A Bayesian Hierarchical Model for Blind Audio Source Separation

Yaron Laufer and Sharon Gannot
Faculty of Engineering, Bar-Ilan University, Ramat-Gan, 5290002, Israel
{yaron.laufer,sharon.gannot}@biu.ac.il

***Abstract*—This paper presents a fully Bayesian hierarchical model for blind audio source separation in a noisy environment. Our probabilistic approach is based on Gaussian priors for the speech signals, Gamma hyperpriors for the speech precisions and a Gamma prior for the noise precision. The time-varying acoustic channels are modelled with a linear-Gaussian state-space model. The inference is carried out using a variational Expectation-Maximization (VEM) algorithm, leading to a variant of the multi-speaker multichannel Wiener filter (MCWF) to separate and enhance the audio sources, and a Kalman smoother to infer the acoustic channels. The VEM speech estimator can be decomposed into two stages: A multi-speaker linearly constrained minimum variance (LCMV) beamformer followed by a variational multi-speaker postfilter. The proposed algorithm is evaluated in a static scenario using recorded room impulse responses (RIRs) with two reverberation levels, showing superior performance compared to competing methods.**

***Index Terms*—Audio source separation, Variational EM.**

I. Introduction

A fundamental problem in the field of audio signal processing is the blind separation of multiple speakers from a mixture signal, recorded in a noisy environment. A common solution is to construct the multi-speaker MCWF beamformer, which is the multichannel minimum mean squared error (MMSE) estimator of the desired speakers, assuming that the signal components are Gaussians [1]. The design of the MCWF requires the estimation of several acoustic parameters, namely the relative transfer function (RTF) matrix, the source power spectral density (PSD) and the noise covariance matrix. These parameters can be jointly estimated using the maximum likelihood (ML) or maximum a posteriori (MAP) criteria. When the resulting optimization problem is complex, the Expectation Maximization (EM) algorithm [2] is a solution that iteratively decomposes the problem into several smaller optimization problems involving subsets of parameters, which are solved separately.

In the Bayesian framework, model parameters are viewed as random variables having a prior probability density function (PDF), rather than deterministic unknown parameters. This approach allows us to include prior knowledge and to explore uncertainty in the model. Rather than point estimates, the inference process is based on the entire posterior PDF, and thus the obtained estimators are more robust and less sensitive to local maxima [3]. Hierarchical Bayesian models use a multi-level modeling to capture important dependencies among parameters. However, the posterior distribution might be intractable in complex Bayesian models, and thus the EM algorithm cannot be applied. The variational approach [3]–[5] circumvents this difficulty by approximating the posterior. Recently, several works utilized the VEM method for speech enhancement [6]–[9] and speech dereverberation [10], [11].

In [12], the problem of blind audio source separation is addressed using the EM algorithm, which simultaneously estimates the speech signals and the model parameters. The source signals in the short-time Fourier transform (STFT) domain are modelled as complex-Gaussians, and their PSD with a non-negative matrix factorization (NMF) model. The acoustic transfer function (ATF) matrix, assumed to be time-invariant, is modelled as a deterministic unknown parameter. Bayesian extensions for the NMF factors can be found in, e.g., [13], [14]. Adopting the NMF framework of [12], the problem of separating moving sound sources is addressed in [15], by modeling the time-varying ATF matrix as a set of temporally-linked latent variables, parametrized with a first-order Markov model. Due to the complex structure, the posterior distribution is intractable, and the variational approach is adopted. By modeling the speech signals and the acoustic channels as latent variables, their posterior distribution are jointly estimated in the E-step. However, the authors did not adopt a fully Bayesian model, since the source PSD and the noise covariance matrix are still deterministic unknown parameters, for which point estimates are computed in the M-step. Moreover, the results are highly sensitive to the initial values of the NMF parameters. Note also that only the case of spatially white noise field is considered in [12], [15].

Recently, several deep neural network (DNN)-based methods were proposed for source separation, e.g. [16], [17]. These methods require training data, specifically reverberant utterances in multiple acoustic environments, thus motivating the use of blind source separation techniques, as the proposed method. Other common blind methods are independent vector analysis (IVA) [18] and independent low-rank matrix analysis (ILRMA) [19], which unifies IVA and NMF frameworks.

In this work, we propose a fully Bayesian model for blind separation of moving speakers in noisy conditions. We introduce a hierarchical model based on Gaussian priors for the speech signals and Gamma hyperpriors for the speech precisions. For the time-varying RTF, the probabilistic model

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme, Grant Agreement No. 871245. Yaron Laufer is supported by the Adams Fellowship Program of the Israel Academy of Sciences and Humanities.

of [15] is adopted. The noise is modelled as a spatially homogeneous sound field, and a Gamma prior is assumed on the noise precision. As the precisions are modelled as latent random variables, their posterior distribution are inferred through a VEM algorithm. Inspired by the decomposition of the multi-speaker MMSE estimator, i.e. the MCWF, into a multi-speaker LCMV beamformer followed by a subsequent multi-speaker Wiener postfilter [20], we show that the VEM multi-speaker estimator has an analogous decomposition. Similarly to the MCWF, it includes an LCMV beamformer as an initial stage. However, the multi-speaker Wiener postfilter is substituted by a variational multi-speaker postfilter, which takes into account the uncertainty in the RTF estimate and weights accordingly the LCMV outputs.

