

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

Dionyssos Kounades-Bastian, Laurent Girin, Xavier Alameda-Pineda, Sharon Gannot, 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 block-wise version of a state-of-the-art baseline method.

Index Terms—Audio-source separation, time-varying mixing filters, moving sources, Kalman smoother, 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 (SCA) 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], [12], [13], [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 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 EM algorithm.

D. Kounades-Bastian is with INRIA Grenoble Rhône-Alpes, France. E-mail: dionyssos.kounades-bastian@inria.fr

L. Girin is with INRIA Grenoble Rhône-Alpes, France, and with Univ. Grenoble Alpes, GIPSA-lab, Grenoble, France. E-mail: laurent.girin@gipsa-lab.grenoble-inp.fr

X. Alameda-Pineda is with University of Trento, Italy. E-mail: xavier.alamedapineda@unitn.it

S. Gannot is with Bar Ilan University, Faculty of Engineering, Israel. E-mail: Sharon.Gannot@biu.ac.il

R. Horaud is with INRIA Grenoble Rhône-Alpes, France. E-mail: radu.horaud@inria.fr

D. Kounades-Bastian, L. Girin and R. Horaud acknowledge support from the European FP7 STREP project EARS #609465 and from the European Research Council through the ERC Advanced Grant VHIA #340113.

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.

Interesting enough, most of the block-wise approaches use ICA, either in the temporal domain [19] (limited to anechoic setups), [20], [21], [22], [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 sound 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 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 well 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 propose here. Therefore, we propose an approximate solution for the joint estimation of the model parameters and inference of the latent variables. We derive a variational EM (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 temporal frames of the input mixtures. To summarize, the proposed method makes use of 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 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.

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 using simulated data 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 time-frequency representation, after applying the short-time Fourier transform (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) .

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 $c_{k,f\ell}$, $k \in \mathcal{K}_j$, i.e.:

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

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:¹

$$p(\mathbf{c}_{f\ell}) = \mathcal{N}_c\left(\mathbf{c}_{f\ell}; \mathbf{0}, \text{diag}_K(w_{fk} h_{k\ell})\right), \quad (2)$$

¹The proper complex Gaussian distribution is defined as $\mathcal{N}_c(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = |\pi \boldsymbol{\Sigma}|^{-1} \exp(-[\mathbf{x} - \boldsymbol{\mu}]^H \boldsymbol{\Sigma}^{-1} [\mathbf{x} - \boldsymbol{\mu}])$, 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].

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

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

Eq. (??) corresponds to the modeling of the $F \times L$ source PSD matrix with the NMF model, which is widely used in audio analysis, audio source separation, and speech enhancement [9], [37], [17], [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 (??).

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:

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

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 (??) 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. $p(\mathbf{b}_{f\ell}) = \mathcal{N}_c(\mathbf{b}_{f\ell}; \mathbf{0}, v_f \mathbf{I}_I)$, with $v_f \in \mathbb{R}^+$ 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 $p(\mathbf{x}_{f\ell} | \mathbf{A}_{f\ell}, \mathbf{s}_{f\ell}) = \mathcal{N}_c(\mathbf{x}_{f\ell}; \mathbf{A}_{f\ell} \mathbf{s}_{f\ell}, v_f \mathbf{I}_I)$.

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

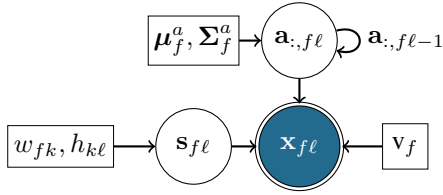


Fig. 1. Graphical model for time-varying convolutive mixtures with NMF source model. Latent variables are represented with circles, observations with double circles, deterministic parameters with rectangles, and temporal dependencies with self loops.

$\mathbf{a}_{:,f\ell} \in \mathbb{C}^{IJ}$, i.e. $\mathbf{a}_{:,f\ell} = \text{vec}(\mathbf{A}_{f\ell}) = [\mathbf{a}_{1,f\ell}^\top \dots \mathbf{a}_{J,f\ell}^\top]^\top$. 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:

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

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

where the mean vector $\boldsymbol{\mu}_f^a \in \mathbb{C}^{IJ}$ and the *evolution* covariance matrix $\Sigma_f^a \in \mathbb{C}^{IJ \times IJ}$ are parameters to be estimated. Σ_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 $\Sigma_f^a \rightarrow 0\mathbf{I}_{IJ}$. 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 audio source separation 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. (??) defines the *observation model* of the Kalman smoother.² 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.

III. VEM FOR SOURCE SEPARATION

In this section, we present the proposed variational EM 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:

²The vectorized form of the latent mixing filters can be made explicit in the observation model by rewriting it as $\mathbf{x}_{f\ell} = (\mathbf{s}_{f\ell}^\top \otimes \mathbf{I}_I) \mathbf{a}_{:,f\ell} + \mathbf{b}_{f\ell}$, with \otimes denoting the Kronecker matrix product.

