0}, \mathrm{diag}_K\,(w_{fk}h_{k\ell})\Big), \tag{2}$$

where $\mathbf{0}$ denotes the zero-vector, $\mathrm{diag}_K(d_k)$ denotes the $K \times K$ diagonal matrix with entries $[d_1 \dots d_k \dots d_K]^\top$, and the source vector pdf writes:

$$p(\mathbf{s}_{f\ell}) = \mathcal{N}_c\left(\mathbf{s}_{f\ell}; \mathbf{0}, \mathrm{diag}_J\left(\sum_{k \in \mathcal{K}_j} w_{fk}h_{k\ell}\right)\right). \tag{3}$$

Eq. (3) corresponds to the modeling of the $F \times L$ source PSD matrix with the NMF model, which is widely used in audio analysis, ASS, and speech enhancement [9], [17], [37], [38]. NMF is empirically verified to adequately model a large range of sounds by providing harmonic as well as non-harmonic patterns activated over time. Note that both source and component vectors are treated as latent variables linked by (1).

B. The Mixture Model

In many source separation methods, including [12], the mixture signal is modeled as a time-invariant convolutive noisy mixture of the source signals. Let us denote the I-channel mixture signal in the TF domain by $\mathbf{x}_{f\ell} = [x_{1,f\ell} \dots x_{I,f\ell}]^\top \in \mathbb{C}^I$. Relying on the so-called narrow-band assumption (i.e. the impulse responses of the channel are shorter than the TF analysis window), $\mathbf{x}_{f\ell}$ writes [39], [40]: $\mathbf{x}_{f\ell} = \mathbf{A}_f\mathbf{s}_{f\ell} + \mathbf{b}_{f\ell}$, where $\mathbf{b}_{f\ell} = [b_{1,f\ell} \dots b_{I,f\ell}]^\top \in \mathbb{C}^I$ is a zero-mean complex-Gaussian residual noise, and $\mathbf{A}_f = [\mathbf{a}_{1,f} \dots \mathbf{a}_{J,f}] \in \mathbb{C}^{I \times J}$ is the mixing matrix (a column $\mathbf{a}_{j,f} \in \mathbb{C}^I$ is the mixing vector for source j). This way, the mixing matrix depends only on the frequency f but not on the time frame ℓ, meaning that the filters are assumed to be time-invariant. Since we are expressly interested in modeling time-varying filters, the mixing equation naturally becomes:

$$\mathbf{x}_{f\ell} = \mathbf{A}_{f\ell}\mathbf{s}_{f\ell} + \mathbf{b}_{f\ell}, \tag{4}$$

with $\mathbf{A}_{f\ell}$ being both frequency- and time-dependent. This equation allows us to cope with possible source/sensor movements and other environmental changes. Note that (4) accounts for temporal variations of the channel across frames, though it assumes that the channel is not varying within an individual frame, which is a reasonable assumption for a wide variety of applications. For simplicity $\mathbf{b}_{f\ell}$ is assumed here to be stationary and isotropic, i.e. $p(\mathbf{b}_{f\ell}) = \mathcal{N}_c(\mathbf{b}_{f\ell}; \mathbf{0}, \mathrm{v}_f\mathbf{I}_I)$, with $\mathrm{v}_f \in \mathbb{R}^+$ being a parameter to be estimated, and $\mathbf{I}_I$ denoting the identity matrix of size I. The conditional data distribution is thus given by $p(\mathbf{x}_{f\ell}|\mathbf{A}_{f\ell}, \mathbf{s}_{f\ell}) = \mathcal{N}_c(\mathbf{x}_{f\ell}; \mathbf{A}_{f\ell}\mathbf{s}_{f\ell}, \mathrm{v}_f\mathbf{I}_I)$.

[1] http://ieeexplore.ieee.org

[2] https://team.inria.fr/perception/research/vemove/

[3] The proper complex Gaussian distribution is defined as $\mathcal{N}_c(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = |\pi\boldsymbol{\Sigma}|^{-1} \exp\Big(-[\mathbf{x} - \boldsymbol{\mu}]^\mathrm{H}\boldsymbol{\Sigma}^{-1}[\mathbf{x} - \boldsymbol{\mu}]\Big)$, with $\mathbf{x}, \boldsymbol{\mu} \in \mathbb{C}^I$ and $\boldsymbol{\Sigma} \in \mathbb{C}^{I \times I}$ being the argument, mean vector, and covariance matrix respectively [36].

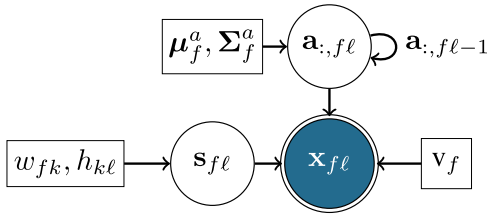

Fig. 1. Graphical model for time-varying convolutive mixtures with NMF source model. Latent variables are represented with circles, observations with double circles, deterministic parameters with rectangles, and temporal dependencies with self loops.

C. The Channel Model

A straightforward extension of [12] to time-varying linear filters is unfeasible. Indeed, instead of estimating the $I \times J \times F$ complex parameters of all $\mathbf{A}_f$, one would have to estimate the $I \times J \times F \times L$ complex parameters of all $\mathbf{A}_{f\ell}$ (with only $I \times F \times L$ observations). In order to circumvent this issue, we model the mixing matrix $\mathbf{A}_{f\ell}$ as a latent variable and parameterize its temporal evolution, with much less parameters.

For this purpose, we first vectorize $\mathbf{A}_{f\ell}$ by vertically concatenating its J columns $\{\mathbf{a}_{j,f\ell}\}_{j=1}^J$ into a single vector $\mathbf{a}_{:,f\ell} \in \mathbb{C}^{IJ}$, i.e. $\mathbf{a}_{:,f\ell} = \text{vec}(\mathbf{A}_{f\ell}) = [\mathbf{a}_{1,f\ell}^\top \ldots \mathbf{a}_{J,f\ell}^\top]^\top$. In the following $\mathbf{a}_{:,f\ell}$ is referred to as the *mixing vector*. Then we assume that for every frequency f the sequence of the L unobserved mixing vectors $\{\mathbf{a}_{:,f\ell}\}_{\ell=1}^L$ is ruled by a first-order LDS, where both the prior distribution and the process noise are assumed complex Gaussian. Formally, this writes:

$$p(\mathbf{a}_{:,f\ell}|\mathbf{a}_{:,f\ell-1}) = \mathcal{N}_c(\mathbf{a}_{:,f\ell}; \mathbf{a}_{:,f\ell-1}, \boldsymbol{\Sigma}_f^a), \quad (5)$$

$$p(\mathbf{a}_{:,f1}) = \mathcal{N}_c(\mathbf{a}_{:,f1}; \boldsymbol{\mu}_f^a, \boldsymbol{\Sigma}_f^a), \quad (6)$$

where the mean vector $\boldsymbol{\mu}_f^a \in \mathbb{C}^{IJ}$ and the *evolution* covariance matrix $\boldsymbol{\Sigma}_f^a \in \mathbb{C}^{IJ\times IJ}$ are parameters to be estimated. $\boldsymbol{\Sigma}_f^a$ is expected to reflect the amplitude of variations in the channel. Importantly, the time-invariant mixing model of [12] corresponds to the particular case in the proposed model when $\boldsymbol{\Sigma}_f^a \to \mathbf{0}_{IJ\times IJ}$. Indeed, in that case the latent state $\mathbf{a}_{:,f\ell}$ collapses to $\mathbf{a}_{:,f1}$ and hence the mixing matrix $\mathbf{A}_{f\ell}$ reduces to its time-invariant version $\mathbf{A}_f$. The complete graphical model of the proposed probabilistic model for ASS of time-varying convolutive mixtures is given in Fig. 1.

The standard way to perform inference in LDS is the *Kalman smoother* (or the *Kalman filter* if only causal observations are used). Eq. (4) defines the *observation model* of the Kalman smoother.[4] However, since part of the observation model, for instance $\mathbf{s}_{f\ell}$, is a latent variable, the direct application of the classical Kalman technique is infeasible in our case. In other words, we need to infer both latent variables: the mixing filters and the sources/components. For this purpose, in the next section we introduce a VEM procedure that alternates between (i) the complex Kalman smoother to infer the mixing filters sequence, (ii) the Wiener filter to estimate the sources and (iii) update rules for the parameters. Importantly, this result is a consequence of the joint effect of the proposed model and the variational approximation.

[4]The vectorized form of the latent mixing filters can be made explicit in the observation model by rewriting it as $\mathbf{x}_{f\ell} = \left(\mathbf{s}_{f\ell}^\top \otimes \mathbf{I}_I\right)\mathbf{a}_{:,f\ell} + \mathbf{b}_{f\ell}$, with $\otimes$ denoting the Kronecker matrix product.

III. VEM for Source Separation

In this section, we present the proposed VEM algorithm that alternates between the inference of the latent variables and the update of the parameters. We start with stating the principle of VEM. Then we present the E-step, farther decomposed in an E-A step for the mixing vector sequence and an E-S/C step for source/component coefficients, and then the M-step. The following notations are introduced: $\mathbb{E}_q$ is the expectation with respect to q, $\hat{\mathbf{z}} = \mathbb{E}_{q(\mathbf{z})}[\mathbf{z}]$ is the posterior mean vector of a random vector $\mathbf{z}$, $\boldsymbol{\Sigma}^{\eta z} = \mathbb{E}_{q(\mathbf{z})}[(\mathbf{z}-\hat{\mathbf{z}})(\mathbf{z}-\hat{\mathbf{z}})^{\mathrm{H}}]$ is its posterior covariance matrix, and $\mathbf{Q}^{\eta z} = \mathbb{E}_{q(\mathbf{z})}[\mathbf{z}\mathbf{z}^{\mathrm{H}}] = \boldsymbol{\Sigma}^{\eta z} + \hat{\mathbf{z}}\hat{\mathbf{z}}^{\mathrm{H}}$ is its second-order posterior moment. In general, superscript η denotes parameters of *posterior* distributions, whereas no superscript denotes parameters of *prior* distributions. The posterior mean is the estimate of the corresponding latent variable, provided by our algorithm. Also, let $\Sigma_{kg,f\ell}$ denote the (k,g)th entry of matrix $\boldsymbol{\Sigma}_{f\ell}$. Let $\overset{ct}{=}$ denote equality up to an additive term that is independent of the variable at stake, and let $\text{tr}\{\cdot\}$ denote the trace operator. For brevity $\mathbf{a}_{:,f1:L} = \{\mathbf{a}_{:,f\ell}\}_{\ell=1}^L$ denotes the whole sequence of mixing vectors at frequency f.

A. Variational Inference Principle

EM is a standard procedure to find maximum likelihood (ML) estimates in the presence of hidden variables [33], [41]. By alternating between the evaluation of the posterior distribution of the hidden variables (E-step) and the maximization of the expected complete-data log-likelihood (M-step), EM provides ML parameter estimates from the set of observations $\{\mathbf{x}_{f\ell}\}_{f,\ell=1}^{F,L}$. In this work the set of hidden variables $\mathcal{H} = \{\mathbf{a}_{:,f\ell}, \mathbf{s}_{f\ell}, \mathbf{c}_{f\ell}\}_{f,\ell=1}^{F,L}$ consists of the mixing vectors and the source (or the component) coefficients. The parameter set $\theta = \{\boldsymbol{\mu}_f^a, \boldsymbol{\Sigma}_f^a, w_{fk}, h_{k\ell}, \mathrm{v}_f\}_{f,\ell,k=1}^{F,L,K}$ consists of the channel evolution parameters, the source NMF parameters, and the variance of the sensor noise.

In our case, the posterior distribution of the latent variables, $q(\mathcal{H}) = p(\mathcal{H}|\{\mathbf{x}_{f\ell}\}_{f,\ell=1}^{F,L};\theta)$ cannot be expressed in closed-form. Therefore we develop a variational inference procedure [33], [42], based on the following principle. First, $q(\mathcal{H})$ is assumed to factorize into marginal posterior distributions over a partition of the latent variables. An approximation of the marginal posterior distribution of a subset of latent variables $\mathcal{H}_0 \subseteq \mathcal{H}$ is then computed with

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

where $q(\mathcal{H}/\mathcal{H}_0)$ is the approximation of the joint posterior distribution of all hidden variables, except the subset $\mathcal{H}_0$. Subsequently, $q(\mathcal{H})$ can be inferred in an alternating manner for each $\mathcal{H}_0 \subset \mathcal{H}$. In the present work, we assume that the mixing filters and the source coefficients are conditionally independent given

the observations. Therefore, the posterior distribution[5] naturally factorizes as:

$$q(\mathcal{H}) \approx \prod_{f=1}^{F} q(\mathbf{a}_{:,f1:L}) \prod_{f,\ell=1}^{F,L} q(\mathbf{s}_{f\ell}). \tag{8}$$

Note that the factorization over frequency (for both sources and filters) and over time (for the sources) arises naturally from the prior distributions and from the observation model (4).