II. Problem Formulation

A. Signal Model

Consider a mixture of J speakers received by N microphones, in a noisy acoustic environment. We work in the STFT domain, where $k \in \{1, \ldots, K\}$ denotes the frequency band, and $\ell \in \{1, \ldots, L\}$ is the time frame. The N-channel measurement signal $\mathbf{x}(\ell, k) = [x_1(\ell, k), \cdots, x_N(\ell, k)]^\top$ writes

$$\mathbf{x}(\ell, k) = \mathbf{A}(\ell, k)\mathbf{s}(\ell, k) + \mathbf{u}(\ell, k), \tag{1}$$

where $\mathbf{s}(\ell, k) = [s_1(\ell, k), s_2(\ell, k), \cdots, s_J(\ell, k)]^\top$ is the vector of the speech signals as received by the first microphone (designated as the reference microphone), $\mathbf{A}(\ell, k) = [\mathbf{a}_1(\ell, k), \cdots, \mathbf{a}_J(\ell, k)]$ is the $N \times J$ RTF matrix and $\mathbf{u}(\ell, k) = [u_1(\ell, k), \cdots, u_N(\ell, k)]^\top$ is the additive noise.

The J speech signals are modelled as independent zero-mean Gaussian random variables, having distinct precisions, denoted by $\boldsymbol{\tau}(\ell, k) = [\tau_1(\ell, k), \cdots, \tau_J(\ell, k)]^\top$. The PDF of the speech vector $\mathbf{s}$ therefore writes $p\big(\boldsymbol{s}(\ell, k)|\boldsymbol{\tau}(\ell, k)\big) = \mathcal{N}_c\Big(\mathbf{s}(\ell, k); \mathbf{0}, \operatorname{diag}^{-1}\big(\boldsymbol{\tau}(\ell, k)\big)\Big)$. The noise is modelled as a zero-mean multivariate Gaussian with $p\big(\mathbf{u}(\ell, k)|\boldsymbol{\Phi}_\mathbf{u}(k)\big) = \mathcal{N}_c\big(\mathbf{u}(\ell, k); \mathbf{0}, \boldsymbol{\Phi}_\mathbf{u}(k)\big)$. The noise is assumed to be a spatially homogeneous sound field, i.e. $\boldsymbol{\Phi}_\mathbf{u}(k) = \beta^{-1}(k)\boldsymbol{\Gamma}(k)$, where $\beta(k)$ is the inverse power of the noise and $\boldsymbol{\Gamma}(k)$ is a spatial coherence matrix, assumed to be known. The conditional data distribution is therefore given by $p\big(\mathbf{x}(\ell, k)|\mathbf{A}(\ell, k), \mathbf{s}(\ell, k), \beta(k)\big) = \mathcal{N}_c\big(\mathbf{x}(\ell, k); \mathbf{A}(\ell, k)\mathbf{s}(\ell, k), \beta^{-1}(k)\boldsymbol{\Gamma}(k)\big)$.

Let $\mathbf{v}(\ell, k)$ denote the column-wise vectorization of $\mathbf{A}(\ell, k)$, i.e. $\mathbf{v}(\ell, k) \triangleq \operatorname{vec}(\mathbf{A}(\ell, k)) = [\mathbf{a}_1^\top(\ell, k) \cdots \mathbf{a}_J^\top(\ell, k)]^\top \in \mathbb{C}^{NJ}$. In many realistic scenarios, the audio channel might be time-varying. Thus, the RTF matrix is modelled as a set of temporally-linked continuous latent variables, parameterized with a first-order linear dynamical system (LDS), as in [15]:

$$p\big(\mathbf{v}(1, k)\big) = \mathcal{N}_c\big(\mathbf{v}(1, k); \boldsymbol{\mu}_\mathbf{v}(k), \boldsymbol{\Phi}_\mathbf{v}(k)\big), \tag{2}$$

$$p\big(\mathbf{v}(\ell, k)|\mathbf{v}(\ell-1, k)\big) = \mathcal{N}_c\big(\mathbf{v}(\ell, k); \mathbf{v}(\ell-1, k), \boldsymbol{\Phi}_\mathbf{v}(k)\big), \tag{3}$$

with the mean vector $\boldsymbol{\mu}_\mathbf{v}(k) \in \mathbb{C}^{NJ}$ and the covariance matrix $\boldsymbol{\Phi}_\mathbf{v}(k) \in \mathbb{C}^{NJ \times NJ}$. For brevity, $\mathbf{v}(1{:}L, k) = \{\mathbf{v}(\ell, k)\}_{\ell=1}^{L}$ denotes the entire sequence of RTFs at frequency k.

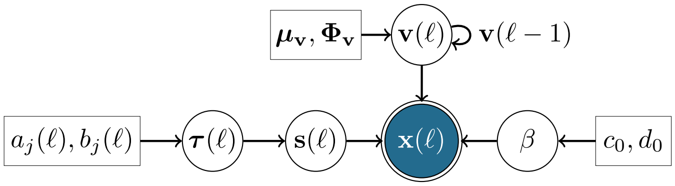

Fig. 1: Graphical model (Frequency index is omitted).