\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} , $\Sigma^{\eta z} = \mathbb{E}_{q(\mathbf{z})}[(\mathbf{z} - \hat{\mathbf{z}})(\mathbf{z} - \hat{\mathbf{z}})^\text{H}]$ is its posterior covariance matrix, and $\mathbf{Q}^{\eta z} = \mathbb{E}_{q(\mathbf{z})}[\mathbf{z}\mathbf{z}^\text{H}] = \Sigma^{\eta z} + \hat{\mathbf{z}}\hat{\mathbf{z}}^\text{H}$ 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 $\Sigma_{f\ell}$. Let $\stackrel{ct}{=}$ 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 [41], [33]. 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 $\theta = \{\boldsymbol{\mu}_f^a, \Sigma_f^a, w_{fk}, h_{k\ell}, v_f\}_{f,\ell,k=1}^{F,L,K}$ 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 joint, possibly also factorized, 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 naturally factorizes as:

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

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

B. E-A Step

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

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

We have:

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

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

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

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

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

Functional $\mathcal{N}_c(\boldsymbol{\mu}_{f\ell}^{ia}; \mathbf{a}_{:,f\ell}, \boldsymbol{\Sigma}_{f\ell}^{ia})$ can be viewed as an *instantaneous* distribution of a *measured* vector $\boldsymbol{\mu}_{f\ell}^{ia}$, conditioned to the *hidden* variable $\mathbf{a}_{:,f\ell}$. Henceforth one recognizes that (??) represents an LDS with continuous hidden state variables $\{\mathbf{a}_{:,f\ell}\}_{\ell=1}^L$, transition distribution given by (??), initial distribution given by (??), and *emission* distribution given by $\mathcal{N}_c(\boldsymbol{\mu}_{f\ell}^{ia}; \mathbf{a}_{:,f\ell}, \boldsymbol{\Sigma}_{f\ell}^{ia})$. 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 (??) is described below. Because of the form of (??), 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 $\boldsymbol{\mu}_{f\ell}^{\phi a} \in \mathbb{C}^{IJ}$ and covariance matrix $\boldsymbol{\Sigma}_{f\ell}^{\phi a} \in \mathbb{C}^{IJ \times IJ}$ of this distribution are calculated as:

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

$$\boldsymbol{\mu}_{f\ell}^{\phi a} = \boldsymbol{\Sigma}_{f\ell}^{\phi a} \left(\boldsymbol{\Sigma}_{f\ell}^{ia-1} \boldsymbol{\mu}_{f\ell}^{ia} + (\boldsymbol{\Sigma}_{f\ell-1}^{\phi a} + \boldsymbol{\Sigma}_f^a)^{-1} \boldsymbol{\mu}_{f\ell-1}^{\phi a} \right). \quad (14)$$

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

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

$$\boldsymbol{\Sigma}_{f\ell}^{\beta a} = \boldsymbol{\Sigma}_f^a + \boldsymbol{\Sigma}_{f\ell}^{\zeta a}, \quad (16)$$

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

where $\boldsymbol{\Sigma}_{f\ell}^{\zeta a} \in \mathbb{C}^{IJ \times IJ}$ 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]:

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

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

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

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

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

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

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

Note here the role of $\boldsymbol{\Sigma}_{f\ell}^{\zeta a}$ that is to describe the uncertainty of $\boldsymbol{\mu}_{f\ell+1}^{\beta a}$ but without incorporating the additional uncertainty of the transition variance $\boldsymbol{\Sigma}_f^a$, 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 (??), the posterior distribution of the sources writes:

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

Using (??), the expectation in (??) computes:

$$\mathbb{E}_{q(\mathbf{a}_{\cdot, f\ell})} [\log p(\mathbf{x}_{f\ell} | \mathbf{A}_{f\ell}, \mathbf{s}_{f\ell})] \stackrel{ct}{=} \frac{1}{v_f} \text{tr} \left\{ \mathbf{s}_{f\ell} (\hat{\mathbf{A}}_{f\ell}^H \mathbf{x}_{f\ell})^H + (\hat{\mathbf{A}}_{f\ell}^H \mathbf{x}_{f\ell}) \mathbf{s}_{f\ell}^H - \mathbf{U}_{f\ell}^\top \mathbf{s}_{f\ell} \mathbf{s}_{f\ell}^H \right\}, \quad (24)$$

where $\hat{\mathbf{A}}_{f\ell} = \mathbb{E}_{q(\mathbf{a}_{\cdot, f\ell})} [\mathbf{A}_{f\ell}] \in \mathbb{C}^{I \times J}$ is a matrix constructed from $\hat{\mathbf{a}}_{\cdot, f\ell}$ (i.e. the reverse operation of column-wise vectorization), and $\mathbf{U}_{f\ell} \in \mathbb{C}^{J \times J}$ is defined by $\mathbf{U}_{f\ell} = \mathbb{E}_{q(\mathbf{a}_{\cdot, f\ell})} [\mathbf{A}_{f\ell}^H \mathbf{A}_{f\ell}]^\top$. Of course, $\mathbf{U}_{f\ell}$ is closely related to $\mathbf{Q}_{f\ell}^{\eta a}$. If we segment $\mathbf{Q}_{f\ell}^{\eta a}$ into J^2 non-overlapping $I \times I$ subblocks $\mathbf{Q}_{pq, f\ell}^{\eta a}$, we have:

$$\mathbf{U}_{f\ell} = \begin{bmatrix} \text{tr}\{\mathbf{Q}_{11, f\ell}^{\eta a}\} & \dots & \text{tr}\{\mathbf{Q}_{1J, f\ell}^{\eta a}\} \\ \vdots & & \vdots \\ \text{tr}\{\mathbf{Q}_{J1, f\ell}^{\eta a}\} & \dots & \text{tr}\{\mathbf{Q}_{JJ, f\ell}^{\eta a}\} \end{bmatrix}. \quad (25)$$