B. E-A Step

Using (7) it is straightforward to show that the joint posterior distribution of the mixing vector sequence writes:

$$q(\mathbf{a}_{:,f1:L}) \propto p(\mathbf{a}_{:,f1:L}) \prod_{\ell=1}^{L} \exp\left(\mathbb{E}_{q(\mathbf{s}_{f\ell})}\left[\log p(\mathbf{x}_{f\ell}|\mathbf{A}_{f\ell},\mathbf{s}_{f\ell})\right]\right). \tag{9}$$

We have:

$$\begin{aligned}\mathbb{E}_{q(\mathbf{s}_{f\ell})}\left[\log p(\mathbf{x}_{f\ell}|\mathbf{A}_{f\ell},\mathbf{s}_{f\ell})\right] &\overset{ct}{=} \\ -\operatorname{tr}\left\{\mathbb{E}_{q(\mathbf{s}_{f\ell})}\left[(\mathbf{x}_{f\ell}-\mathbf{A}_{f\ell}\mathbf{s}_{f\ell})(\mathbf{x}_{f\ell}-\mathbf{A}_{f\ell}\mathbf{s}_{f\ell})^{\mathrm{H}}\right]\mathrm{v}_f{}^{-1}\right\} &\overset{ct}{=} \\ -\operatorname{tr}\left\{\frac{\mathbf{I}_I}{\mathrm{v}_f}\left(\mathbf{A}_{f\ell}-\mathbf{M}^{\iota a}_{f\ell}\right)\mathbf{Q}^{\eta s}_{f\ell}\left(\mathbf{A}_{f\ell}-\mathbf{M}^{\iota a}_{f\ell}\right)^{\mathrm{H}}\right\}, \end{aligned} \tag{10}$$

where $\mathbf{M}^{\iota a}_{f\ell} = \mathbf{x}_{f\ell}\hat{\mathbf{s}}^{\mathrm{H}}_{f\ell}(\mathbf{Q}^{\eta s}_{f\ell})^{-1} \in \mathbb{C}^{I\times J}$, with $\hat{\mathbf{s}}_{f\ell}$ and $\mathbf{Q}^{\eta s}_{f\ell}$ provided by the E-S step in Section III-C. By defining $\boldsymbol{\mu}^{\iota a}_{f\ell} = \mathrm{vec}(\mathbf{M}^{\iota a}_{f\ell}) \in \mathbb{C}^{IJ}$, (10) can be reorganized as:

$$\begin{aligned}\mathbb{E}_{q(\mathbf{s}_{f\ell})}\left[\log p(\mathbf{x}_{f\ell}|\mathbf{A}_{f\ell},\mathbf{s}_{f\ell})\right] &\overset{ct}{=} \\ -(\mathbf{a}_{:,f\ell}-\boldsymbol{\mu}^{\iota a}_{f\ell})^{\mathrm{H}}\left(\mathbf{Q}^{\eta s\top}_{f\ell}\otimes\frac{\mathbf{I}_I}{\mathrm{v}_f}\right)(\mathbf{a}_{:,f\ell}-\boldsymbol{\mu}^{\iota a}_{f\ell}). \end{aligned} \tag{11}$$

Let us define $\boldsymbol{\Sigma}^{\iota a}_{f\ell} = \left(\mathbf{Q}^{\eta s\top}_{f\ell}\otimes\mathbf{I}_I\mathrm{v}_f^{-1}\right)^{-1} \in \mathbb{C}^{IJ\times IJ}$. This matrix is Hermitian positive definite and (11) characterizes a complex Gaussian distribution with mean $\boldsymbol{\mu}^{\iota a}_{f\ell}$ and covariance $\boldsymbol{\Sigma}^{\iota a}_{f\ell}$. By substituting (11) in (9), we obtain:

$$q(\mathbf{a}_{:,f1:L}) \propto p(\mathbf{a}_{:,f1:L}) \prod_{\ell=1}^{L} \mathcal{N}_c(\boldsymbol{\mu}^{\iota a}_{f\ell};\mathbf{a}_{:,f\ell},\boldsymbol{\Sigma}^{\iota a}_{f\ell}). \tag{12}$$

Functional $\mathcal{N}_c(\boldsymbol{\mu}^{\iota a}_{f\ell};\mathbf{a}_{:,f\ell},\boldsymbol{\Sigma}^{\iota a}_{f\ell})$ can be viewed as an *instantaneous* distribution of a *measured* vector $\boldsymbol{\mu}^{\iota a}_{f\ell}$, conditioned to the *hidden* variable $\mathbf{a}_{:,f\ell}$. Henceforth one recognizes that (12) represents an LDS with continuous hidden state variables $\{\mathbf{a}_{:,f\ell}\}_{\ell=1}^{L}$, transition distribution given by (5), initial distribution given by (6), and *emission* distribution given by $\mathcal{N}_c(\boldsymbol{\mu}^{\iota a}_{f\ell};\mathbf{a}_{:,f\ell},\boldsymbol{\Sigma}^{\iota a}_{f\ell})$. Subsequently the marginal posterior distribution of each hidden state, $q(\mathbf{a}_{:,f\ell})$, can be calculated recursively using a *forward-backward* algorithm [33], aka *Kalman smoother*.

1) Forward-Backward Algorithm: Given the LDS parameters, a forward-backward algorithm computes an estimate $\hat{\mathbf{a}}_{:,f\ell}$ for all ℓ by taking into account all causal measurements (from 1 to ℓ) and anti-causal measurements (from $\ell+1$ to L). The implementation of the forward-backward algorithm thus consists of a recursive forward pass and a recursive backward pass. Different variants for this algorithm are available. The *forward-backward* procedure that we specifically designed to infer (12) is described below. Because of the form of (5), all covariance updates of this forward-backward algorithm are computable using only additions and matrix inversion. Indeed it is desirable to avoid subtractions and matrix multiplications of covariance matrices since these operations do not guarantee that (with Hermitian operands) the resulting matrix is Hermitian. As a result, the proposed Kalman smoother was found to be very stable from a numerical point of view. In addition, since all distributions under consideration are complex Gaussian, the outcome of the forward-backward recursions will also be complex Gaussian [33].

The forward pass recursively provides the joint distribution of the state variable and the causal observations. The mean vector $\boldsymbol{\mu}^{\phi a}_{f\ell} \in \mathbb{C}^{IJ}$ and covariance matrix $\boldsymbol{\Sigma}^{\phi a}_{f\ell} \in \mathbb{C}^{IJ\times IJ}$ of this distribution are calculated as:

$$\boldsymbol{\Sigma}^{\phi a}_{f\ell} = \left(\boldsymbol{\Sigma}^{\iota a\,-1}_{f\ell} + \left(\boldsymbol{\Sigma}^{\phi a}_{f\ell-1}+\boldsymbol{\Sigma}^{a}_{f}\right)^{-1}\right)^{-1}, \tag{13}$$

$$\boldsymbol{\mu}^{\phi a}_{f\ell} = \boldsymbol{\Sigma}^{\phi a}_{f\ell}\left(\boldsymbol{\Sigma}^{\iota a\,-1}_{f\ell}\boldsymbol{\mu}^{\iota a}_{f\ell} + \left(\boldsymbol{\Sigma}^{\phi a}_{f\ell-1}+\boldsymbol{\Sigma}^{a}_{f}\right)^{-1}\boldsymbol{\mu}^{\phi a}_{f\ell-1}\right). \tag{14}$$

The backward pass recursively provides the distribution of the anti-causal observations given the current state. The mean vector $\boldsymbol{\mu}^{\beta a}_{f\ell} \in \mathbb{C}^{IJ}$ and covariance matrix $\boldsymbol{\Sigma}^{\beta a}_{f\ell} \in \mathbb{C}^{IJ\times IJ}$ of this distribution are calculated as:

$$\boldsymbol{\Sigma}^{\zeta a}_{f\ell} = \left(\boldsymbol{\Sigma}^{\iota a}_{f\ell+1}{}^{-1} + \boldsymbol{\Sigma}^{\beta a}_{f\ell+1}{}^{-1}\right)^{-1}, \tag{15}$$

$$\boldsymbol{\Sigma}^{\beta a}_{f\ell} = \boldsymbol{\Sigma}^{a}_{f} + \boldsymbol{\Sigma}^{\zeta a}_{f\ell}, \tag{16}$$

$$\boldsymbol{\mu}^{\beta a}_{f\ell} = \boldsymbol{\Sigma}^{\zeta a}_{f\ell}\left(\boldsymbol{\Sigma}^{\iota a}_{f\ell+1}{}^{-1}\boldsymbol{\mu}^{\iota a}_{f\ell+1} + \boldsymbol{\Sigma}^{\beta a}_{f\ell+1}{}^{-1}\boldsymbol{\mu}^{\beta a}_{f\ell+1}\right), \tag{17}$$

where $\boldsymbol{\Sigma}^{\zeta a}_{f\ell} \in \mathbb{C}^{IJ\times IJ}$ is an intermediate matrix that enables to express the backward recursion without subtractions.

2) Posterior Estimate of the Mixing Vector: Let us now calculate the *smoothed* estimate $\hat{\mathbf{a}}_{:,f\ell}$. By composing the forward and the backward estimates, the marginal (frame-wise) posterior distribution of $\mathbf{a}_{:,f\ell}$ writes [33]:

$$q(\mathbf{a}_{:,f\ell}) = \mathcal{N}_c(\mathbf{a}_{:,f\ell};\hat{\mathbf{a}}_{:,f\ell},\boldsymbol{\Sigma}^{\eta a}_{f\ell}), \tag{18}$$

with $\boldsymbol{\Sigma}^{\eta a}_{f\ell} \in \mathbb{C}^{IJ\times IJ}$ and $\hat{\mathbf{a}}_{:,f\ell} \in \mathbb{C}^{IJ}$ computed as:

$$\boldsymbol{\Sigma}^{\eta a}_{f\ell} = \left(\boldsymbol{\Sigma}^{\phi a}_{f\ell}{}^{-1} + \boldsymbol{\Sigma}^{\beta a}_{f\ell}{}^{-1}\right)^{-1}, \tag{19}$$

$$\hat{\mathbf{a}}_{:,f\ell} = \boldsymbol{\Sigma}^{\eta a}_{f\ell}\left(\boldsymbol{\Sigma}^{\phi a}_{f\ell}{}^{-1}\boldsymbol{\mu}^{\phi a}_{f\ell} + \boldsymbol{\Sigma}^{\beta a}_{f\ell}{}^{-1}\boldsymbol{\mu}^{\beta a}_{f\ell}\right). \tag{20}$$

3) Joint Posterior Distribution of a Pair of Successive Mixing Vectors: This joint distribution will be needed to update $\boldsymbol{\Sigma}^{a}_{f}$ in Section III-F. Let $\mathbf{a}_{:,f\{\ell+1,\ell\}} = \left[\mathbf{a}^{\top}_{:,f\ell+1},\mathbf{a}^{\top}_{:,f\ell}\right]^{\top} \in \mathbb{C}^{2IJ}$ denote the joint variable. By marginalizing out all mixing vectors except $\mathbf{a}_{:,f\ell+1},\mathbf{a}_{:,f\ell}$ in (12), the joint posterior distribution $q(\mathbf{a}_{:,f\{\ell+1,\ell\}})$ can be identified to be also a Gaussian distribution with mean vector $\boldsymbol{\mu}^{\xi a}_{f\ell} \in \mathbb{C}^{2IJ}$ and covariance matrix

[5]From now on, we abuse the language and refer to q as the posterior distribution, even if technically it is only a variational approximation of it.

$\boldsymbol{\Sigma}^{\xi a}_{f\ell} \in \mathbb{C}^{2IJ\times 2IJ}$ computed as:

$$\boldsymbol{\Sigma}^{\xi a}_{f\ell} = \begin{bmatrix} \boldsymbol{\Sigma}^{\zeta a}_{f\ell^{-1}} + \boldsymbol{\Sigma}^{a^{-1}}_{f} & -\boldsymbol{\Sigma}^{a^{-1}}_{f} \\ -\boldsymbol{\Sigma}^{a^{-1}}_{f} & \boldsymbol{\Sigma}^{\phi a^{-1}}_{f\ell} + \boldsymbol{\Sigma}^{a^{-1}}_{f} \end{bmatrix}^{-1}, \tag{21}$$

$$\boldsymbol{\mu}^{\xi a}_{f\ell} = \boldsymbol{\Sigma}^{\xi a}_{f\ell} \left[\left(\boldsymbol{\Sigma}^{\zeta a^{-1}}_{f\ell} \boldsymbol{\mu}^{\beta a}_{f\ell+1} \right)^{\top}, \left(\boldsymbol{\Sigma}^{\phi a^{-1}}_{f\ell} \boldsymbol{\mu}^{\phi a}_{f\ell} \right)^{\top} \right]^{\top}. \tag{22}$$

Note here the role of $\boldsymbol{\Sigma}^{\zeta a}_{f\ell}$ that is to describe the uncertainty of $\boldsymbol{\mu}^{\beta a}_{f\ell+1}$ but without incorporating the additional uncertainty of the transition variance $\boldsymbol{\Sigma}^{a}_{f}$, as the transition from $\mathbf{a}_{:,f\ell}$ to $\mathbf{a}_{:,f\ell+1}$ is explicitly defined by the joint variable $\mathbf{a}_{:,f\{\ell+1,\ell\}}$.