B. Conjugate Priors

In the Bayesian framework, it is common to introduce probabilistic priors over the latent variables, which allows us to take into account the uncertainty in the model. We therefore establish a generative hierarchical model by introducing priors with unknown parameters on the precisions of the speakers and the noise. The conjugate prior for the precision of a univariate Gaussian is the Gamma distribution [5]. Hence, the prior for the speech precisions is given by:

$$p\big(\boldsymbol{\tau}(\ell, k)\big) = \prod_{j=1}^{J} \operatorname{Gam}\big(\tau_j(\ell, k); a_j(\ell, k), b_j(\ell, k)\big). \tag{4}$$

Similarly, we assume a Gamma prior for the noise precision:

$$p\big(\beta(k)\big) = \operatorname{Gam}\big(\beta(k); c_0(k), d_0(k)\big). \tag{5}$$

The proposed hierarchical model is illustrated in Fig. 1.

III. Variational EM for Source Separation

In this work, the set of observations is denoted by $\mathcal{X} = \{\mathbf{x}(\ell, k)\}_{\ell,k=1}^{L,K}$, the set of hidden variables consists of $\mathcal{H} = \{\mathbf{s}(\ell, k), \mathbf{v}(\ell, k), \boldsymbol{\tau}(\ell, k), \beta(k)\}_{\ell,k=1}^{L,K}$, and the parameter set consists of $\Theta = \big\{a_j(\ell, k), b_j(\ell, k), c_0(k), d_0(k), \boldsymbol{\mu}_\mathbf{v}(k), \boldsymbol{\Phi}_\mathbf{v}(k)\big\}_{\ell,k,j=1}^{L,K,J}$. Bayesian inference aims to infer the latent variables according to their posterior mean (PM). To this end, it is required to compute the posterior distribution of the hidden variables $p(\mathcal{H}|\mathcal{X}; \Theta) = p(\mathcal{X}, \mathcal{H}; \Theta)/p(\mathcal{X}; \Theta)$. In our model, the complete-data distribution writes

$$\begin{aligned} p(\mathcal{X}, \mathcal{H}; \Theta) = \prod_{\ell,k=1}^{L,K} \Big[& p\big(\mathbf{x}(\ell, k)|\mathbf{A}(\ell, k), \mathbf{s}(\ell, k), \beta(k)\big) \\ & \times p\big(\mathbf{s}(\ell, k)|\boldsymbol{\tau}(\ell, k)\big) \prod_{j=1}^{J} p\big(\tau_j(\ell, k); a_j(\ell, k), b_j(\ell, k)\big)\Big] \\ & \times \prod_{k=1}^{K} \Big[p\big(\beta(k); c_0(k), d_0(k)\big) p\big(\mathbf{v}(1, k); \boldsymbol{\mu}_\mathbf{v}(k), \boldsymbol{\Phi}_\mathbf{v}(k)\big) \\ & \times \prod_{\ell=2}^{L} p\big(\mathbf{v}(\ell, k); \mathbf{v}(\ell-1, k), \boldsymbol{\Phi}_\mathbf{v}(k)\big)\Big]. \end{aligned} \tag{6}$$

Due to the complex form of (6), the likelihood $p(\mathcal{X}; \Theta) = \int p(\mathcal{X}, \mathcal{H}; \Theta) d\mathcal{H}$ cannot be computed analytically and thus exact inference becomes intractable. We therefore propose a variational inference procedure, which approximates the posterior $q(\mathcal{H}) \approx p(\mathcal{H}|\mathcal{X}; \Theta)$. According to the *mean field theory* [4], [21], we assume that the speech signals, RTF matrix, speech precisions and noise precision are conditionally independent given the observations. Hence, the approximate posterior distribution can be factorized as:

$$q(\mathcal{H}){=}\prod_{\ell,k=1}^{L,K}\Big[q\big(\mathbf{s}(\ell,k)\big)q\big(\boldsymbol{\tau}(\ell,k)\big)\Big]\prod_{k=1}^{K}\Big[q\big(\beta(k)\big)q\big(\mathbf{v}(1{:}L,k)\big)\Big]. \tag{7}$$

The VEM procedure consists in iterating the following two steps until convergence. In the E-Step, the approximate posterior distribution of each subset $\mathcal{H}_i \subset \mathcal{H}$ is computed by [5]:

$$\ln q(\mathcal{H}_i) = \mathbb{E}_{q(\mathcal{H}/\mathcal{H}_i)}[\ln p(\mathcal{X},\mathcal{H};\Theta)] + \text{const}, \tag{8}$$

where $q(\mathcal{H}/\mathcal{H}_i)$ is the approximate joint posterior distribution of all latent variables, excluding $\mathcal{H}_i$. In the subsequent M-step, the expected log-likelihood of the complete-data $\mathcal{L}(\Theta) = \mathbb{E}_{q(\mathcal{H})}[\ln p(\mathcal{X},\mathcal{H};\Theta)]$ is maximized w.r.t. the parameters Θ. For brevity, the frequency bin index k is henceforth omitted.