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

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

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

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

$$\hat{\mathbf{s}}_{f\ell} = \boldsymbol{\Sigma}_{f\ell}^{\eta s} \hat{\mathbf{A}}_{f\ell}^H \frac{\mathbf{x}_{f\ell}}{v_f}. \quad (28)$$

Remarkably, (??) 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 $(\boldsymbol{\Sigma}_{f\ell}^{\eta s})^{-1}$ is built by summation of (i) the sensor precision $1/v_f$ distributed over the sources with the unit-less quantity $\mathbf{U}_{f\ell}^\top$, 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 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:

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

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

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

$$\hat{\mathbf{c}}_{f\ell} = \boldsymbol{\Sigma}_{f\ell}^{\eta c} \mathbf{G}^\top \hat{\mathbf{A}}_{f\ell}^H \frac{\mathbf{x}_{f\ell}}{v_f}. \quad (31)$$

Again, (??) is a Wiener filtering estimator, here at the source

component vector level. Note that left-multiplication of both sides of (??) by \mathbf{G} naturally leads to (??).

D. Outline of the Maximization Step

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

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

Notice that (??) 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:

$$\begin{aligned} v_f &= \frac{1}{LI} \sum_{\ell=1}^L \left(\mathbf{x}_{f\ell}^H \mathbf{x}_{f\ell} - \mathbf{x}_{f\ell}^H \hat{\mathbf{A}}_{f\ell} \hat{\mathbf{s}}_{f\ell} \right. \\ &\left. - (\hat{\mathbf{A}}_{f\ell} \hat{\mathbf{s}}_{f\ell})^H \mathbf{x}_{f\ell} + \text{tr}\{\mathbf{U}_{f\ell}^\top \mathbf{Q}_{f\ell}^{\eta s}\} \right), \end{aligned} \quad (33)$$

which resembles the estimator obtained in [12].

F. M-A Step

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

$$\boldsymbol{\mu}_f^a = \hat{\mathbf{a}}_{f1}. \quad (34)$$

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

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

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

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.:

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

Derivating (??) w.r.t. the entries of Σ_f^a , and setting the result to zero, yields [43]:

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

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:

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

that can be iteratively applied until convergence.

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 $\{\mathbb{E}_{q(\mathbf{a}_{:,f\ell}, \mathbf{s}_{f\ell})}[\mathbf{a}_{j,f\ell} \mathbf{s}_{j,f\ell}] = \hat{\mathbf{a}}_{j,f\ell} \hat{\mathbf{s}}_{j,f\ell}\}_{f,\ell=1}^{F,L}$, 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).

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.³

1) *Simplifying the LDS measurement vector*: The forward-backward procedure requires the quantity $(\Sigma_{f\ell}^{\iota a -1} \boldsymbol{\mu}_{f\ell}^{\iota a})$, appearing in (??) and (??). This can be computed as:

$$\Sigma_{f\ell}^{\iota a -1} \boldsymbol{\mu}_{f\ell}^{\iota a} = \frac{1}{v_f} \text{vec}(\mathbf{x}_{f\ell} \hat{\mathbf{S}}_{f\ell}^H), \quad (39)$$

thus sparing the inversion of $\Sigma_{f\ell}^{\iota a}$.

³If the paper is accepted for publication, the Matlab code of the proposed algorithm will be released publicly.