C. E-S Step and E-C Step

From (7), the posterior distribution of the sources writes:

$$q(\mathbf{s}_{f\ell}) \propto p(\mathbf{s}_{f\ell}) \exp\left(\mathbb{E}_{q(\mathbf{a}_{:,f\ell})} \left[\log p(\mathbf{x}_{f\ell} | \mathbf{A}_{f\ell}, \mathbf{s}_{f\ell}) \right] \right). \tag{23}$$

Using (4), the expectation in (23) computes:

$$\mathbb{E}_{q(\mathbf{a}_{:,f\ell})} \left[\log p(\mathbf{x}_{f\ell} | \mathbf{A}_{f\ell}, \mathbf{s}_{f\ell}) \right] \stackrel{ct}{=} \frac{1}{\mathrm{v}_f} \mathrm{tr}\Big\{ \mathbf{s}_{f\ell} \big(\hat{\mathbf{A}}^{\mathrm{H}}_{f\ell} \mathbf{x}_{f\ell} \big)^{\mathrm{H}} + \big(\hat{\mathbf{A}}^{\mathrm{H}}_{f\ell} \mathbf{x}_{f\ell} \big) \mathbf{s}^{\mathrm{H}}_{f\ell} - \mathbf{U}_{f\ell} \mathbf{s}_{f\ell} \mathbf{s}^{\mathrm{H}}_{f\ell} \Big\}, \tag{24}$$

where $\hat{\mathbf{A}}_{f\ell} = \mathbb{E}_{q(\mathbf{a}_{:,f\ell})}[\mathbf{A}_{f\ell}] \in \mathbb{C}^{I\times J}$ is a matrix constructed from $\hat{\mathbf{a}}_{:,f\ell}$ (i.e. the reverse operation of column-wise vectorization), and $\mathbf{U}_{f\ell} = \mathbb{E}_{q(\mathbf{a}_{:,f\ell})}[\mathbf{A}^{\mathrm{H}}_{f\ell}\mathbf{A}_{f\ell}] \in \mathbb{C}^{J\times J}$. Of course, $\mathbf{U}_{f\ell}$ is closely related to $\mathbf{Q}^{\eta a}_{f\ell}$. Indeed, if we define $\mathbf{Q}^{\eta a}_{jr,f\ell} = \mathbb{E}_{q(\mathbf{a}_{:,f\ell})}[\mathbf{a}_{j,f\ell}\mathbf{a}^{\mathrm{H}}_{r,f\ell}]$ as the (j,r)-th $I \times I$ subblock of $\mathbf{Q}^{\eta a}_{f\ell}$, then each entry $U_{jr,f\ell}$ of $\mathbf{U}_{f\ell}$ is simply given by:

$$U_{jr,f\ell} = \mathbb{E}_{q(\mathbf{a}_{:,f\ell})}[\mathbf{a}^{\mathrm{H}}_{j,f\ell}\mathbf{a}_{r,f\ell}] = \mathrm{tr}\Big\{ \mathbf{Q}^{\eta a}_{rj,f\ell} \Big\}. \tag{25}$$

Eq. (24) is an incomplete quadratic form in $\mathbf{s}_{f\ell}$. Combining in (23) this quadratic form with the quadratic form of the source prior $p(\mathbf{s}_{f\ell})$, we obtain a multivariate Gaussian:

$$q(\mathbf{s}_{f\ell}) = \mathcal{N}_c(\mathbf{s}_{f\ell}; \hat{\mathbf{s}}_{f\ell}, \boldsymbol{\Sigma}^{\eta s}_{f\ell}), \tag{26}$$

with mean vector $\hat{\mathbf{s}}_{f\ell} \in \mathbb{C}^J$ and covariance matrix $\boldsymbol{\Sigma}^{\eta s}_{f\ell} \in \mathbb{C}^{J\times J}$ given by:

$$\boldsymbol{\Sigma}^{\eta s}_{f\ell} = \left[\mathrm{diag}_J \left(\frac{1}{\sum_{k\in\mathcal{K}_j} w_{fk} h_{k\ell}} \right) + \frac{\mathbf{U}_{f\ell}}{\mathrm{v}_f} \right]^{-1}, \tag{27}$$

$$\hat{\mathbf{s}}_{f\ell} = \boldsymbol{\Sigma}^{\eta s}_{f\ell} \hat{\mathbf{A}}^{\mathrm{H}}_{f\ell} \frac{\mathbf{x}_{f\ell}}{\mathrm{v}_f}. \tag{28}$$

Remarkably, (28) has a form similar to the source estimator in [12], namely a Wiener filtering estimator, with two notable differences. First, in [12] the mixing matrix is an estimated parameter, whereas here it is the posterior expectation $\hat{\mathbf{A}}_{f\ell}$ of the latent mixing matrix. Second, the source posterior precision matrix $(\boldsymbol{\Sigma}^{\eta s}_{f\ell})^{-1}$ is built by summation of (i) the sensor precision $1/\mathrm{v}_f$ distributed over the sources with the unit-less quantity $\mathbf{U}_{f\ell}$, and of (ii) the diagonal prior precision of the source coefficients given by the NMF model (as in [12]). In other words, the a posteriori uncertainty of the sources encompasses the a priori uncertainty (the NMF), the channel noise (v_f), and the channel uncertainty ($\mathbf{U}_{f\ell}$).

A similar E-step can be applied to the source components $\mathbf{c}_{f\ell}$. This will be used in Section III-G to optimize the NMF parameters. For this aim, we simply replace $\mathbf{A}_{f\ell}$ with $\mathbf{A}_{f\ell}\mathbf{G}$, and $p(\mathbf{s}_{f\ell})$ with $p(\mathbf{c}_{f\ell})$, obtaining again a complex Gaussian for the posterior distribution of the components:

$$q(\mathbf{c}_{f\ell}) = \mathcal{N}_c(\mathbf{c}_{f\ell}; \hat{\mathbf{c}}_{f\ell}, \boldsymbol{\Sigma}^{\eta c}_{f\ell}), \tag{29}$$

with parameters $\hat{\mathbf{c}}_{f\ell} \in \mathbb{C}^K$ and $\boldsymbol{\Sigma}^{\eta c}_{f\ell} \in \mathbb{C}^{K\times K}$ given by:

$$\boldsymbol{\Sigma}^{\eta c}_{f\ell} = \left[\mathrm{diag}_K \left(\frac{1}{w_{fk} h_{k\ell}} \right) + \mathbf{G}^{\top} \frac{\mathbf{U}_{f\ell}}{\mathrm{v}_f} \mathbf{G} \right]^{-1}, \tag{30}$$

$$\hat{\mathbf{c}}_{f\ell} = \boldsymbol{\Sigma}^{\eta c}_{f\ell} \mathbf{G}^{\top} \hat{\mathbf{A}}^{\mathrm{H}}_{f\ell} \frac{\mathbf{x}_{f\ell}}{\mathrm{v}_f}. \tag{31}$$

Again, (31) is a Wiener filtering estimator, here at the source component vector level. Note that left-multiplication of both sides of (31) by $\mathbf{G}$ naturally leads to (28).

D. Outline of the Maximization Step

Once we have the posterior distributions of the variables in $\mathcal{H}$, the *expected complete-data log-likelihood* $\mathcal{L}(\theta) = \mathbb{E}_{q(\mathcal{H})} \log p\big(\mathcal{H}, \{\mathbf{x}_{f\ell}\}^{F,L}_{f,\ell=1}; \theta\big)$ is maximized with respect to the parameters. The analytic expression of $\mathcal{L}(\theta)$ is

$$\begin{aligned} \mathcal{L}(\theta) = & \sum_{f,\ell=1}^{F,L} \mathbb{E}_{q(\mathbf{a}_{:,f\ell})q(\mathbf{s}_{f\ell})} \left[\log \mathcal{N}_c(\mathbf{x}_{f\ell}; \mathbf{A}_{f\ell}\mathbf{s}_{f\ell}, \mathrm{v}_f \mathbf{I}_I) \right] \\ & + \sum_{f,\ell=1}^{F,L} \mathbb{E}_{q(\mathbf{c}_{f\ell})} \left[\log \mathcal{N}_c \left(\mathbf{c}_{f\ell}; \mathbf{0}, \mathrm{diag}_K \left(w_{fk} h_{k\ell} \right) \right) \right] \\ & + \sum_{f=1}^{F} \Bigg(\sum_{\ell=1}^{L-1} \mathbb{E}_{q(\mathbf{a}_{:,f\{\ell+1,\ell\}})} \left[\log \mathcal{N}_c \left(\mathbf{a}_{:,f\ell+1}; \mathbf{a}_{:,f\ell}, \boldsymbol{\Sigma}^{a}_{f} \right) \right] \\ & + \mathbb{E}_{q(\mathbf{a}_{:,f1})} \left[\log \mathcal{N}_c \left(\mathbf{a}_{:,f1}; \boldsymbol{\mu}^{a}_{f}, \boldsymbol{\Sigma}^{a}_{f} \right) \right] \Bigg). \end{aligned} \tag{32}$$

Notice that (32) can be optimized w.r.t. the microphone noise parameters, the channel parameters, or the NMF parameters, independently.

E. M-V Step

Derivating $\mathcal{L}(\theta)$ w.r.t. v_f, and setting the result to zero, leads to the following update:

$$\begin{aligned} \mathrm{v}_f = \frac{1}{LI} \sum_{\ell=1}^{L} \Big(& \mathbf{x}^{\mathrm{H}}_{f\ell}\mathbf{x}_{f\ell} - \mathbf{x}^{\mathrm{H}}_{f\ell}\hat{\mathbf{A}}_{f\ell}\hat{\mathbf{s}}_{f\ell} \\ & - \big(\hat{\mathbf{A}}_{f\ell}\hat{\mathbf{s}}_{f\ell} \big)^{\mathrm{H}} \mathbf{x}_{f\ell} + \mathrm{tr}\big\{ \mathbf{U}_{f\ell}\mathbf{Q}^{\eta s}_{f\ell} \big\} \Big), \end{aligned} \tag{33}$$

which resembles the estimator obtained in [12].

F. M-A Step

Optimizing $\mathcal{L}(\theta)$ w.r.t. the prior mean $\boldsymbol{\mu}^{a}_{f}$ results in the following update:

$$\boldsymbol{\mu}^{a}_{f} = \hat{\mathbf{a}}_{f1}. \tag{34}$$

The ML initial vector is thus the posterior mean vector for $\ell = 1$. The way the E-A step was designed, (34) becomes rather important.