A. *E-*$\mathbf{s}$ *Step*

The approximate posterior PDF of the sources is obtained from (6) and (8) by identifying the terms that depend on $\mathbf{s}(\ell)$:

$$\begin{aligned}\ln q\big(\mathbf{s}(\ell)\big) \propto \mathbb{E}_{q(\mathbf{A}(\ell))q(\tau(\ell))q(\beta)}\Big[&\ln p\big(\mathbf{x}(\ell)|\mathbf{A}(\ell),\mathbf{s}(\ell),\beta\big)\\ &+\ln p\big(\mathbf{s}(\ell)|\boldsymbol{\tau}(\ell)\big)\Big].\end{aligned} \tag{9}$$

It can be shown that (9) yields a Gaussian distribution $q\left(\mathbf{s}(\ell)\right) = \mathcal{N}_c\big(\mathbf{s}(\ell);\hat{\mathbf{s}}(\ell),\boldsymbol{\Sigma}_{\mathbf{s}}(\ell)\big)$, with

$$\boldsymbol{\Sigma}_{\mathbf{s}}(\ell) = \Big(\hat{\beta}\mathbb{E}[\mathbf{A}^{\mathrm{H}}(\ell)\boldsymbol{\Gamma}^{-1}\mathbf{A}(\ell)] + \operatorname{diag}\big(\hat{\boldsymbol{\tau}}(\ell)\big)\Big)^{-1}, \tag{10}$$

$$\hat{\mathbf{s}}(\ell) = \boldsymbol{\Sigma}_{\mathbf{s}}(\ell)\hat{\mathbf{A}}^{\mathrm{H}}(\ell)\hat{\beta}\boldsymbol{\Gamma}^{-1}\mathbf{x}(\ell), \tag{11}$$

where $\hat{\mathbf{A}}(\ell),\hat{\beta},\hat{\boldsymbol{\tau}}(\ell)$ are posterior statistics that will be defined in the following sections. For brevity, let $\hat{\mathbf{T}}(\ell) \triangleq \operatorname{diag}\big(\hat{\boldsymbol{\tau}}(\ell)\big)$. The remaining expectation term in (10) can be simplified as follows. Let $\hat{\mathbf{U}}(\ell) \triangleq \mathbb{E}[\mathbf{A}^{\mathrm{H}}(\ell)\boldsymbol{\Gamma}^{-1}\mathbf{A}(\ell)] \in \mathbb{C}^{J\times J}$ [15]. In Section III-B, the vectorized PM of $\mathbf{A}(\ell)$ and the corresponding covariance matrix are defined by $\hat{\mathbf{v}}(\ell) \in \mathbb{C}^{NJ}$ and $\boldsymbol{\Sigma}_{\mathbf{v}}(\ell) \in \mathbb{C}^{NJ\times NJ}$, respectively. Let $\boldsymbol{\Sigma}_{\mathbf{v}}^{rj}(\ell)$ denote the (r,j)-th $N\times N$ sub-block of $\boldsymbol{\Sigma}_{\mathbf{v}}(\ell)$, where $r,j = 1,\ldots,J$. Note that $\Big[\hat{\mathbf{U}}(\ell)\Big]_{jr} = \mathbb{E}[\mathbf{a}_j^{\mathrm{H}}\boldsymbol{\Gamma}^{-1}\mathbf{a}_r(\ell)] = \hat{\mathbf{a}}_j^{\mathrm{H}}(\ell)\boldsymbol{\Gamma}^{-1}\hat{\mathbf{a}}_r(\ell) + \operatorname{Tr}\left[\boldsymbol{\Sigma}_{\mathbf{v}}^{rj}(\ell)\boldsymbol{\Gamma}^{-1}\right]$. We conclude that $\hat{\mathbf{U}}(\ell) = \hat{\mathbf{A}}^{\mathrm{H}}(\ell)\boldsymbol{\Gamma}^{-1}\hat{\mathbf{A}}(\ell) + \boldsymbol{\Psi}(\ell)$, where $\hat{\mathbf{A}}$ is the matrix form of $\hat{\mathbf{v}}$, and $[\boldsymbol{\Psi}(\ell)]_{jr} \triangleq \operatorname{Tr}\left[\boldsymbol{\Sigma}_{\mathbf{v}}^{rj}(\ell)\boldsymbol{\Gamma}^{-1}\right]$. It follows that

$$\begin{aligned}\hat{\mathbf{s}}(\ell) = \Big(&\hat{\mathbf{A}}^{\mathrm{H}}(\ell)\hat{\beta}\boldsymbol{\Gamma}^{-1}\hat{\mathbf{A}}(\ell) + \hat{\beta}\boldsymbol{\Psi}(\ell) + \hat{\mathbf{T}}(\ell)\Big)^{-1}\\ &\times\hat{\mathbf{A}}^{\mathrm{H}}(\ell)\hat{\beta}\boldsymbol{\Gamma}^{-1}\mathbf{x}(\ell).\end{aligned} \tag{12}$$

The speech signals can be estimated by the PM, namely $\hat{\mathbf{s}}(\ell)$, with the covariance matrix $\boldsymbol{\Sigma}_{\mathbf{s}}(\ell)$. The second-order posterior moment is defined by $\mathbf{Q}_{\mathbf{s}}(\ell) = \hat{\mathbf{s}}(\ell)\hat{\mathbf{s}}^{\mathrm{H}}(\ell) + \boldsymbol{\Sigma}_{\mathbf{s}}(\ell)$. Note that in the single-speaker case, the solution reduces to that of [7].