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 $\hat{\mathbf{a}}_{:,f\ell}$, $\Sigma_{f\ell}^{\eta a}$.
repeat
 Variational E-step
 Calculate $\mathbf{Q}_{f\ell}^{\eta a} = \Sigma_{f\ell}^{\eta a} + \hat{\mathbf{a}}_{:,f\ell} \hat{\mathbf{a}}_{:,f\ell}^H$ and $\mathbf{U}_{f\ell}$ with (??).
 E-S step: Compute $\Sigma_{f\ell}^{\eta s}$ with (??) and $\hat{\mathbf{s}}_{f\ell}$ with (??).
 Then compute $\mathbf{Q}_{f\ell}^{\eta s} = \Sigma_{f\ell}^{\eta s} + \hat{\mathbf{s}}_{f\ell} \hat{\mathbf{s}}_{f\ell}^H$.
 E-C step: Compute $\Sigma_{kk,f\ell}^{\eta c}$ with (??) and $\hat{c}_{k,f\ell}$ with (??).
 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 $(\Sigma_{f\ell}^{\iota a -1} \boldsymbol{\mu}_{f\ell}^{\iota a})$ with (??).
 Compute $\Sigma_{f\ell}^{\iota a -1} = \mathbf{Q}_{f\ell}^{\eta s \top} \otimes \mathbf{I}_{Iv_f^{-1}}$.
 E-A step (Forward Pass):
 Initialize $\Sigma_{f1}^{\phi a} = (\Sigma_{f1}^{\iota a -1} + \Sigma_f^{\eta a -1})^{-1}$.
 Initialize $\boldsymbol{\mu}_{f1}^{\phi a} = \Sigma_{f1}^{\phi a} (\Sigma_{f1}^{\iota a -1} \boldsymbol{\mu}_{f1}^{\iota a} + \Sigma_f^{\eta a -1} \boldsymbol{\mu}_f^{\eta a})$.
 for $\ell : 2$ to L
 Compute $\Sigma_{f\ell}^{\phi a}$ with (??), then $\boldsymbol{\mu}_{f\ell}^{\phi a}$ with (??).
 end
 E-A step (Backward Pass):
 Initialize $\Sigma_{fL}^{\beta a} = \Sigma_{fL}^{\phi a}$ and $\boldsymbol{\mu}_{fL}^{\beta a} = \boldsymbol{\mu}_{fL}^{\phi a}$.
 for $\ell : L - 1$ to 1
 Compute $\Sigma_{f\ell}^{\zeta a}$ by (??), then $\Sigma_{f\ell}^{\beta a}$ by (??),
 and $\boldsymbol{\mu}_{f\ell}^{\beta a}$ by (??).
 end
 E-A step (Posterior Marginal Statistics):
 Compute $\Sigma_{f\ell}^{\eta a}$ with (??).
 Then compute $\hat{\mathbf{a}}_{:,f\ell}$ with (??).
 E-A step (Pairwise Joint Posterior):
 Compute $\Sigma_{f\ell}^{\xi a}$ with (??).
 Then compute $\boldsymbol{\mu}_{f\ell}^{\xi a}$ with (??).
 Then compute $\mathbf{Q}_{f\ell}^{\xi a}$ with (??).
M-step
 M-v step: Update v_f with (??).
 M-A step: Update $\boldsymbol{\mu}_f^a$ with (??), update Σ_f^a with (??).
 M-C step: Alternately update w_{fk} and $h_{k\ell}$ with (??).
until convergence
return the estimated source images $\hat{\mathbf{a}}_{j,f\ell} \hat{\mathbf{s}}_{j,f\ell}$, $j \in [1, J]$.

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

3) *Avoiding $K \times K$ matrix construction*: Eq. (??) 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 $\Sigma_{f\ell}^{\eta^s}$ and $\hat{\mathbf{S}}_{f\ell}$ (which use operations only on $J \times J$ arrays). Applying the Woodbury identity to (??) and some algebraic manipulations, one obtains:

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

where j_k is the index of the source that the k^{th} 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 $\Sigma_{f\ell}^{\eta^c}$:

$$\hat{c}_{k, f\ell} = w_{fk} h_{k\ell} \left[\hat{\mathbf{A}}_{f\ell}^H \frac{\mathbf{x}_{f\ell}}{\nu_f} - \mathbf{U}_{f\ell}^\top \frac{\hat{\mathbf{S}}_{f\ell}}{\nu_f} \right]_{j_k}, \quad (41)$$

where $[\cdot]_{j_k}$ is the j_k^{th} element of the $J \times 1$ vector in brackets. Interestingly, (??) 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, (??) makes clear that if the value of ν_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 ν_f in cases where the NMF parameters are quite correctly initialized.

4) *Ordering the steps*: When building a (V)EM algorithm, the question of ordering the steps execution arises. Likewise the majority of EMs, our algorithm is sensitive to initialization (discussed in Section V-C). 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: $\Sigma_{f\ell}^{\eta^a}, \hat{\mathbf{a}}_{:, f\ell}$. 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.

5) *NMF scaling*: When estimating the NMF parameters using (??), 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. $w_{fk} \leftarrow w_{fk} / \sum_f w_{fk}$, 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.

6) *Loss of Hermitian properties*: We enforce matrices $\mathbf{U}_{f\ell}$ and Σ_f^a to be Hermitian with $\Sigma_f^a \leftarrow \frac{1}{2} (\Sigma_f^a + \Sigma_f^{aH})$. Both these matrices are theoretically Hermitian, but this has to be assured numerically as well.

7) *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 $\mathcal{O}(16FL(IJ)^3 + 5FLG + F(L-1)(2IJ)^3)$. The experiments of this paper were conducted with a HP Z800 desktop 4-core computer (8 threads) Xeon E5620 CPU at 2.4GHz and 17.6GB 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.