As for $\boldsymbol{\Sigma}_f^a$, the terms of $\mathcal{L}(\theta)$ that depend on this parameter reduce to:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\Sigma}_f^a) &\equiv \sum_{\ell=1}^{L-1} \mathbb{E}_{q(\mathbf{a}_{:,f\{\ell+1,\ell\}})}\left[\log \mathcal{N}_c\left(\mathbf{a}_{:,f\ell+1}; \mathbf{a}_{:,f\ell}, \boldsymbol{\Sigma}_f^a\right)\right] \\ &\quad + \mathbb{E}_{q(\mathbf{a}_{:,f1})}\left[\log \mathcal{N}_c\left(\mathbf{a}_{:,f1}; \boldsymbol{\mu}_f^a, \boldsymbol{\Sigma}_f^a\right)\right] \\ &\overset{ct}{=} -\operatorname{tr}\left\{\boldsymbol{\Sigma}_f^{a^{-1}} \boldsymbol{\Sigma}_{f1}^{\eta a}\right\} \\ &\quad - \operatorname{tr}\left\{\begin{bmatrix} \boldsymbol{\Sigma}_f^{a^{-1}} & -\boldsymbol{\Sigma}_f^{a^{-1}} \\ -\boldsymbol{\Sigma}_f^{a^{-1}} & \boldsymbol{\Sigma}_f^{a^{-1}} \end{bmatrix} \mathbf{Q}_f^{\xi a}\right\} - L \log |\boldsymbol{\Sigma}_f^a| \\ &= -L \log |\boldsymbol{\Sigma}_f^a| \\ &\quad - \operatorname{tr}\left\{\boldsymbol{\Sigma}_f^{a^{-1}}\left[\boldsymbol{\Sigma}_{f1}^{\eta a} + \mathbf{Q}_{11,f}^{\xi a} - \mathbf{Q}_{12,f}^{\xi a} - \mathbf{Q}_{21,f}^{\xi a} + \mathbf{Q}_{22,f}^{\xi a}\right]\right\}.\end{aligned} \tag{35}$$

In the above equation $\mathbf{Q}_f^{\xi a} \in \mathbb{C}^{2IJ \times 2IJ}$ is the cumulate second-order joint posterior moment of $\mathbf{a}_{:,f\{\ell+1,\ell\}}$, and the four $\mathbf{Q}_{nm,f}^{\xi a}$ matrices are its $IJ \times IJ$ non-overlapping principal subblocks, i.e.:

$$\mathbf{Q}_f^{\xi a} = \sum_{\ell=1}^{L-1}\left(\boldsymbol{\Sigma}_{f\ell}^{\xi a} + \boldsymbol{\mu}_{f\ell}^{\xi a}(\boldsymbol{\mu}_{f\ell}^{\xi a})^{\mathrm{H}}\right) = \begin{bmatrix} \mathbf{Q}_{11,f}^{\xi a} & \mathbf{Q}_{12,f}^{\xi a} \\ \mathbf{Q}_{21,f}^{\xi a} & \mathbf{Q}_{22,f}^{\xi a} \end{bmatrix}. \tag{36}$$

Derivating (35) w.r.t. the entries of $\boldsymbol{\Sigma}_f^a$, and setting the result to zero, yields [43]:

$$\boldsymbol{\Sigma}_f^a = \frac{1}{L}\left(\mathbf{Q}_{11,f}^{\xi a} - \mathbf{Q}_{12,f}^{\xi a} - \mathbf{Q}_{21,f}^{\xi a} + \mathbf{Q}_{22,f}^{\xi a} + \boldsymbol{\Sigma}_{f1}^{\eta a}\right). \tag{37}$$

G. M-C Step and M-S Step

The joint optimization of $\mathcal{L}(\theta)$ over w_{fk} and $h_{k\ell}$ is non-convex. However alternate maximization is a classical solution to solve for a locally-optimal set of NMF parameters [17]. Calculating the derivatives of $\mathcal{L}(\theta)$ w.r.t. to w_{fk} and $h_{k\ell}$ and setting the result to zero leads to the following update formulae:

$$w_{fk} = \frac{1}{L}\sum_{\ell=1}^{L} \frac{Q_{kk,f\ell}^{\eta c}}{h_{k\ell}}, \quad h_{k\ell} = \frac{1}{F}\sum_{f=1}^{F} \frac{Q_{kk,f\ell}^{\eta c}}{w_{fk}}. \tag{38}$$

This formulae can be iteratively applied until convergence, although in an effort to avoid local optima, each of w_{fk}, $h_{k\ell}$ was updated only once at each VEM iteration.

H. Estimation of Source Images

As is often the case in source separation, the proposed framework suffers from the well-known scale ambiguity, namely the source signals and the mixing matrices can only be estimated up to (frequency-dependent) compensating multiplicative factors [1]. To alleviate this problem and to be able to assess the performance of source separation, we consider the separation of the source images, i.e. the source signals as recorded by the microphones [13], [44], instead of the (monophonic) source signals. For this purpose, the inverse STFT is applied to $\{\mathbb{E}_{q(\mathbf{a}_{:,f\ell},\mathbf{s}_{f\ell})}[\mathbf{a}_{j,f\ell}s_{j,f\ell}] = \hat{\mathbf{a}}_{j,f\ell}\hat{s}_{j,f\ell}\}_{f,\ell=1}^{F,L}$, where $\hat{\mathbf{a}}_{j,f\ell}$ is the j-th column of $\hat{\mathbf{A}}_{f\ell}$. The complete VEM separating J sound sources from an I-channel time-varying mixture is outlined in Algorithm 1 (omitting STFT and inverse STFT for clarity).

Algorithm 1: Proposed VEM for the separation of sound sources mixed with time-varying filters

input $\{\mathbf{x}_{f\ell}\}_{f,\ell=1}^{F,L}$, partition matrix $\mathbf{G}$, initial parameters θ.
initialize posterior statistics $\hat{\mathbf{a}}_{:,f\ell}, \boldsymbol{\Sigma}_{f\ell}^{\eta a}$.
repeat
  Variational E-step
  Calculate $\mathbf{Q}_{f\ell}^{\eta a} = \boldsymbol{\Sigma}_{f\ell}^{\eta a} + \hat{\mathbf{a}}_{:,f\ell}\hat{\mathbf{a}}_{:,f\ell}^{\mathrm{H}}$ and $\mathbf{U}_{f\ell}$ with (25).
  E-S step: Compute $\boldsymbol{\Sigma}_{f\ell}^{\eta s}$ with (27) and $\hat{\mathbf{s}}_{f\ell}$ with (28).
    Then compute $\mathbf{Q}_{f\ell}^{\eta s} = \boldsymbol{\Sigma}_{f\ell}^{\eta s} + \hat{\mathbf{s}}_{f\ell}\hat{\mathbf{s}}_{f\ell}^{\mathrm{H}}$.
  E-C step: Compute $\Sigma_{kk,f\ell}^{\eta c}$ with (40) and $\hat{c}_{k,f\ell}$ with (41).
    Then compute $Q_{kk,f\ell}^{\eta c} = \Sigma_{kk,f\ell}^{\eta c} + |\hat{c}_{k,f\ell}|^2$.
  E-A step (Instantaneous Quantities):
    Compute $(\boldsymbol{\Sigma}_{f\ell}^{\iota a^{-1}}\boldsymbol{\mu}_{f\ell}^{\iota a})$ with (39).
    Compute $\boldsymbol{\Sigma}_{f\ell}^{\iota a^{-1}} = \mathbf{Q}_{f\ell}^{\eta s^\top} \otimes \mathbf{I}_I \mathrm{v}_f^{-1}$.
  E-A step (Forward Pass):
    Initialize $\boldsymbol{\Sigma}_{f1}^{\phi a} = \left(\boldsymbol{\Sigma}_{f1}^{\iota a^{-1}} + \boldsymbol{\Sigma}_f^{a^{-1}}\right)^{-1}$.
    Initialize $\boldsymbol{\mu}_{f1}^{\phi a} = \boldsymbol{\Sigma}_{f1}^{\phi a}\left(\boldsymbol{\Sigma}_{f1}^{\iota a^{-1}}\boldsymbol{\mu}_{f1}^{\iota a} + \boldsymbol{\Sigma}_f^{a^{-1}}\boldsymbol{\mu}_f^a\right)$.
    for $\ell : 2$ to L
      Compute $\boldsymbol{\Sigma}_{f\ell}^{\phi a}$ with (13), then $\boldsymbol{\mu}_{f\ell}^{\phi a}$ with (14).
    end
  E-A step (Backward Pass):
    Initialize $\boldsymbol{\Sigma}_{fL}^{\beta a} = \boldsymbol{\Sigma}_{fL}^{\phi a}$ and $\boldsymbol{\mu}_{fL}^{\beta a} = \boldsymbol{\mu}_{fL}^{\phi a}$.
    for $\ell : L-1$ to 1
      Compute $\boldsymbol{\Sigma}_{f\ell}^{\zeta a}$ with (15).
      Then compute $\boldsymbol{\Sigma}_{f\ell}^{\beta a}$ with (16).
      Then compute $\boldsymbol{\mu}_{f\ell}^{\beta a}$ with (17).
    end
  E-A step (Posterior Marginal Statistics):
    Compute $\boldsymbol{\Sigma}_{f\ell}^{\eta a}$ with (19).
    Then compute $\hat{\mathbf{a}}_{:,f\ell}$ with (20).
  E-A step (Pairwise Joint Posterior):
    Compute $\boldsymbol{\Sigma}_{f\ell}^{\xi a}$ with (21).
    Then compute $\boldsymbol{\mu}_{f\ell}^{\xi a}$ with (22).
    Then compute $\mathbf{Q}_f^{\xi a}$ with (36).
  M-step
  M-v step: Update v_f with (33).
  M-A step: Update $\boldsymbol{\mu}_f^a$ with (34), update $\boldsymbol{\Sigma}_f^a$ with (37).
  M-C step: Alternately update w_{fk} and $h_{k\ell}$ with (38).
until convergence
return the estimated source images $\hat{\mathbf{a}}_{j,f\ell}\hat{s}_{j,f\ell}, j \in [1, J]$.

IV. Implementation Issues

In this section we present some simplifications that our algorithm admits, we give physical interpretations, and we discuss some numerical stability issues.

A. Simplifying the LDS Measurement Vector

The forward-backward procedure requires the quantity $(\boldsymbol{\Sigma}_{f\ell}^{\iota a\,-1}\boldsymbol{\mu}_{f\ell}^{\iota a})$, appearing in (14) and (17). This can be computed as:

$$\boldsymbol{\Sigma}_{f\ell}^{\iota a\,-1}\boldsymbol{\mu}_{f\ell}^{\iota a} = \frac{1}{\mathrm{v}_f}\mathrm{vec}\big(\mathbf{x}_{f\ell}\hat{\mathbf{s}}_{f\ell}^{\mathrm{H}}\big), \tag{39}$$

thus sparing the inversion of $\boldsymbol{\Sigma}_{f\ell}^{\iota a}$.

B. Initializing the Forward and Backward Recursions

The forward-backward algorithm needs to set $\boldsymbol{\Sigma}_{f1}^{\phi a}$ and $\boldsymbol{\mu}_{f1}^{\phi a}$ for the first frame, and to set $\boldsymbol{\Sigma}_{fL}^{\beta a}$ and $\boldsymbol{\mu}_{fL}^{\beta a}$ for the last frame. We observed faster convergence with the following choice. At each VEM iteration, we set $\boldsymbol{\Sigma}_{f1}^{\phi a} = (\boldsymbol{\Sigma}_{f1}^{\iota a\,-1} + \boldsymbol{\Sigma}_f^{a^{-1}})^{-1}$ and $\boldsymbol{\mu}_{f1}^{\phi a} = \boldsymbol{\Sigma}_{f1}^{\phi a}(\boldsymbol{\Sigma}_{f1}^{\iota a\,-1}\boldsymbol{\mu}_{f1}^{\iota a} + \boldsymbol{\Sigma}_f^{a^{-1}}\boldsymbol{\mu}_f^a)$. Then, we run the forward pass first. After it is completed we set $\boldsymbol{\Sigma}_{fL}^{\beta a} = \boldsymbol{\Sigma}_{fL}^{\phi a}$, $\boldsymbol{\mu}_{fL}^{\beta a} = \boldsymbol{\mu}_{fL}^{\phi a}$, to initialize the backward pass.

C. Avoiding $K \times K$ Matrix Construction

Eq. (30) is computationally demanding as it requires the construction of a $K \times K$ matrix (recall that $K \gg J$). Yet, it has been shown in Section III-G that one needs only the diagonal entries of $\mathbf{Q}_{f\ell}^{\eta c}$. Therefore we derive an alternative expression for $\Sigma_{kk,f\ell}^{\eta c}$ and $\hat{c}_{k,f\ell}$ that builds on the already computed $\boldsymbol{\Sigma}_{f\ell}^{\eta s}$ and $\hat{\mathbf{s}}_{f\ell}$ (which use operations only on $J \times J$ arrays). Applying the *Woodbury* identity to (30) and some algebraic manipulations, one obtains:

$$\Sigma_{kk,f\ell}^{\eta c} = w_{fk}h_{k\ell}\left(1 - \frac{w_{fk}h_{k\ell}\left[\mathbf{U}_{f\ell}\boldsymbol{\Sigma}_{f\ell}^{\eta s}\right]_{j_k j_k}}{\mathrm{v}_f\sum_{\rho\in\mathcal{K}_{j_k}} w_{f\rho}h_{\rho\ell}}\right), \tag{40}$$

where j_k is the index of the source that the kth component belongs to, and $[\cdot]_{j_k j_k}$ is the j_k^{th} diagonal element of the $J \times J$ matrix in brackets. Additionally, $\hat{c}_{k,f\ell}$ can be expressed in a very simple way, independently of $\boldsymbol{\Sigma}_{f\ell}^{\eta c}$:

$$\hat{c}_{k,f\ell} = w_{fk}h_{k\ell}\left[\hat{\mathbf{A}}_{f\ell}^{\mathrm{H}}\frac{\mathbf{x}_{f\ell}}{\mathrm{v}_f} - \mathbf{U}_{f\ell}\frac{\hat{\mathbf{s}}_{f\ell}}{\mathrm{v}_f}\right]_{j_k}, \tag{41}$$

where $[\cdot]_{j_k}$ is the j_k^{th} element of the $J \times 1$ vector in brackets. Interestingly, (41) shows that $\hat{c}_{k,f\ell}$ is some kind of inpainting onto the mixture signal, whose purpose is to equalize the filtered mixture with the sources. Besides, (40) makes clear that if the value of v_f is high enough, the posterior variance of $c_{k,f\ell}$ remains close to its prior value $w_{fk}h_{k\ell}$. This justifies the use of a high initial value for v_f in cases where the NMF parameters are quite correctly initialized.