The form of (12) resembles the multi-speaker MCWF [1]:

$$\begin{aligned}\hat{\mathbf{s}}_{\text{MCWF}}(\ell) = \Big(&\hat{\mathbf{A}}^{\mathrm{H}}(\ell)\hat{\beta}\boldsymbol{\Gamma}^{-1}\hat{\mathbf{A}}(\ell) + \hat{\mathbf{T}}(\ell)\Big)^{-1}\\ &\times\hat{\mathbf{A}}^{\mathrm{H}}(\ell)\hat{\beta}\boldsymbol{\Gamma}^{-1}\mathbf{x}(\ell),\end{aligned} \tag{13}$$

except the term $\hat{\beta}\boldsymbol{\Psi}(\ell)$. In [20], the multi-speaker MCWF is decomposed into a multi-speaker LCMV beamformer followed by multi-speaker Wiener postfilter. In a similar way, $\hat{\mathbf{s}}(\ell)$ in (12) can be recast as

$$\begin{aligned}\hat{\mathbf{s}}(\ell){=}&\underbrace{\Big(\hat{\mathbf{A}}^{\mathrm{H}}(\ell)\hat{\beta}\boldsymbol{\Gamma}^{-1}\hat{\mathbf{A}}(\ell) + \hat{\beta}\boldsymbol{\Psi}(\ell) + \hat{\mathbf{T}}(\ell)\Big)^{-1}\hat{\mathbf{A}}^{\mathrm{H}}(\ell)\hat{\beta}\boldsymbol{\Gamma}^{-1}\hat{\mathbf{A}}(\ell)}_{\mathbf{H}^{\mathrm{H}}(\ell)}\\ &\times\underbrace{\Big(\hat{\mathbf{A}}^{\mathrm{H}}(\ell)\boldsymbol{\Gamma}^{-1}\hat{\mathbf{A}}(\ell)\Big)^{-1}\hat{\mathbf{A}}^{\mathrm{H}}(\ell)\boldsymbol{\Gamma}^{-1}}_{\mathbf{W}_{\text{LCMV}}^{\mathrm{H}}(\ell)}\mathbf{x}(\ell).\end{aligned} \tag{14}$$

Due to RTF estimation errors, there might be a leakage between sources at the output of the LCMV stage. The multi-speaker Wiener postfilter reduces this leakage and enhances the LCMV outputs [20]. However, it treats the RTF estimator $\hat{\mathbf{A}}(\ell)$ as a point estimator, and ignores its uncertainty. In contrast, the proposed method views the RTF as a random variable, whose posterior distribution encapsulates the uncertainty about the parameter. Thus, it includes a multi-speaker postfilter $\mathbf{H}(\ell)$ that takes into account the uncertainty level, expressed by $\boldsymbol{\Sigma}_{\mathbf{v}}(\ell)$, and weights accordingly the LCMV outputs. When $\boldsymbol{\Sigma}_{\mathbf{v}}(\ell) \to \mathbf{0}$, then $\boldsymbol{\Psi}(\ell) \to \mathbf{0}$ and $\hat{\mathbf{s}}(\ell)$ reduces to $\hat{\mathbf{s}}_{\text{MCWF}}(\ell)$.

B. *E-*$\mathbf{v}$ *Step*

With (8), the joint posterior PDF of the RTF sequence writes

$$\begin{aligned}\ln q(\mathbf{v}(1{:}L)) \propto \sum_{\ell=1}^{L}\mathbb{E}_{q(\mathbf{s}(\ell))q(\beta)}&\Big[\ln p\big(\mathbf{x}(\ell)|\mathbf{A}(\ell),\mathbf{s}(\ell),\beta\big)\Big]\\ &+\ln p\big(\mathbf{v}(1{:}L)\big).\end{aligned} \tag{15}$$

The first term of (15) can be reduced to a Gaussian distribution [15] $\mathcal{N}_c\big(\boldsymbol{\mu}_L(\ell);\mathbf{v}(\ell),\boldsymbol{\Phi}_L(\ell)\big)$, with $\boldsymbol{\mu}_L(\ell) = \operatorname{vec}\left(\mathbf{x}(\ell)\hat{\mathbf{s}}^{\mathrm{H}}(\ell)\mathbf{Q}_{\mathbf{s}}^{-1}(\ell)\right)$, and $\boldsymbol{\Phi}_L(\ell) = \Big(\mathbf{Q}_{\mathbf{s}}^{\top}(\ell)\otimes\hat{\beta}\boldsymbol{\Gamma}^{-1}\Big)^{-1}$, where $\otimes$ denotes the Kronecker product.[1] Since the second term of (15) is also a Gaussian, it follows that (15) is a first-order LDS [5] over $\{\mathbf{v}(\ell)\}_{\ell=1}^{L}$. Thus, the marginal posterior PDF is Gaussian, $q(\mathbf{v}(\ell)) = \mathcal{N}_c(\mathbf{v}(\ell);\hat{\mathbf{v}}(\ell),\boldsymbol{\Sigma}_{\mathbf{v}}(\ell))$, which can be recursively calculated with the *Kalman smoother* [5], using $\boldsymbol{\mu}_L(\ell)$, $\boldsymbol{\Phi}_L(\ell)$ and $\boldsymbol{\Phi}_{\mathbf{v}}$. Hence, the RTF can be estimated by $\hat{\mathbf{v}}(\ell)$, with uncertainty $\boldsymbol{\Sigma}_{\mathbf{v}}(\ell)$. The pair-wise joint posterior distribution of two successive frames, required to update $\boldsymbol{\Phi}_{\mathbf{v}}$ in Section III-E, is obtained by marginalizing out all other frames in (15): $q\big(\mathbf{v}(\ell),\mathbf{v}(\ell{-}1)\big) = \mathcal{N}_c\big(\left[\mathbf{v}(\ell)^{\top},\mathbf{v}(\ell-1)^{\top}\right]^{\top};\mathbf{a}_{\xi}(\ell),\boldsymbol{\Sigma}_{\xi}(\ell)\big)$, where $\mathbf{a}_{\xi}(\ell) \in \mathbb{C}^{2NJ}$ and $\boldsymbol{\Sigma}_{\xi}(\ell) \in \mathbb{C}^{2NJ\times 2NJ}$. The second-order joint posterior moment is $\boldsymbol{Q}_{\xi}(\ell) = \boldsymbol{\Sigma}_{\xi}(\ell) + \mathbf{a}_{\xi}(\ell)\mathbf{a}_{\xi}^{\mathrm{H}}(\ell)$.