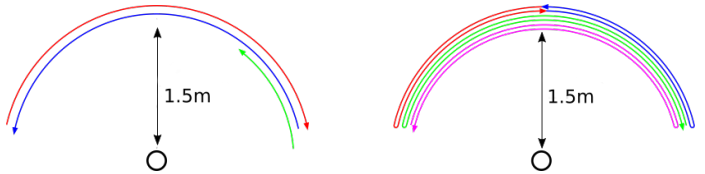


Fig. 2. Type I (left) and II (right) source trajectories. In Type I, Sources s_1 (red) and s_2 (blue) move from $-\vartheta$ to ϑ and from ϑ to $-\vartheta$ respectively, while Source s_3 moves from 85° to 45° . In Type II, sources move: from 0° to $-\vartheta$ and back (s_1 , red), from 0° to ϑ and 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, $\vartheta = 75^\circ$.

V. EXPERIMENTAL STUDY

A. Simulation Setup

To assess the performance of the proposed model and associated VEM algorithm, we conducted experiments with simulated 2-channel underdetermined time-varying convolutive mixtures of 3 or 4 speech signals. The source signals were monochannel 16 KHz signals randomly taken from the TIMIT database [45]. Each source signal was convolved with a set of impulse responses (see below) to produce the corresponding ground truth source image. These source images were then truncated to 32,768 samples (silence excluded), so that the spectrogram had $L = 128$ time frames. The images of the 3 or 4 sources were added to provide the mix signal.

As for the mixing filters, we used the binaural room impulse responses (BRIRs) from the University of Surrey [46]. These BRIRs were recorded with a dummy head equipped with 2 ear microphones, placed in a large lecture theatre of dimensions $23.5 \text{ m} \times 18.8 \text{ m} \times 4.6 \text{ m}$, attaining a reverberation-time of $\text{RT}_{60} \approx 0.68 \text{ s}$ [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. Each movement was then described by a sequence of 32,768 BRIRs. Due to memory limitations, we truncated the original 16,000-tap BRIRs to either 512 or 4,096 taps. Choosing two different lengths enables to evaluate the adequacy of the narrow-band assumption. Note that the recorded BRIRs almost vanish after 4,096 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 different setups for the movement of the sources around the dummy head. More precisely, we defined two mixture types, drawn in Fig. 2. Source s_3 in Type I mixtures always goes from 85° to 45° . The amplitude of the trajectory of all other sources is varied with $\vartheta \in \{15^\circ, 30^\circ, 45^\circ, 60^\circ, 75^\circ, 90^\circ\}$. Each trajectory is covered at fixed speed, within the 2s of the signal duration. In practice we used four different combinations of mixture type, filter tap length and number of sources, namely: *I-512-3, I-4096-3, II-512-3*,⁴ and *II-512-4*.

The STFT was applied to the mixed signal with a 512-sample, 50%-overlap, sine window, leading to $L = 128$

⁴In this case we discarded the fourth source (green plot in Fig. 2).

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.

Finally, two standard audio source separation 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]. In practice we used the `bss_eval_image` Matlab function dedicated to multichannel signals⁵ [48]. Each reported measure is the average over 10 experiments with different source signals, and different NMF initializations (see below).

B. Baseline method

The chosen baseline is a block-wise adaptation of the state-of-the-art method in [12]. We reutilized the implementation provided by the authors⁶ and followed 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 .

C. Initialization

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

Initialization is known to be a crucial step for the performance of 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). On the other hand, providing initial values that are too close to the ground truth is not an appropriate strategy since it does not reflect the amount of information available in real-world applications. We chose to relax the initialization strategy proposed in [12] to make it more realistic.

1) *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 $R \in \{20\text{dB}, 10\text{dB}, 0\text{dB}\}$, with 0dB meaning here equal power of signal $s_j(t)$ and of the sum of all interfering source signals. Note that $R = 20\text{dB}$ is a quite favorable initialization, whereas $R = 0\text{dB}$ 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.

2) *Mixing vectors*: As for the initialization of the mixing vectors $\hat{\mathbf{a}}_{:,f\ell}$, 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}_f^p (after applying a 512-point FFT). For the proposed method, this initial \mathbf{A}_f^p was 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.

3) *Other parameters*: The remaining parameters were initialized as follows: $\Sigma_{f\ell}^{\eta^a} = 10^3 \mathbf{I}_{JJ}$, $\mu_f^a = \hat{\mathbf{a}}_{:,f1}$, $\Sigma_f^a = \mathbf{I}_{JJ}$, $\forall f, \ell$. As for the sensor noise variance ν_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 ν_f , namely 1,000 times the (L, I) -average PSD of the mixture.

D. Results

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

1) *Detailed results for $\vartheta = 75^\circ$* : Fig. 3 represents the evolution of average SDR measures with the (V)EM iterations, for $\vartheta = 75^\circ$, 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

⁵http://bass-db.gforge.inria.fr/bss_eval/.

⁶<http://www.unice.fr/cfevotte/publications.html>.