D. Ordering the Steps

When building a (V)EM algorithm, the question of ordering the steps execution arises. Like the majority of EMs, our algorithm is sensitive to initialization (discussed in Section V-A4). We observed in practice that our algorithm is much more sensitive to the initialization of the NMF parameters $w_{fk}, h_{k\ell}$ than to the initialization of (the posterior parameters of) the mixing vectors: $\boldsymbol{\Sigma}_{f\ell}^{\eta a}, \hat{\mathbf{a}}_{:,f\ell}$. Therefore we choose to first infer the source/component statistics by running E-S/C and then infer the sequence of mixing vectors by running E-A. As for the M-steps, they are independent and so they can be executed in any order after the E-steps.

E. NMF Scaling

When estimating the NMF parameters using (38), an arbitrary scale can circulate between $w_{fk}, h_{k\ell}$ of a component [12]. Therefore one can consider scaling one of the factors, e.g. $w_{fk} \leftarrow w_{fk}/\sum_n w_{nk}$, so to have unit $L1$-norm vectors, and reciprocally scaling the other factors, e.g. $h_{k\ell} \leftarrow h_{k\ell}\sum_f w_{fk}$ for compensation.

F. Numerical Stability

We enforce matrices $\mathbf{U}_{f\ell}$ and $\boldsymbol{\Sigma}_f^a$ to be Hermitian with $\boldsymbol{\Sigma}_f^a \leftarrow \frac{1}{2}(\boldsymbol{\Sigma}_f^a + \boldsymbol{\Sigma}_f^{a\,\mathrm{H}})$. We also regularized the updates of v_f and of $\boldsymbol{\Sigma}_{f\ell}^{\xi a}$, by adding 10^{-7} and $10^{-7}\mathbf{I}_{2IJ}$ respectively.

G. Computational Complexity

Counting only matrix multiplications, inversions and the solution of linear systems (assuming cubic complexity) the complexity order of the proposed VEM algorithm is $\mathcal{O}\big(16FL(IJ)^3 + 5FLG + F(L-1)(2IJ)^3\big)$. The experiments of this paper were conducted with a *HP Z800* desktop 4-core computer (8 threads) *Xeon E5620* CPU at 2.4 GHz and 17.6 GB of RAM. To process a 2s 16KHz stereo mixture, with $J = 3$, $K = 75$, $F = 512$, $L = 128$ our non-optimized implementation needs 30s per iteration, running in *MATLAB R2014a*, on *Fedora 20*. On the same data, the block-wise adaptation of the baseline method requires 4s for a complete iteration (an iteration for all blocks of frames). Hence, with this set-up, the complexity of the proposed method is about 8 times larger than the complexity of the baseline method.

V. Experimental Study

To assess the performance of the proposed model and associated VEM algorithm, we conducted a series of experiments with 2-channel time-varying convolutive mixtures of speech signals. Initialization is known to be a crucial step for the performance of (V)EM algorithms. In a general manner, EM-like algorithms have severe difficulties to converge to a relevant solution in totally blind setups (i.e. random initialization). A first series of experiments was thus conducted with simulated mixtures and artificially controlled (semi-blind) initialization of the VEM in order to extensively investigate its performance independently of initialization problems. Then a second series was conducted using a state-of-the art blind source separation method based on binary masking for the initialization. This latter configuration was first applied on simulated mixtures and then real-world recorded mixtures.

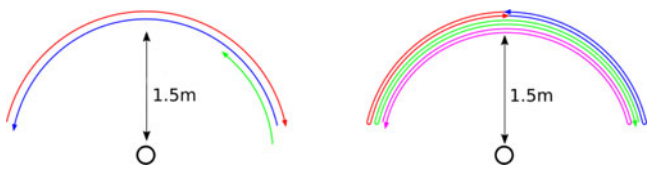

Fig. 2. Type I (left) and II (right) source trajectories for the experiments with semi-blind initialization. In Type I, Sources s_1 (red) and s_2 (blue) move from $-\vartheta$ to ϑ and from ϑ to $-\vartheta$, respectively, while Source s_3 moves from $85°$ to $45°$. In Type II, sources move: from $0°$ to $-\vartheta$ and back (s_1, red), from $0°$ to ϑ and back (s_2, blue), from $-\vartheta$ to ϑ and back (s_3, purple) and from ϑ to $-\vartheta$ and back (s_4, green); note that s_3 and s_4 move twice as fast as s_1 and s_2. In this example, $\vartheta = 75°$.

A. Experiments with Semi-Blind Initialization

1) Simulation Setup: The source signals were monochannel 16 kHz signals randomly taken from the TIMIT database [45]. Each source signal was convolved with a binaural room impulse responses (BRIRs) from [46] to produce the corresponding ground truth source image. The images of the 3 or 4 sources were added to provide the mix signal. The BRIRs were recorded with a dummy head equipped with 2 ear microphones, placed in a large lecture theatre of dimensions 23.5 m $\times$ 18.8 m $\times$ 4.6 m, and reverberation time $\mathrm{RT}_{60} \approx 0.68$ s [46]. We used a subset of (time-invariant) BRIRs with azimuthal source-to-head angle varying from $-90°$ to $90°$ with a $5°$ step. Continuous circular movements were simulated by interpolating the BRIRs at the sample level using up-sampling, delay compensation, linear interpolation, delay restoration, and downsampling. Due to memory limitations, we truncated the original 16000-tap BRIRs to either 512 or 4096 taps. Choosing two different lengths enables to evaluate the adequacy of the narrow-band assumption. Note that the recorded BRIRs almost vanish after 4096 samples, but not after 512 samples.

To assess the potential of the proposed algorithm to infer the time-varying frequency responses of the mixing filters, we devised two setups for the movement of the sources around the dummy head, drawn in Fig. 2. In Type I mixtures, Source s_3 always goes from $85°$ to $45°$. The amplitude of the trajectory of all other sources is varied with $\vartheta \in \{15°, 30°, 45°, 60°, 75°, 90°\}$. Each trajectory is covered at fixed speed, within the approximate 2s of signal duration (all signals are truncated to 32768 samples). We used four combinations of mixture type, filter tap length and number of sources, namely: *I-512-3*, *I-4096-3*, *II-512-3*,[6] and *II-512-4*.

The STFT was applied to the mixed signal with a 512-sample, 50%-overlap, sine window, leading to $L = 128$ observation frames. The number of components per source was set to $|\mathcal{K}_j| = 25$. The correct number of sources in the mixture (3 or 4) was provided to the algorithms in all experiments, along with the component-to-source partition $\mathcal{K}$. The number of iterations for all methods was fixed to 100.

2) Performance Measures: Two standard ASS objective measures were calculated between the estimated and ground truth source images, namely: signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR) [47].[7] In practice we used the `bss_eval_image` Matlab function dedicated to multichannel signals[8] [48]. Each reported measure is the average over 10 experiments with different source signals, and different NMF initializations (see below).

3) Baseline Method: The chosen baseline is a block-wise adaptation of the state-of-the-art method in [12]. We adapted the implementation provided by the authors,[9] following the line described in the introduction. We first segmented the sequence of $L = 128$ frames of the input mix into P blocks of $L_p = L/P$ consecutive frames, and applied the baseline method to each block independently (i.e. to each $I \times F \times L_p$ subarray of mixture coefficients). Hence for each block we obtain a subarray of the source image STFT coefficients estimates. Then by concatenating the successive subarrays and applying inverse STFT with overlap-add we obtain complete time-domain estimates of the source images. As mentioned in the introduction, the block size L_p must assume a good trade-off between local stationarity of mixing filters and a sufficient number of data to construct relevant statistics. The method in [12] was found to be very sensitive to the above constraint. For the simulations, we used $P = 4$ ($\Leftrightarrow L_p = 32$). This value showed better overall performance over the entire range of ϑ.

4) Initialization: The proposed VEM requires initializing $\{w_{fk}, h_{k\ell}, \hat{\mathbf{a}}_{:,f\ell}, \boldsymbol{\Sigma}^{\eta a}_{f\ell}, \boldsymbol{\Sigma}^{a}_{f}, \boldsymbol{\mu}^{a}_{f}, \mathrm{v}_f\}^{F,L,K}_{f,\ell,k=1}$. The baseline method requires initializing $\{w_{fk}, h_{k\ell}, \mathbf{A}^p_f, \mathrm{v}_f\}^{F,L,K}_{f,\ell,k=1}$. Note that all P blocks share the same w_{fk}, each block has its own set of $\mathbf{A}^p_f, \mathrm{v}_f$ and also a subset of $h_{k\ell}$ (though an additional block index is omitted for clarity).

NMF parameters: The initial values for the NMF parameters $\{w_{fk}, h_{k\ell}\}$, $k \in \mathcal{K}_j$ of a given source j are calculated by applying the KL-NMF algorithm [17] to the monochannel power spectrogram of source j, with random initialization. In order to assess the robustness of the proposed method to "realistic" initialization, KL-NMF is applied to a corrupted version of the source spectrogram. For this, the time-domain source signal $s_j(t)$ is first summed with all other interfering source signals with a controlled signal-to-noise ratio (SNR) R. We tested three different levels of corruption, namely $R \in \{20\,\mathrm{dB}, 10\,\mathrm{dB}, 0\,\mathrm{dB}\}$, with 0 dB meaning here equal power of signal $s_j(t)$ and of the sum of all interfering source signals. Note that $R = 20$ dB is a quite favorable initialization, whereas $R = 0$ dB tends towards more realism. This NMF initialization process is applied independently to all sources $j \in [1, J]$. The same resulting NMF initial parameters are used for both the proposed and baseline methods.

Mixing vectors: As for the initialization of $\hat{\mathbf{a}}_{:,f\ell}$, we used two different strategies. In the first one, for each source and each block of the baseline method, the time-interpolated BRIR corresponding to the center of the block was selected for the initialization of the corresponding column of $\mathbf{A}^p_f$ (after applying a 512-point FFT). For the proposed method, this initial $\mathbf{A}^p_f$ was

[6] In this case we discarded the fourth source (green plot in Fig. 2).

[7] We do not report and discuss signal-to-artefact ratio (SAR) measures in this subsection, due to space limitation.

[8] http://bass-db.gforge.inria.fr/bss_eval/.

[9] http://www.unice.fr/cfevotte/publications.html.

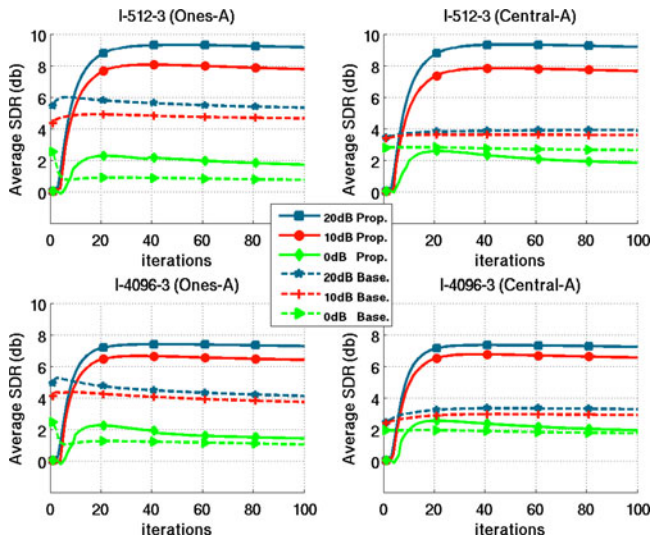

Fig. 3. Overall-sources average SDR vs iterations. For different initialization schemes: (top): *I-512-3*, (bottom): *I-4096-3*, (left) column is with *Ones-A* initialization, (right) is with *Central-A*. All experiments are at $\vartheta = 75°$.

replicated at each frame of the block, then vectorized, and set as initial $\hat{\mathbf{a}}_{:,f\ell}$. Applying this process to each block results in a complete initial sequence of L mixing vectors $\hat{\mathbf{a}}_{:,f\ell}$. In the following, we refer to this strategy as *Central-A*. The second strategy, called *Ones-A*, consists of setting all the entries of $\mathbf{A}_f^p$ and $\hat{\mathbf{a}}_{:,f\ell}$ to 1, $\forall f, \ell$. Obviously, this is a truly blind and challenging setup. Note that in all cases, both proposed and baseline algorithms were initialized with the same amount of filter information.