C. *E-*$\boldsymbol{\tau}$ *Step*

Using (8), the posterior PDF of the speech precisions writes

$$\begin{aligned}\ln q(\boldsymbol{\tau}(\ell)) \propto \mathbb{E}_{q(\mathbf{s}(\ell))}&\Big[\ln p\big(\mathbf{s}(\ell)|\boldsymbol{\tau}(\ell)\big)\Big]\\ &+\sum_{j=1}^{J}\ln p\big(\tau_j(\ell);a_j(\ell),b_j(\ell)\big),\end{aligned} \tag{16}$$

[1] The proof follows by standard properties of the Kronecker product [22].

which is a product of J independent Gamma distributions: $q(\boldsymbol{\tau}(\ell)) = \prod_{j=1}^{J} \text{Gam}\big(\tau_j(\ell); a_{p,j}(\ell), b_{p,j}(\ell)\big)$, with

$$a_{p,j}(\ell) = a_j(\ell) + 1 \ , \ b_{p,j}(\ell) = b_j(\ell) + \widehat{|s_j(\ell)|^2}, \qquad (17)$$

where $\widehat{|s_j(\ell)|^2} = [\mathbf{Q}_{\mathbf{s}}(\ell)]_{jj}$. Thus, the PM estimate for the precision of the jth speaker writes:

$$\hat{\tau}_j(\ell) = \frac{a_{p,j}(\ell)}{b_{p,j}(\ell)} = \frac{a_j(\ell)+1}{b_j(\ell) + \widehat{|s_j(\ell)|^2}}, \quad j = 1, \ldots, J. \qquad (18)$$

Note that if $\tau_j(\ell)$ was modelled as a deterministic unknown parameter, the following point estimator was obtained $\hat{\tau}_{j,D}(\ell) = 1/\widehat{|s_j(\ell)|^2}$. Interestingly, when using a non-informative Gamma prior, i.e. $a_j(\ell) = b_j(\ell) = 0$ [5], the VEM posterior estimate in (18) coincides with the deterministic estimate.

D. E-β Step

Similarly, the posterior PDF of the noise precision writes:

$$\begin{aligned}\ln q(\beta) \propto \sum_{\ell=1}^{L} \mathbb{E}_{q(\mathbf{s}(\ell))q(\mathbf{A}(\ell))} \Big[\ln p\big(\mathbf{x}(\ell)|\mathbf{A}(\ell), \mathbf{s}(\ell), \beta\big)\Big] \\ + \ln p(\beta; c_0, d_0), \qquad (19)\end{aligned}$$

leading to a Gamma distribution: $q(\beta) = \text{Gam}(\beta; c_p, d_p)$, with

$$c_p = c_0 + NL, \qquad (20)$$

$$\begin{aligned}d_p = d_0 + \sum_{\ell=1}^{L} \Bigg(\mathbf{x}^{\text{H}}(\ell)\boldsymbol{\Gamma}^{-1}\mathbf{x}(\ell) - 2\Re\Big\{\mathbf{x}^{\text{H}}(\ell)\boldsymbol{\Gamma}^{-1}\hat{\mathbf{A}}(\ell)\hat{\mathbf{s}}(\ell)\Big\} \\ + \text{Tr}\Big[\mathbf{Q}_{\mathbf{s}}(\ell)\Big(\hat{\mathbf{A}}^{\text{H}}(\ell)\boldsymbol{\Gamma}^{-1}\hat{\mathbf{A}}(\ell) + \boldsymbol{\Psi}(\ell)\Big)\Big]\Bigg). \quad (21)\end{aligned}$$

Hence, the PM estimate for the noise precision writes:

$$\hat{\beta} = \frac{c_p}{d_p} = \frac{c_0 + NL}{d_p}. \qquad (22)$$

Treating the inverse noise power as a deterministic unknown parameter as in [15], leads to the following point estimator:

$$\begin{aligned}\hat{\beta}_D^{-1} = \frac{1}{NL} \sum_{\ell=1}^{L} \Bigg(\mathbf{x}^{\text{H}}(\ell)\boldsymbol{\Gamma}^{-1}\mathbf{x}(\ell) - 2\Re\Big\{\mathbf{x}^{\text{H}}(\ell)\boldsymbol{\Gamma}^{-1}\hat{\mathbf{A}}(\ell)\hat{\mathbf{s}}(\ell)\Big\} \\ + \text{Tr}\Big[\mathbf{Q}_{\mathbf{s}}(\ell)\Big(\hat{\mathbf{A}}^{\text{H}}(\ell)\boldsymbol{\Gamma}^{-1}\hat{\mathbf{A}}(\ell) + \boldsymbol{\Psi}(\ell)\Big)\Big]\Bigg). \quad (23)\end{aligned}$$