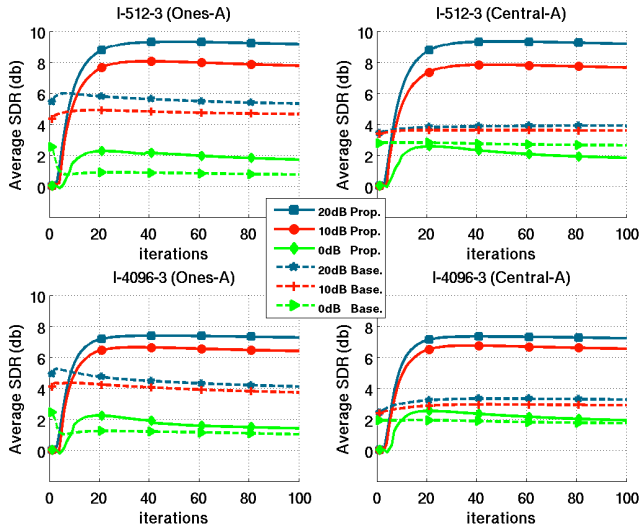


Fig. 3. Overall-sources average SDR vs iterations. For different initialization schemes: (top): *I-512-3*, (bottom): *I-4096-3*, (left) column is with *Ones-A* initialization, (right) is with *Central-A*. All experiments are at $\vartheta = 75^\circ$.

of STFT frames. Also, the baseline method has less parameters to estimate. Let us start the detailed analysis with *I-512-3 Central-A*. We see that the proposed method has an average performance of $\text{SDR} \approx 9.5\text{dB}$ for $R = 20\text{dB}$. The SDR score slightly degrades to about 8dB for $R = 10\text{dB}$, and then more abruptly decreases to about 2dB for $R = 0\text{dB}$, which shows the importance of NMF parameter initialization. SDR scores of the baseline method at $R = 20\text{dB}$, 10dB , and 0dB go from 4 to 2.5dB . Therefore, the proposed VEM largely outperforms the baseline method for $R = 20\text{dB}$ and 10dB , though in this example, the baseline performs slightly better at $R = 0\text{dB}$ ($\approx +0.5\text{dB}$ over the proposed method).

Regarding the influence of mixing vectors initialization, i.e. *Ones-A* vs *Central-A*, the proposed algorithm proves to be remarkably robust to poor mixing filter initialization: The blind initialization strategy *Ones-A* provides results that are similar to the *Central-A* configuration. 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 $R = 20\text{dB}$ and 10dB are again largely below the scores of the proposed method. However, and quite surprisingly, the baseline method behaves better (by about $0.4\text{--}0.7\text{dB}$) in the *Ones-A* (blind) configuration compared to the *Central-A* configuration, for $R = 20\text{dB}$ and 10dB . 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 $R = 0\text{dB}$, the filter information delivered by *Central-A* seems more useful, since the performance of the baseline method in the *Ones-A* configuration are about 2dB 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 $R = 0\text{dB}$.

As for the influence of the length of the BRIRs, we see

that, unsurprisingly, the performance of both proposed and baseline algorithms decrease when the BRIRs go from 512-tap to 4096-tap responses. For $R = 20\text{dB}$ and 10dB , we can observe that the decrease is of about $1.5\text{--}2\text{dB}$ for the proposed method, independently of the mixing vectors initialization. The decrease is lower for the baseline method ($\approx 1\text{dB}$), but this is probably related to the fact that the baseline scores are lower. For $R = 0\text{dB}$, 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 $R = 20\text{dB}$ and 10dB . All this manifests that (??) 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 $\vartheta = 75^\circ$ 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 $R = 20\text{dB}$ and $R = 10\text{dB}$, 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 $R = 20\text{dB}$ and $R = 10\text{dB}$ 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 2.1dB (s_2 in *II-512-4* at $R = 10\text{dB}$) to 4.0dB (s_1 in *II-512-3* at $R = 20\text{dB}$). SIR improvement of the proposed method over the baseline ranges from 2.1dB (s_2 in *I-512-3* at $R = 10\text{dB}$) to an impressive 5.9dB (s_3 in *I-512-3* at $R = 20\text{dB}$). 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 4.4dB (s_3 and s_4 at $R = 20\text{dB}$). At $R = 0\text{dB}$ 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 0dB are all higher for the proposed method than for the baseline method, except for s_2 in mixture *I-4096-3* (only 0.2dB below the baseline though). The improvement is however more limited than for $R = 20\text{dB}$ and $R = 10\text{dB}$ (maximum improvement is here 1.3dB). Finally, at $R = 0\text{dB}$, 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 going from 0.7dB to 1.7dB) and SIR (improvement going from 0.4dB to 2dB).

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 source⁷. Subtracting the scores

⁷We 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 2nd decimal figure), when it slightly degrades the corresponding SDR scores.

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

R	Mixture	SDR								SIR							
		Proposed				Baseline				Proposed				Baseline			
		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

in Table I and Table 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 a SDR gain ranging from 3.9dB to 7.8dB, and an SIR gain ranging from 4.1dB to 5.8dB. 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. At $R = 0$ dB, the SDR gain for s_1 , s_2 , s_3 and s_4 , is 9.6dB, 9.3dB, 8.7dB and 7.9dB, respectively, whereas the SIR gain is 10.5dB, 9.6dB, 9.4dB and 9.3dB, respectively. This tends to show that the proposed method is robust to different source speed, which we verify more systematically in the next subsection. Importantly, we remind that $R = 0$ dB represents a very reasonable amount of source initial information. Thus, in such a difficult configuration as reverberant 2×4 time-varying mixture, highly-corrupted initial source parameters, blind filter initialization, an SIR gain of about 9–10.5dB is very satisfactory.