Other parameters: The remaining parameters were initialized as follows: $\mathbf{\Sigma}_{f\ell}^{\eta a} = 10^3 \mathbf{I}_{IJ}, \boldsymbol{\mu}_f^a = \hat{\mathbf{a}}_{:,f1}, \mathbf{\Sigma}_f^a = \mathbf{I}_{IJ}, \forall f, \ell$. As for the sensor noise variance v_f, the baseline method showed the best performance when initialized with 1% of the (L, I)-average PSD of the mixture, as suggested in [12]. Our method behaved best with a much higher initial value for v_f, namely 1000 times the (L, I)-average PSD of the mixture.

5) Results: We first discuss detailed results for a particular (but representative) value of ϑ, namely $\vartheta = 75°$. Then we report the performance of the proposed ASS algorithm w.r.t. the variation of ϑ and generalize the discussion.

Fig. 3 represents the evolution of average SDR measures with the (V)EM iterations, for $\vartheta = 75°$, and Mix-I. Let us recall that SDR is a general indicator that balances separation performance (i.e. interfering source rejection) and signal distortion (reconstruction artifacts). Each line is the result of averaging over the 3 sources, and over 10 different runs with different source signals. The two upper plots correspond to mix *I-512-3* and the two lower plots correspond to mix *I-4096-3*. The two left plots were initialized with the *Ones-A* strategy and the two right plots were initialized with *Central-A*.

In a general manner, the curves show that the baseline method converges faster than the proposed method, which is natural since the baseline method functions on blocks of STFT frames and the proposed method uses the complete sequence of STFT frames. Also, the baseline method has less parameters to estimate. In *I-512-3* (*Central-A*), the proposed method has an average performance of SDR ≈ 9.5 dB for $R = 20$ dB. The SDR score slightly degrades to about 8 dB for $R = 10$ dB, and then more abruptly decreases to about 2 dB for $R = 0$ dB. SDR scores of the baseline method at $R = 20$ dB, 10 dB, and 0 dB go from 4 to 2.5 dB. Therefore, the proposed VEM largely outperforms the baseline method for $R = 20$ dB and 10 dB, though in this example, the baseline performs slightly better at $R = 0$ dB ($\approx +0.5$ dB over the proposed method).

Regarding the influence of the initialization of the mixing vectors initialization, *Ones-A* vs. *Central-A*, the proposed algorithm proves to be remarkably robust to poor mixing filter initialization, since *Ones-A* provides similar results to *Central-A*. Hence, the proposed algorithm is able to correctly infer the mixing vectors from blind initialization, given that some reasonable amount of information on source PSD is provided (for instance by the NMF initialization). As for the baseline, its scores for $R = 20$ dB and 10 dB are again largely below the scores of the proposed method. However, and quite surprisingly, the baseline method behaves better (by about 0.4–0.7 dB) in the *Ones-A* (blind) configuration compared to the *Central-A* configuration, for $R = 20$ dB and 10 dB. This result is a bit difficult to interpret, but a possible explanation is that we measure the performance using the source images, rather than the monochannel source signals. Nevertheless for $R = 0$ dB, the filter information delivered by *Central-A* seems more useful, since the performance of the baseline method in the *Ones-A* configuration is about 2 dB lower than for *Central-A*. As a result, in the *Ones-A* configuration, the SDR scores of the proposed VEM are above the scores of the baseline method for all tested R values, including $R = 0$ dB.

As for the influence of the length of the BRIRs, we see that, unsurprisingly, the performance of both proposed and baseline algorithms decreases when the BRIRs go from 512-tap to 4096-tap responses. For $R = 20$ dB and 10 dB, we can observe that the decrease is of about 1.5–2 dB for the proposed method, independently of the mixing vectors initialization. The decrease is lower for the baseline method (≈ 1 dB), but this is probably related to the fact that the baseline scores are lower. For $R = 0$ dB, the influence of the BRIRs length on the performance of the proposed method is quite moderate, but this is also probably because the SDR scores are much lower than for $R = 20$ dB and 10 dB. All this manifests that (5) becomes a less appropriate model as the reverberation increases. Note that this is a recurrent problem in ASS in general. Our VEM is not intended to deal with this problem, but these experiments show that our VEM can provide quite remarkable SDR scores in a configuration that is *very* difficult in many aspects (underdetermined, time-varying, reverberant).

Table I provides results (at iteration 100) that are detailed per source (still averaged over 10 mixtures), and extended to SIR, for $\vartheta = 75°$ and *Ones-A* filter initialization. Output SIR scores focus on the ability of an ASS method to reject interfering sources. We first see from Table I that for $R = 20$ dB and $R = 10$ dB, the proposed VEM outperforms the baseline in both SDR and SIR for all configurations. In other words, the hierarchy discussed when analyzing Fig. 3 for $R = 20$ dB and $R = 10$ dB

TABLE I
AVERAGE SDR AND SIR MEASURES FOR $\vartheta = 75°$, *Ones-A*

		SDR								SIR							
		Proposed				Baseline				Proposed				Baseline			
R	Mixture	s_1	s_2	s_3	s_4	s_1	s_2	s_3	s_4	s_1	s_2	s_3	s_4	s_1	s_2	s_3	s_4
20 dB	I-512-3	**9.3**	**10.4**	**7.9**	–	5.5	6.5	4.0	–	**14.9**	**16.0**	**14.3**	–	10.5	12.3	8.4	–
	I-4096-3	**7.7**	**7.9**	**6.2**	–	4.7	4.6	3.0	–	**13.0**	**13.7**	**11.3**	–	10.0	9.9	6.6	–
	II-512-3	**8.4**	**8.2**	**9.5**	–	4.4	4.5	5.7	–	**13.6**	**13.8**	**16.1**	–	8.6	9.1	12.2	–
	II-512-4	**7.0**	**6.6**	**7.6**	**9.2**	3.8	3.9	4.9	5.8	**11.4**	**11.8**	**14.2**	**15.7**	7.4	8.7	9.8	11.3
10 dB	I-512-3	**7.9**	**9.1**	**6.3**	–	4.8	6.0	3.1	–	**12.8**	**13.6**	**12.9**	–	9.4	11.5	7.2	–
	I-4096-3	**6.9**	**7.1**	**5.2**	–	4.2	4.4	2.5	–	**11.4**	**11.7**	**9.7**	–	9.0	9.2	5.7	–
	II-512-3	**7.1**	**6.9**	**8.2**	–	3.8	4.0	5.3	–	**11.5**	**12.2**	**13.9**	–	7.5	8.5	11.3	–
	II-512-4	**6.1**	**6.0**	**6.9**	**8.2**	3.7	3.9	4.6	5.4	**10.4**	**10.6**	**12.8**	**13.7**	6.8	8.1	8.8	10.7
0 dB	I-512-3	**2.4**	**2.7**	**0.0**	–	1.1	2.3	−1.2	–	**4.3**	4.4	−0.4	–	3.7	**5.9**	**0.0**	–
	I-4096-3	**2.0**	1.9	**0.3**	–	1.8	**2.1**	-0.8	–	4.2	3.6	**−0.2**	–	**4.9**	**5.1**	-0.5	–
	II-512-3	**1.1**	**1.1**	**2.7**	–	0.0	0.4	1.7	–	**2.5**	2.1	3.9	–	2.0	**3.3**	**4.2**	–
	II-512-4	**1.8**	**1.7**	**3.4**	**3.8**	0.7	1.0	1.7	2.3	**4.2**	**3.6**	**5.3**	**5.8**	2.7	3.2	3.3	4.6

extends to per-source results, to Mix-II, and to SIR (at least for *Ones-A*). SDR improvement of the proposed method over the baseline ranges from 2.1 dB (s_2 in *II-512-4* at $R = 10$ dB) to 4.0 dB (s_1 in *II-512-3* at $R = 20$ dB). SIR improvement of the proposed method over the baseline ranges from 2.1 dB (s_2 in *I-512-3* at $R = 10$ dB) to an impressive 5.9 dB (s_3 in *I-512-3* at $R = 20$ dB). The results are particularly remarkable for the 4-source mixture configuration, with a range of output score similar to the 3-source configuration, and improvement over the baseline method up to 4.4 dB (s_3 and s_4 at $R = 20$ dB). At $R = 0$ dB the SIR results are more deteriorated for the 3-source configurations: they do not seem to indicate which method performs best (in terms of SIR). However, the SDR scores at 0 dB are all higher for the proposed method than for the baseline method, except for s_2 in mixture *I-4096-3* (only 0.2 dB below the baseline though). The improvement is however more limited than for $R = 20$ dB and $R = 10$ dB (maximum improvement is here 1.3 dB). Finally, at $R = 0$ dB, it can be noted that for the 4-source mixture, the proposed method outperforms the baseline method for all sources, and for both SDR (improvement ranges from 0.7 dB to 1.7 dB) and SIR (improvement ranges from 0.4 dB to 2 dB).

For a given source, the performance of ASS is more adequately described by the *separation gain*, i.e. the difference between output score and input score than by the output score only. Indeed, an input score quantifies how much the target source is corrupted in the input mixture. A source with low input score is more difficult to extract than a source with high input score. We thus display in Table II the input SDR and input SIR scores of each source.[10] Subtracting the scores in Table I and II, we can obtain SDR gains and SIR gains. We comment the results for $R = 0$ dB since it is the most realistic setting (remind that we also are in the *Ones-A* blind configuration for filters). For the 3-source mixtures, the proposed VEM algorithm provides

[10]We can see in this table that the length of BRIRs does not affect the input SIR, i.e. the entries *I-512-3* and *I-4096-3* are the same up to second decimal figure), when it slightly degrades the corresponding SDR scores.

TABLE II
INPUT SDR AND SIR FOR THE FOUR DIFFERENT MIXTURES

	SDR				SIR			
Mixture	s_1	s_2	s_3	s_4	s_1	s_2	s_3	s_4
I-512-3	−3.4	−1.2	−7.6	–	−2.0	−0.5	−5.9	–
I-4096-3	−2.6	−2.0	−7.5	–	−2.0	−0.5	−5.9	–
II-512-3	−5.3	−4.9	−2.1	–	−4.1	−3.7	−1.1	–
II-512-4	−7.8	−7.6	−5.3	−4.1	−6.3	−6.0	−4.1	−3.5

an SDR gain ranging from 3.9 dB to 7.8 dB, and an SIR gain ranging from 4.1 dB to 5.8 dB. As for the 4-source mixture, it is interesting to see that sources s_3 and s_4 score higher than s_1 and s_2 in Table I, although they move twice as fast as s_1 and s_2 and are thus expected to be more difficult to separate. However, they also have higher input scores, so that the separation gain turns out to be quite similar across sources.

We now focus on performance behavior w.r.t. the source velocity, i.e. different values of ϑ. Fig. 4 plots the gain of the proposed method over the baseline method, i.e. the (signed) difference of the proposed method's SDR and the SDR of the baseline. The results shown in Fig. 4 are at $R = 20$ dB, and *Ones-A* strategy (as the latter was shown to be most favorable for the baseline). For *II-512-3*, we observe that except for the 3 sources at $\vartheta = 30°$ and for s_2 at $\vartheta = 90°$, the gain is monotonically increasing for all three sources, starting from about 3 dB at $\vartheta = 15°$ and going up to 3.5–4.5 dB at $\vartheta = 90°$. Therefore, the advantage of the proposed method over the block-wise approach gets larger as the speed of moving sources increases. This makes sense since the block-wise baseline method rely on the assumption that filters are stationary on each block, and this assumption gets mangled as the source speed increases. In contrast, the proposed method seems robust to a large range of source velocity. This trend is also visible on the other plots. For example, for the *I-512-3* mixture, we see that the gain increases with ϑ for s_1 and s_2, from about 3 dB at $\vartheta = 15°$ to about 4 dB

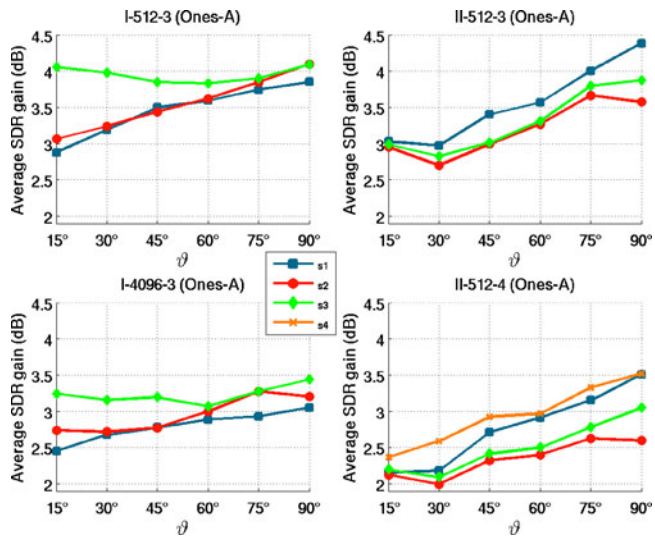

Fig. 4. Average SDR gain of the proposed method over the baseline method, for the four-source mixture, as a function of ϑ ($R = 20$ dB, *Ones-A* initialization).

at $\vartheta = 90°$, whereas the gain for s_3 (whose trajectory remains independent of ϑ) is almost constant at about 4 dB. The decreasing of this latter curve a bit around $\vartheta = 45°$ may be due to the trajectories of s_1 and s_2 interfering with the trajectory of s_3 for $\vartheta \geq 45°$. Additionally, the s_3 curve in configuration *I-512-3* shows that the advantage of the proposed method can be also large for relatively slow sources.