Note that $d_p = d_0 + NL\hat{\beta}_D^{-1}$, hence (22) becomes $\hat{\beta} = \frac{c_0+NL}{d_0+NL\hat{\beta}_D^{-1}}$. Letting c_0, d_0 approach zero, $\hat{\beta} = \hat{\beta}_D$.

E. M Step

The parameters are now estimated by maximizing the expected log-likelihood of the completed data $\mathcal{L}(\Theta) = \mathbb{E}_{q(\mathcal{H})}\left[\ln p(\mathcal{X}, \mathcal{H}; \Theta)\right]$ w.r.t. Θ. We obtain the following updates:

$$a_j(\ell) = \Psi^{-1}\left[\Psi(a_{p,j}(\ell)) + \ln\frac{b_j(\ell)}{b_{p,j}(\ell)}\right], \ \ j = 1, \ldots, J$$

$$b_j(\ell) = \frac{a_j(\ell)}{a_{p,j}(\ell)} b_{p,j}(\ell), \ \ j = 1, \ldots, J$$

$$c_0 = \Psi^{-1}\left[\Psi(c_p) + \ln\frac{d_0}{d_p}\right] \ , \ d_0 = \frac{c_0}{c_p}d_p \ , \ \boldsymbol{\mu}_{\mathbf{v}} = \hat{\mathbf{v}}(1),$$

$$\begin{aligned}\boldsymbol{\Phi}_{\mathbf{v}} = \frac{1}{L}\Bigg(\boldsymbol{\Sigma}_{\mathbf{v}}(1) + \sum_{\ell=2}^{L}\Big(\boldsymbol{Q}_{\xi,11}(\ell) - \boldsymbol{Q}_{\xi,12}(\ell) - \boldsymbol{Q}_{\xi,21}(\ell) \\ + \boldsymbol{Q}_{\xi,22}(\ell)\Big)\Bigg), \qquad (24)\end{aligned}$$

where $\Psi(\cdot)$ is the digamma function and $\boldsymbol{Q}_{\xi,np}(\ell)$, $(n,p) \in \{1,2\}$ are $NJ \times NJ$ non-overlapping subblocks of $\boldsymbol{Q}_{\xi}(\ell)$.

IV. Performance Evaluation

A. Simulation Setup

Our experiments consist of two concurrent speakers in a static scenario. Dynamic scenarios are left for future study. Room impulse responses (RIRs) were downloaded from an open-source database recorded in our lab [23]. The room dimensions are $6 \times 6 \times 2.4$ m and the reverberation time was set to $T_{60} \in \{160, 360\}$ msec. The RIRs correspond to a uniform linear array (ULA) of $N = 8$ microphones with inter-distances of 8 cm. Loudspeakers were positioned at 1 m distance from the array, at different angles in the set $\{-90°, -75°, -60°, \ldots, 90°\}$. The measured signals were generated by convolving the RIRs with TIMIT utterances [24]. The performance is evaluated by averaging over 10 mixtures of two speakers, with randomly selected utterances and angles. An artificial diffuse noise with speech-like spectrum was generated by the method described in [25], with various signal to noise ratio (SNR) levels. The sampling rate was 16 kHz and the STFT frame length was 128 ms with 75% overlap.

B. Performance Measures and Competing Methods

The source separation performance is evaluated in terms of two common objective measures, namely signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR) [26]. The reported results are average scores over the 2 sources and over the 10 experiments described above.

The performance of the proposed algorithm is compared to the method in [15], denoted as NMF-VEM.[2] For both methods, the RTF of each speaker was initialized with a simplified RTF matrix, using only the time difference of arrival (TDOA) w.r.t. the first microphone. The rest of the parameters were initialized blindly, except for the NMF parameters in the NMF-VEM, for which a semi-blind procedure was applied, where each source was corrupted by adding the other source with equal power [15]. The number of VEM iterations was 5.

We also compare the proposed method with ILRMA [19].[3] Note that ILRMA is restricted to time-invariant mixing sys-

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

[3] Available at https://github.com/d-kitamura/ILRMA

TABLE I: SDR Scores

	$T_{60} = 160$msec				$T_{60} = 360$msec			
Alg.\SNR	-5dB	0dB	5dB	10dB	-5dB	0dB	5dB	10dB
Unprocessed	-6.39	-3.19	-1.25	-0.37	-6.36	-3.16	-1.24	-0.37
NMF-VEM [15]	3.88	6.69	8.75	9.94	2.06	4.20	5.53	6.18
ILRMA [19]	-1.61	3.25	7.55	**10.60**	-2.37	1.99	5.64	**7.90**
Proposed	**6.78**	**8.75**	**9.53**	10.36	**4.95**	**6.45**	**6.94**	7.42