TABLE II
INPUT SDR AND SIR FOR THE 4 DIFFERENT MIXTURES.

Mixture	SDR				SIR			
	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

2) *Effect of ϑ* : 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

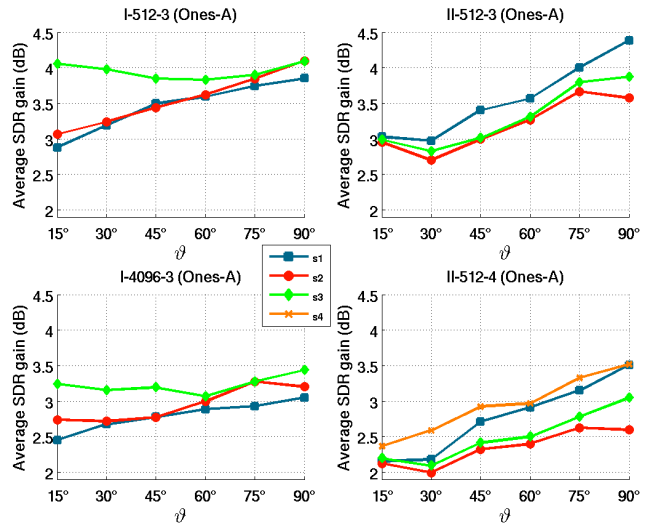


Fig. 4. Average SDR gain of the proposed method over the baseline method, for the 4-source mixture, as a function of ϑ ($R = 20$ dB, *Ones-A* initialization).

$R = 20$ dB, and *Ones-A* strategy (as the latter was shown to be most favorable for the baseline).

Let us start with *II-512-3*. We observe that, except for the 3 sources at $\vartheta = 30^\circ$ and for s_2 at $\vartheta = 90^\circ$, the gain is monotonically increasing for all three sources, starting from about 3dB at $\vartheta = 15^\circ$ and going up to 3.5–4.5dB at $\vartheta = 90^\circ$. 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 general trend is also visible on the other plots of Fig. 4, though with configuration specificities. For example, for the *I-512-3* mixture, we see that it the gain increases quite regularly

with ϑ for s_1 and s_2 , from about 3dB at $\vartheta = 15^\circ$ to about 4dB at $\vartheta = 90^\circ$, whereas the gain for s_3 (whose trajectory remains independent of ϑ) is almost constant at about 4dB. The decreasing of this latter curve a bit around $\vartheta = 45^\circ$ may be due to the trajectories of s_1 and s_2 interfering with the trajectory of s_3 for $\vartheta \geq 45^\circ$. However, the curve increases again after $\vartheta = 60^\circ$ and this is more difficult to interpret. Note that even if the general trend of those curves is to show that the difference between the proposed method and the baseline method increases with source velocity, 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 (s_3 is the slower source for $\vartheta \geq 20^\circ$). For the *II-512-4* configuration, we observe that the behavior of the gain is very similar to *II-512-3*, though with smaller values. The addition of the fourth source makes the source and mixing filter estimation more difficult for both algorithms, thus somehow bringing the performances of the two methods closer. In the same line, going from 512-tap mixing filters to 4096-tap mixing filters also bring the performances of the two methods closer, as seen by comparing the bottom-left figure with the top-left figure. This is probably also because demixing mixture *I-4096-3* is more difficult than demixing *I-512-3* for both methods.