B. Experiments with Blind Initialization

In this section, we report the second series of experiments, that were conducted with blind initialization. This series of experiments consists of two parts: the first part deals with simulated 3-speaker mixtures, and the second part deals with a 2-speaker mixture made of real recordings. We first present the blind initialization method, that is common to all these new experiments, and then we detail the set-ups and results in the next subsections.

1) Blind Initialization: In these new experiments, the initialization of the proposed VEM algorithm (and of the baseline method) relies on the use of a state-of-the art blind source separation method based on source localization and binary masking. More specifically, we adapted the sound source localization method of [49], which is a good representative of recently proposed probabilistic methods based on mixture models of acoustic feature distribution parameterized by source position, see e.g. [6], [50]–[52]. The method in [49] relies on a mixture of complex Gaussian distributions (CGMM) that is used to compare the measured normalized relative transfer function (NRTF) at a pair of microphones with the expected NRTF as predicted by a source at a candidate position and a direct-path propagation model (there is one CGMM component for each candidate source position on a predefined grid). Combining the measures obtained at different microphone pairs into an EM algorithm enables to estimate the priors of the CGMM components. Then selecting the J first maxima of the priors amounts to localize the J sources. It also delivers the associated mixing vectors (corresponding to the direct path between sources and microphones). We adapted this method to the use of one pair of microphone, delivering J source direction estimates (in azimuth) and corresponding mixing vectors. We further combined it with a binary mask for source separation, inspired by [53]. For each TF bin, the masks are obtained by comparing the measured NRTF with the NRTF corresponding to the J candidate source directions; the source obtaining the largest posterior value in the CGMM among the J selected components has its mask set to 1 while the other sources have their mask set to 0. Then for each source, the mask is classically applied to the mixture STFT to obtain an estimate of the corresponding source image STFT. Importantly, to deal with our time-varying mixing set-up, this process is applied in a block-wise mode, similarly to what is done with the baseline method (see Section V-A3). Mixing vectors estimated on each block are replicated and catenated to form the initial $\hat{\mathbf{a}}_{:,f\ell}$ L-sequence. For each source j, the block-wise estimates of source image STFT vectors obtained by the binary masking are also concatenated, transformed to absolute squared values, averaged across channels, and supplied to the KL-NMF algorithm [17] to provide initial NMF parameter estimates for the complete sequence of L frames. This blind source separation method has been shown to be robust to short blocks, and therefore we can use here more blocks (of course shorter blocks) than in Section V-A3. This method was thus applied with 16 blocks (to process 2-second signals, with 50% overlap, hence one block is 250 ms long). Note that the baseline method that is plugged onto the initialization method is still run with $P = 4$ blocks. Note also that, as in Section V-A, the same information is used for the initialization of the proposed VEM and for the initialization of the baseline method.

2) Simulation Set-Up: The new simulation set-up is an underdetermined stereo setup of $J = 3$ simulated moving speakers (two male and one female from TIMIT). Since the blind initialization method relies on a free-field direct-path propagation model, we replaced the dummy head binaural recordings of Section V-A with the room impulse response simulator of AudioLabs Erlangen,[11] based on the image method [54]. We defined a 2-microphone set-up with omnidirectional microphones, spaced by $d = 50$ cm. The simulated room had the same size as the one in Section V-A1. In Section V-A1, we had simulated sources trajectoires that were crossing multiple times, to test the proposed method in a difficult scenario. However, the binary-mask initialization method is applied on blocks of time-frames, and it may be subject to source permutation across blocks.[12] To avoid this problem, we simulated a new setup where the trajectories of the $J = 3$ sources are not crossing each other: The 3 speech sources are all moving in circle of $\vartheta = 60°$ in 2 s, from $-65°$ to $-5°$ for s_1, from $-30°$ to $30°$ for s_2 and from $5°$ to $65°$ for s_3, at about 1.5 m of the microphone pair center (see Fig. 5 – left). We simulated two reverberation times, namely $T_{60} = 680$ ms

[11] Available at www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator.

[12] Note however that it is not subject to source permutation across frequency bins since all frequencies are jointly considered in the CGMM model, see [49] for details.

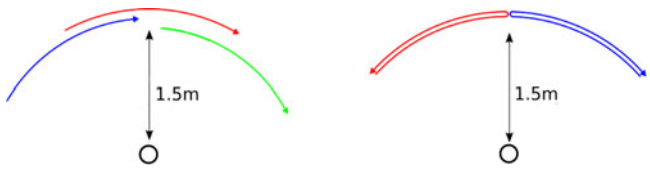

Fig. 5. Source trajectories for the experiments with blind initialization: Simulations (left) and real recordings (right).

(same as in Section V-A) and $T_{60} = 270$ ms (the corresponding mixtures are denoted respectively as *Mix-680* and *Mix-270*). We also tested the mixtures as is (noiseless case) and corrupted with additive white Gaussian noise at SNR$= 4$ dB. This resulted in 4 configurations. All reported measures are average results over 10 mixtures using different speech signals from TIMIT.

3) Real Recordings Set-Up: Real recordings were made in a 20 m^2 reverberant room ($T_{60} \approx 500$ ms), using $I = 2$ omnidirectional microphones in free field, placed in the center of the room, and spaced by $d = 30$ cm. For real recordings, the blind initialization method was shown to be much less efficient to separate 3 speakers, compared to the simulated experiments, but still worked very well for 2 speakers. We thus limited the present experiments with 2 speakers. Two speakers (one female, one male) were thus asked to pronounce spontaneous speech while moving on a circle at 1.5 m from the microphones, of about $45°$, two-way opposite motions, starting respectively at about $45°$ and $-45°$ (see Fig. 5 – right). The trajectory was traveled within 2 s, hence the speaker movement was pretty fast. The two speakers were recorded separately, and the signals were added, so that we could calculate separation scores.

4) Results of Simulations: Measures are reported in Table III for the input mixed signals, the initial source estimates after the binary masking, the estimates using the baseline method and the estimates using the proposed method. In addition to the SDR and SIR measures, we also report here SAR which measure the quantity of artefacts introduced on the separated signal by the separation method. Note that relatively homogeneous input SDR scores across sources (around -3 dB and -5 dB for the noiseless and noisy case respectively for both *Mix-270* and *Mix-680*) indicate that all sources have roughly the same power in the mix.

Let us start with the most reverberant condition *Mix-680*. At SNR $= \infty$, the average SDR (across sources) attained by the binary masking method is approximatively 3 dB, hence an SDR gain of about 6 dB over input signals. The corresponding average SIR gain is 7.8 dB, and the output average SAR is about 7 dB.[13] For this setting, the baseline method does not seem able to efficiently exploit the information provided by the blind initialization: The overall performance is comparable to the binary masking (SDR is even very slightly decreased for two sources). Regarding the proposed method, there is a significant improvement over both the binary mask initialization and the baseline method. In detail, the proposed method outperforms the baseline method by 0.5 dB to 1 dB SDR, by 0.5 dB to 1.9 dB SIR, and by 1.1 dB to 1.4 dB SAR (averaged across sources). With the addition of noise (SNR $= 4$ dB), all performance measures drop significantly, which was expected. For example, the average SDR for the binary masking is 2.3 dB lower than for the noiseless condition. Here, the baseline method slightly improves the binary masking scores, by 0.3 dB SDR, 0.1 dB SIR, and 1.5 dB SAR. More importantly, the proposed method outperforms the baseline method by 1.1 dB SDR, 0.9 dB SIR, and 3 dB SAR. Note that under noisy conditions, there is more margin for improvement over the binary masking since the latter provides worse estimates than in the noiseless case.

For *Mix-270*, i.e. moderate reverberations, we obtain significantly higher separation scores for all methods, as expected. For example, at SNR $= \infty$, the SDR for the binary masking (averaged across sources) is about 6 dB, hence an SDR gain of about 9 dB over input signals. Output SIR and SAR are within 9.2 dB to 10.8 dB (with an SIR gain going up to 13.8 dB). These scores (the SIR measures in particular) confirm what is well-known in the literature: Binary-masking techniques show good separation performance in low-to-moderate reverberant conditions. They place our block-wise binary masking method at the level of state-of-the-art methods based on the same principles (two-microphone source localization and binary masking), e.g. [6], [50]–[52], even though it is applied on quite short blocks (250 ms of mixture signal). Again, the baseline method exhibits comparable scores with the binary masking, here slightly better on the average. In addition, the proposed method significantly outperforms the baseline method, by 1.4 dB SDR, 2.2 dB SIR, and 1.8 dB SAR. The proposed method obtains SIR gains with respect to inputs as high as 16.4 dB (source s_2), which, we believe, is remarkable in a blind, underdetermined, dynamic setup, be it simulated. At SNR $= 4$ dB, we observe the same trends as for *Mix-680*: the baseline method improves more neatly over the binary masking, and the proposed method, again, significantly improves over the baseline method (by 1.7 dB SDR, 1.7 dB SIR, and 3.6 dB SAR).

5) Results of Real Recordings: The last three columns of Table III report the performance measures obtained on the real recordings with two sources. We first notice that even if we mix two sources instead of three, the gain performance of the binary masking method is less notable that in our simulated scenarios. This is evidence that separating (two) moving sources from real recordings remains quite a challenging scenario, even for state-of-the-art sound processing techniques. The baseline method shows some SDR improvement (≈ 0.5 dB) and SAR improvement (> 2 dB) for both sources over the binary masking. However, the baseline SIR scores degrade when compared to the binary-masking initialization. The proposed method exhibits positive gains when compared both with the binary-masking initialization and with the baseline method. Indeed, SAR scores of the proposed method are equivalent to the baseline method and notably better than the initialization. SDR improves by more than 1 dB when compared to the initialization, and by 0.7 dB to 0.9 dB when compared to the baseline method. SIR improves by 0.2 dB to 0.7 dB when compared to the initialization and by 0.7 dB to 1.1 dB when compared to the baseline method. Such

[13] It makes little sense to provide SAR gain, since, as source signals are intact in the mix, the input SAR is $= \infty$ and source separation can only lead to SAR decrease.

TABLE III
AVERAGE MEASURES USING BLIND INITIALIZATION, FOR SIMULATIONS AND REAL RECORDINGS (ALL UNITS ARE dB)

		Simulated Mix-270						Simulated Mix-680						real recordings		
SNR		∞			4			∞			4			N/A		
Method	Src	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
Input	s_1	-2.3	-1.9	$+\infty$	-4.5	-1.9	4.6	-3.5	-2.9	$+\infty$	-5.5	-2.9	4.6	0.0	0.2	$+\infty$
	s_2	-3.8	-3.0	$+\infty$	-5.7	-3.0	4.6	-2.7	-1.9	$+\infty$	-4.8	-2.0	4.6	0.0	0.2	$+\infty$
	s_3	-3.1	-2.5	$+\infty$	-5.1	-2.6	4.6	-3.3	-2.7	$+\infty$	-5.3	-2.7	4.6	–	–	–
Bin-Mask	s_1	6.2	10.5	9.5	2.5	7.5	3.4	2.8	5.2	6.1	0.5	2.6	1.7	2.9	7.6	6.3
	s_2	6.2	10.8	9.4	2.0	6.9	3.4	3.8	6.9	8.2	1.2	4.7	3.1	3.1	6.4	6.6
	s_3	5.9	9.9	9.2	1.9	6.0	3.0	2.6	3.8	6.8	0.7	2.7	2.7	–	–	–
Baseline	s_1	6.0	11.1	9.7	3.2	7.9	5.3	2.3	4.9	6.4	0.7	2.6	3.4	3.5	6.7	**8.3**
	s_2	6.7	11.1	10.0	2.9	7.7	5.0	3.8	7.1	8.5	1.6	4.9	4.4	3.6	6.1	9.1
	s_3	5.9	9.7	9.5	2.8	6.7	4.8	2.5	4.4	7.1	1.1	2.8	4.2	–	–	–
Proposed	s_1	**7.5**	**13.4**	**11.5**	**5.0**	**10.0**	**8.9**	**3.3**	**6.8**	**7.8**	**1.9**	**4.0**	**6.3**	**4.2**	**7.8**	**8.3**
	s_2	**7.8**	**13.4**	**11.7**	**4.4**	**9.4**	**8.5**	**4.4**	**8.3**	**9.6**	**2.6**	**5.7**	**7.4**	**4.5**	**7.1**	**9.2**
	s_3	**7.4**	**11.7**	**11.3**	**4.6**	**7.9**	**8.5**	**3.0**	**4.9**	**8.2**	**2.3**	**3.4**	**7.3**	–	–	–

results demonstrate the potential of the proposed approach for real-world applications and encourage us to pursue this line of research.