TABLE II: SIR Scores

	$T_{60} = 160$msec				$T_{60} = 360$msec			
Alg.\SNR	-5dB	0dB	5dB	10dB	-5dB	0dB	5dB	10dB
Unprocessed	0.11	0.11	0.12	0.12	0.10	0.11	0.11	0.11
NMF-VEM [15]	12.47	12.63	12.57	12.48	10.29	10.39	10.35	10.29
ILRMA [19]	13.52	15.79	17.55	17.77	11.61	12.94	14.15	14.54
Proposed	**23.82**	**23.74**	**20.61**	**20.74**	**21.06**	**21.24**	**18.13**	**18.19**

tems. This method was initialized with the same information that was used by the VEM-based methods, using the following initialization procedure. First, we applied principal component analysis (PCA) to reduce the over-determined problem to a determined one, as in [19]. Then, the TDOA-based RTF matrix was multiplied by the PCA projection matrix. The inverse of the resulting square matrix was used as the initial demixing filter. The number of spectral bases was 10 and the number of iterations was 100. According to our tests, when using ILRMA with a blind initialization, the optimal number of bases is 2, as was shown in [19]. However, with our informative initialization, the use of 10 bases yields superior results.

C. Results

SDR and SIR scores are presented in Tables I and II, respectively, for several SNR levels. The best results are highlighted in boldface. The proposed method outperforms the competing methods in all cases, except for the SDR at SNR $= 10$dB. Audio examples are available on our website.[4]

V. Conclusions

In this paper, we presented a Bayesian hierarchical model for blind audio source separation in a noisy environment. A fully Bayesian approach is adopted, where the speech and the noise precisions are included as part of the hidden data. The inference of the latent variables is performed using a VEM algorithm, leading to a variant of the multi-speaker MCWF for separating and enhancing the individual speakers. The speech estimate was decomposed into an LCMV beamformer followed by a variational multi-speaker postfilter. The discussion is supported by an experimental study in a room with two reverberation times and various SNR levels, demonstrating the advantage of the proposed method over competing methods.

References

[1] H. L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory.* New York, USA: Wiley, 2002.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[3] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, “The variational approximation for Bayesian inference,” *IEEE Signal Processing Magazine*, vol. 25, no. 6, pp. 131–146, 2008.

[4] T. S. Jaakkola, “Variational methods for inference and estimation in graphical models,” Ph.D. dissertation, Massachusetts Institute of Technology, 1997.

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

[6] S. Malik, J. Benesty, and J. Chen, “A Bayesian framework for blind adaptive beamforming,” *IEEE Tran. on Signal Processing*, vol. 62, no. 9, pp. 2370–2384, 2014.

[7] Y. Laufer and S. Gannot, “A Bayesian hierarchical model for speech enhancement,” in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2018, pp. 46–50.

[8] ——, “A Bayesian hierarchical model for speech enhancement with time-varying audio channel,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 225–239, 2019.

[9] ——, “A Bayesian hierarchical mixture of Gaussian model for multi-speaker DOA estimation and separation,” in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2020.

[10] D. Schmid, G. Enzner, S. Malik, D. Kolossa, and R. Martin, “Variational Bayesian inference for multichannel dereverberation and noise reduction,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 8, pp. 1320–1335, 2014.

[11] Y. Laufer and S. Gannot, “A Bayesian hierarchical model for speech dereverberation,” in *IEEE International Conference on the Science of Electrical Engineering (ICSEE)*, 2018.

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

[13] T. Virtanen, A. T. Cemgil, and S. Godsill, “Bayesian extensions to non-negative matrix factorisation for audio signal modelling,” in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2008, pp. 1825–1828.

[14] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, “An inverse-gamma source variance prior with factorized parameterization for audio source separation,” in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2016, pp. 136–140.

[15] ——, “A variational EM algorithm for the separation of time-varying convolutive audio mixtures,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 8, pp. 1408–1423, 2016.

[16] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, “Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation,” in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2018, pp. 1–5.

[17] S. E. Chazan, H. Hammer, G. Hazan, J. Goldberger, and S. Gannot, “Multi-microphone speaker separation based on deep DOA estimation,” in *Proceedings of the 27th European Signal Processing Conference (EUSIPCO)*, 2019.

[18] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, “Blind source separation exploiting higher-order frequency dependencies,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 70–79, 2006.

[19] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, “Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1626–1641, 2016.

[20] O. Schwartz, S. Gannot, and E. A. Habets, “Multispeaker LCMV beamformer and postfilter for source separation and noise reduction,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 5, pp. 940–951, 2017.

[21] G. Parisi, *Statistical field theory*. Addison-Wesley, 1988.

[22] A. Graham, *Kronecker products and matrix calculus with applications*. Courier Dover Publications, 2018.

[23] E. Hadad, F. Heese, P. Vary, and S. Gannot, “Multichannel audio database in various acoustic environments,” in *14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014, pp. 313–317.

[24] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, “DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM.” NIST, Gaithersburg, MD, USA, speech disc 1-1.1, 1993.

[25] E. A. Habets and S. Gannot, “Generating sensor signals in isotropic noise fields,” *The Journal of the Acoustical Society of America*, vol. 122, no. 6, pp. 3464–3470, 2007.

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

[4] http://www.eng.biu.ac.il/gannot/speech-enhancement/