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 variational EM (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, as for the initialization, the $R = 0$ dB configuration provided a limited amount of source information initialization, but we have to investigate a completely realistic setup, where the initialization is done using only the mixture signal. For example we may combine the proposed model with an observation vector clustering scheme inspired from, e.g., [27], [49]. Second, the effect of the number of components in this framework should be investigated as well as its dependency on the number of sources. In this line, 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. This could be a crucial step in making the proposed sound source separation algorithm from time-varying convolutive mixtures more robust. Finally, an in-depth study exploring the complex relationship between the physical changes of the recording setup 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*. Academic Press, 2010.
- [2] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time Fourier transform domain," *IEEE Signal Processing Letters*, vol. 14, no. 5, pp. 337–340, 2007.
- [3] E. Vincent, M. Jafari, S. Abdallah, M. Plumbley, and M. Davies, *Probabilistic modeling paradigms for audio source separation*. IGI Global Press, 2010, pp. 162–185.
- [4] A. Hyvärinen, J. Karhunen, and E. Oja, Eds., *Independent Component Analysis*. Wiley and Sons, 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 Journal on Advances in Signal Processing*, p. Article ID 24717, 2007.
- [6] M. Mandel, R. J. Weiss, D. P. Ellis *et al.*, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.
- [7] A. Luttkus, B. Badeau, and G. Richard, "Gaussian processes for underdetermined source separation," *IEEE Trans. on Signal Proc.*, vol. 59, no. 7, pp. 3155–3167, 2011.
- [8] D. Ephraim, Yariv Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol. 33, no. 6, pp. 443–445, 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 *IEEE Int. Conf. Acoustics, Speech, and Signal Proc.*, vol. 6, 2003, pp. 613–616.
- [10] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 14, no. 1, pp. 191–199, 2006.
- [11] C. Févotte and J.-F. Cardoso, "Maximum likelihood approach for blind audio source separation using time-frequency Gaussian source models," in *IEEE Workshop Applcat. Signal Process. to Audio and Acoust. (WASPAA)*, 2005.
- [12] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 18, no. 3, pp. 550–563, 2010.
- [13] N. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 18, no. 7, pp. 1830–1840, 2010.
- [14] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 20, no. 4, pp. 1118–1133, 2012.
- [15] D. Lee and H. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.
- [16] —, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems*, 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 Computation*, 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 Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 69–84, 2011.
- [19] J. Anemüller and T. Gramss, “On-line blind separation of moving sound sources,” in *Proc. Int. Conf. Independent Component Analysis and Blind Source Separation (ICA’99)*, Aussois, France, 1999.
- [20] A. Koutras, E. Dermatas, and G. Kokkinakis, “Blind speech separation of moving speakers in real reverberant environments,” in *Proc. Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, Istanbul, Turkey, 2000.
- [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. Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, Orlando, Florida, 2002.
- [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 Component Analysis and Blind Source Separation (ICA’03)*, 2003.
- [23] R. E. Prieto and J. Pamornpol, “Blind source separation for time-variant mixing systems using piecewise linear approximations,” in *Proc. Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2005.
- [24] R. Mukai, H. Sawada, S. Araki, and S. Makino, “Robust real-time blind source separation for moving speakers in a room,” in *Proc. Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2003.
- [25] W. Addison and S. Roberts, “Blind source separation with non-stationary mixing using wavelets,” in *Proc. Int. Conf. Independent Component Analysis and Blind Source Separation (ICA’06)*, 2006.
- [26] K. Nakadai, H. Nakajima, Y. Hasegawa, and H. Tsujino, “Sound source separation of moving speakers for robot audition,” in *Proc. Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2009.
- [27] S. Araki, H. Sawada, R. Mukai, and S. Makino, “Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors,” *Signal Processing*, 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. Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2009.
- [29] L. Simon and E. Vincent, “A general framework for online audio source separation,” in *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation (LVA)*, 2012.
- [30] S. Markovich-Golan, S. Gannot, and I. Cohen, “Subspace tracking of multiple sources and its application to speakers extraction,” in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Dallas, Texas, USA, 2010.
- [31] E. Weinstein, A. Oppenheim, M. Feder, and J. Buck, “Iterative and sequential algorithms for multisensor signal enhancement,” *IEEE Trans. on Signal Processing*, vol. 42, no. 4, pp. 846–859, 1994.
- [32] T. Higuchi, N. Takamune, N. Tomohiko, and H. Kameoka, “Under-determined blind separation and tracking of moving sources based on DOA-HMM,” in *Proc. Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2014.
- [33] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
- [34] S. Gannot and M. Moonen, “On the application of the unscented Kalman filter to speech processing,” in *IEEE Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Kyoto, Japan, 2003.
- [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 *IEEE Workshop Appl. Signal Process. to Audio and Acoust. (WASPAA)*, 2015.
- [36] F. Neeser and J. Massey, “Proper complex random processes with applications to information theory,” *IEEE Trans. on Information Theory*, vol. 39, no. 4, pp. 1293–1302, 1993.
- [37] T. Virtanen, “Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria,” *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 15, no. 3, pp. 1066–1074, 2007.
- [38] N. Mohammadiha, P. Smaragdis, and A. Leijon, “Supervised and unsupervised speech enhancement using nonnegative matrix factorization,” *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 21, no. 10, pp. 2140–2151, 2013.
- [39] L. Parra and C. Spence, “Convulsive blind separation of non-stationary sources,” *IEEE Trans. on Speech and Audio Proc.*, vol. 8, no. 3, pp. 320–327, 2000.
- [40] S. Gannot, D. Burshtein, and E. Weinstein, “Signal enhancement using beamforming and nonstationarity with applications to speech,” *IEEE Trans. on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.
- [41] G. McLachlan and K. Thriyambakam, *The EM algorithm and extensions*. New-York, USA: John Wiley and sons, 1997.
- [42] V. Smidl and A. Quinn, *The Variational Bayes Method in Signal Processing*. Berlin: Springer-Verlag, 2006.
- [43] A. Hjørungnes and D. Gesbert, “Complex-valued matrix differentiation: Techniques and key results,” *IEEE Transactions on Signal Processing*, vol. 55, no. 6, pp. 2740–2746, June 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,” in *Proc. Convention of the Audio Engineering Society (AES)*, 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,” 1993, linguistic Data Consortium, Philadelphia.
- [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. on Audio, Speech, and Language Proc.*, vol. 14, no. 4, pp. 1462–1469, 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 *Int. Conf. on Independent Component Analysis and Signal Separation*. Springer, 2007, pp. 552–559.
- [49] L. Drude, A. Chinaev, D. H. Tran Vu, and R. Haeb-Umbach, “Source counting in speech mixtures using a variational EM approach for complex Watson mixture models,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014.