VI. CONCLUSION AND FUTURE WORK

In this paper we addressed the challenging task of separating audio sources from underdetermined time-varying convolutive mixtures. We started with the multichannel time-invariant convolutive LGM-NMF framework of [12], and we introduced time-varying filters modeled by a first-order Markov model with complex Gaussian observation and transition distributions. Because the mixture observations do not depend only on the filters, but also on the sources that are latent variables as well, a standard direct application of a Kalman smoother is not possible. We addressed this issue with a variational approximation, assuming that the filters and the sources are conditionally independent with respect to the mixture. This lead to a closed-form VEM, including a variational version of the Kalman smoother, and finally, separating Wiener filters that are constructed from both time-varying estimated source parameters and time-varying estimated mixing filters. Several implementation issues were discussed to facilitate experimental reproducibility. Finally, an extensive evaluation campaign demonstrated the experimental advantage of the proposed approach over a state-of-the-art baseline method in several speech mixtures under different initialization strategies.

These results encourage for further research to improve the proposed model. Firstly, the last series of reported experiments show that the use of realistic blind separation methods for the initialization of our algorithm in the case of more sources than microphones has to be more deeply explored and made more robust to process real recordings. Secondly, in the present study, the number of sources present in the mixture was assumed to be known, although the estimation of this number is a problem on its own. Therefore, developing algorithms capable of estimating the number of active (i.e. emitting) sources varying over time remains an open issue, but is a step closer to realistic applications. We therefore plan to incorporate into the present model the estimation of the sources activity, using diarization latent variables. Finally, an in-depth study exploring the complex relationship between the physical changes of the recording set-up and the mixing filters can be of great help. In particular, a better understanding of how the position of the sources and microphones affect the filters may enable us to incorporate the rationale of the discrete DOA-dependent model in [32] to the proposed continuous latent model, thus using localization cues to help the automatic separation of sound sources.

REFERENCES

[1] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation – Independent Component Analysis and Applications*. New York, NY, USA: Academic, 2010.

[2] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time Fourier transform domain," *IEEE Signal Process. Lett.*, vol. 14, no. 5, pp. 337–340, May 2007.

[3] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, "Probabilistic modeling paradigms for audio source separation," in *Machine Audition: Principles, Algorithms and Systems.* Hershey, PA, USA: Idea Group Inc., pp. 162–185, 2010.

[4] A. Hyvärinen, J. Karhunen, and E. Oja, Eds.,*Independent Component Analysis*. New York, NY, USA: Wiley, 2001.

[5] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "MAP-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and l1-norm minimization," *EURASIP J. Adv. Signal Process.*, vol. 2007, 2006, Art. no. 24717.

[6] M. Mandel, R. J. Weiss, D. P. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 382–394, Feb. 2010.

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

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 6, pp. 443–445, Dec. 1984.

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

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

[11] C. Févotte and J.-F. Cardoso, "Maximum likelihood approach for blind audio source separation using time-frequency Gaussian source models," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, USA, 2005, pp. 78–81.

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

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

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

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

[16] D. Lee and H. Seung, "Algorithms for non-negative matrix factorization," *Adv. Neural Inf. Process. Syst.*, vol. 13, pp. 556–562, 2001.

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

[18] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 69–84, Jan. 2011.

[19] J. Anemüller and T. Gramss, "On-line blind separation of moving sound sources," in *Proc. Int. Conf. Independent Compon. Anal. Blind Source Separation*, Aussois, France, 1999.

[20] A. Koutras, E. Dermatas, and G. Kokkinakis, "Blind speech separation of moving speakers in real reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Istanbul, Turkey, 2000, pp. II1133–II1136.

[21] K. E. Hild II, D. Erdogmus, and J. C. Principe, "Blind source separation of time-varying, instantaneous mixtures using an on-line algorithm," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Orlando, FL, USA, 2002, pp. I-993–I-996.

[22] R. Aichner, H. Buchner, S. Araki, and S. Makino, "On-line time-domain blind source separation of nonstationary convolved signals," in *Proc. Int. Conf. Independent Compon. Anal. Blind Source Separation*, Nara, Japan, 2003, pp. 987–992.

[23] R. E. Prieto and P. Jinachitra, "Blind source separation for time-variant mixing systems using piecewise linear approximations," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Philadelphia, PA, USA, 2005, pp. v/301–v/304.

[24] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Robust real-time blind source separation for moving speakers in a room," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2003, pp. 469–72.

[25] W. Addison and S. Roberts, "Blind source separation with non-stationary mixing using wavelets," in *Proc. Int. Conf. Independent Compon. Anal. Blind Source Separation*, Charleston, SC, USA, 2006.

[26] K. Nakadai, H. Nakajima, Y. Hasegawa, and H. Tsujino, "Sound source separation of moving speakers for robot audition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Taipei, Taiwan, 2009, pp. 3685–3688.

[27] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Process.*, vol. 87, no. 8, pp. 1833–1847, 2007.

[28] B. Loesch and B. Yang, "Online blind source separation based on time-frequency sparseness," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Taipei, Taiwan, 2009, pp. 117–120.

[29] L. Simon and E. Vincent, "A general framework for online audio source separation," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, Tel-Aviv, Israel, 2012, pp. 397–404

[30] S. Markovich-Golan, S. Gannot, and I. Cohen, "Subspace tracking of multiple sources and its application to speakers extraction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Dallas, TX, USA, 2010, pp. 201–204.

[31] E. Weinstein, A. Oppenheim, M. Feder, and J. Buck, "Iterative and sequential algorithms for multisensor signal enhancement," *IEEE Trans. Signal Process.*, vol. 42, no. 4, pp. 846–859, Apr. 1994.

[32] T. Higuchi, N. Takamune, N. Tomohiko, and H. Kameoka, "Underdetermined blind separation and tracking of moving sources based on DOA-HMM," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Florence, Italy, 2014, pp. 3191–3195.

[33] C. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[34] S. Gannot and M. Moonen, "On the application of the unscented Kalman filter to speech processing," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Kyoto, Japan, 2003, pp. 27–30.

[35] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, "A variational EM algorithm for the separation of moving sound sources," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, October 2015, pp. 1–5.

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

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

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

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

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

[41] G. McLachlan and K. Thriyambakam, *The EM Algorithm and Extensions*. New York, NY, USA: Wiley, 1997.

[42] V. Smidl and A. Quinn, *The Variational Bayes Method in Signal Processing*. Berlin, Germany: Springer-Verlag, 2006.

[43] A. Hjorungnes and D. Gesbert, "Complex-valued matrix differentiation: Techniques and key results," *IEEE Trans. Signal Process.*, vol. 55, no. 6, pp. 2740–2746, Jun. 2007.

[44] 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," presented at the Conv. Audio Engineering Society, Budapest, Hungary, 2012.

[45] 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*. Philadelphia, PA, USA: Linguistic Data Consortium, 1993.

[46] C. Hummersone, R. Mason, and T. Brookes, "A comparison of computational precedence models for source separation in reverberant environments," *J. Audio Eng. Soc.*, vol. 61, no. 7/8, pp. 508–520, 2013.

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

[48] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. Rosca, "First stereo audio source separation evaluation campaign: Data, algorithms and results," in *Proc. Int. Conf. Independent Compon. Anal. Signal Separation*, London, U.K., 2007, pp. 552–559.

[49] Y. Dorfan and S. Gannot, "Tree-based recursive expectation-maximization algorithm for localization of acoustic sources," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 10, pp. 1692–1703, Oct. 2015.

[50] T. May, S. Van De Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 1–13, Jan. 2011.

[51] J. Woodruff and D. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1503–1512, Jul. 2012.

[52] J. Traa and P. Smaragdis, "Multichannel source separation and tracking with RANSAC and directional statistics," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 2233–2243, Dec. 2014.

[53] Y. Dorfan, D. Cherkassky, and S. Gannot, "Speaker localization and separation using incremental distributed expectation-maximization," in *Proc. Eur. Signal Process. Conf.*, Nice, France, 2015, pp. 1256–1260.

[54] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

Dionyssos Kounades-Bastian received the M.Sc. degree in computer science from the Engineering School, University of Patras, Patras, Greece, in 2013. He is currently working toward the Ph.D. degree in the Perception Team at INRIA (French Computer Science Research Institute). His research interests include machine learning and signal processing for audio scene analysis.

Laurent Girin received the M.Sc. and Ph.D. degrees in signal processing from the Institut National Polytechnique de Grenoble, Grenoble, France, in 1994 and 1997, respectively. In 1999, he joined the Ecole Nationale Supérieure d'Electronique et de Radioélectricité de Grenoble, as an Associate Professor. He is currently a Professor at Phelma (Physics, Electronics, and Materials Department), Grenoble-INP, where he lectures signal processing theory and applications to audio. His research activity is carried out at GIPSA-Lab (Grenoble Laboratory of Image, Speech, Signal, and Automation). He is also a regular collaborator of INRIA (French Computer Science Research Institute), as an Associate Member of the Perception Team. His research interests include speech and audio processing (analysis, modeling, coding, transformation, synthesis), with a special interest in multimodal speech processing (e.g., audiovisual, articulatory-acoustic, etc.) and speech/audio source separation.

Xavier Alameda-Pineda (M'16) received the M.Sc. degrees in mathematics, and in telecommunications from Barcelona Tech, Barcelona, Spain; and in computer science from Grenoble-INP, Grenoble, France. He did his Ph.D. work in the Perception Team at INRIA (French Computer Science Research Institute), Grenoble, France, and received the Ph.D. degree in mathematics and computer science from Université Joseph Fourier, Saint-Martin-d'Hères, France, in 2013. He is currently a Postdoctoral Fellow at the University of Trento. His research interests include multimodal machine learning and signal processing for scene analysis.

Sharon Gannot (S'92–M'01–SM'06) received the B.Sc. degree (*summa cum laude)* from the Technion Israel Institute of Technology, Haifa, Israel, in 1986 and the M.Sc. (*cum laude*) and Ph.D. degrees from Tel-Aviv University, Tel Aviv-Yafo, Israel, in 1995 and 2000, respectively, all in electrical engineering. In 2001, he held a Postdoctoral position at the Department of Electrical Engineering (ESAT-SISTA) at K. U. Leuven, Belgium. From 2002 to 2003, he held a research and teaching position at the Faculty of Electrical Engineering, Technion-Israel Institute of Technology, Haifa, Israel. He is currently a Full Professor at the Faculty of Engineering, Bar-Ilan University, Israel, where he is heading the Speech and Signal Processing laboratory and the Signal Processing Track. His research interests include multi-microphone speech processing and specifically distributed algorithms for ad hoc microphone arrays for noise reduction and speaker separation; dereverberation; single microphone speech enhancement and speaker localization and tracking. He received the Bar-Ilan University Outstanding Lecturer Award for 2010 and 2014. He has served as an Associate Editor of the EURASIP *Journal of Advances in Signal Processing* during 2003–2012, and as an Editor of several special issues on Multi-Microphone Speech Processing of the same journal. He has also served as a Guest Editor of *ELSEVIER Speech Communication and Signal Processing Journals.* He has served as an Associate Editor of IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING during 2009–2013. He is currently a Senior Area Chair of the same journal. He also serves as a Reviewer of many IEEE journals and conferences. He is a Member of the Audio and Acoustic Signal Processing technical committee of the IEEE since January, 2010. He is currently the committee Vice-Chair. He is also a Member of the Technical and Steering Committee of the International Workshop on Acoustic Signal Enhancement (IWAENC) since 2005 and was the General Co-Chair of IWAENC held at Tel-Aviv, Israel, in August 2010. He has served as the General Co-Chair of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics in October 2013. He was selected (with colleagues) to present a tutorial sessions in ICASSP 2012, EUSIPCO 2012, ICASSP 2013, and EUSIPCO 2013.

Radu Horaud received the B.Sc. degree in electrical engineering, the M.Sc. degree in control engineering, and the Ph.D. degree in computer science from the Institut National Polytechnique de Grenoble, Grenoble, France. He is currently the Director of Research with INRIA Grenoble, Montbonnot Saint-Martin, France, where he is the Founder and Head of the Perception team. His research interests include computer vision, machine learning, audio signal processing, audiovisual analysis, and robotics. He is an Area Editor of the *Elsevier Computer Vision and Image Understanding*, a Member of the advisory board of the *Sage International Journal of Robotics Research*, and an Associate Editor of the *Kluwer International Journal of Computer Vision*. In 2013, he received an ERC Advanced Grant for his project *Vision and Hearing in Action* .