A BAYESIAN HIERARCHICAL MODEL FOR SPEECH ENHANCEMENT

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 addresses the problem of blind adaptive beamforming using a hierarchical Bayesian model. Our probabilistic approach relies on a Gaussian prior for the speech signal and a Gamma hyperprior for the speech precision, combined with a multichannel linear-Gaussian state-space model for the possibly time-varying acoustic channel. Furthermore, we assume a Gamma prior for the ambient noise precision. We present a variational Expectation-Maximization (VEM) algorithm that employs a variant of multi-channel Wiener filter (MCWF) to estimate the sound source and a Kalman smoother to estimate the acoustic channel of the room. It is further shown that the VEM speech estimator can be decomposed into two stages: A multichannel minimum variance distortionless response (MVDR) beamformer and a subsequent single-channel variational postfilter. The proposed algorithm is evaluated in terms of speech quality, for a static scenario with recorded room impulse responses (RIRs). It is shown that a significant improvement is obtained with respect to the noisy signal, and that the proposed algorithm outperforms a baseline algorithm. In terms of channel alignment, a superior channel estimate is demonstrated compared to the causal Kalman filter.

Index Terms— Adaptive beamforming, Kalman smoother, variational EM.

1. INTRODUCTION

Speech enhancement deals with the reconstruction of a speech source from microphone signals recorded in a noisy and reverberant environment, with commercial applications in devices as mobile phones, hands-free systems or hearing aids. Multichannel Wiener beamformer is a common approach that exploits the spatial diversity of the acoustic channels for enhancing the desired source while suppressing sounds from other directions.

The design of beamformers requires that certain parameters are available for their computation, namely the relative transfer function (RTF) of the speaker and the covariance matrices of the background noise and the speaker [1]. Numerous methods exist for estimating these parameters. Some approaches are based on speech presence probability (SPP), by first determining the time-frequency bins dominated by either speech or noise, and then estimating independently the model parameters [2–5]. Other approaches jointly estimate all parameters according to some criterion, such as maximum likelihood (ML) or maximum a posteriori (MAP) criteria [6–8]. When the resulting optimization problems cannot be solved in closed-form, the EM algorithm [9, 10] is a solution that breaks down the problem into the signal estimate and the parameters estimate that are iteratively solved.

The Bayesian approach defines prior distributions over each parameter, and thus provides an elegant way to explore uncertainty in the model and to incorporate prior knowledge into the learning process. Hierarchical Bayesian models, where observations are modeled conditionally on certain parameters, which themselves are given a probabilistic specification in terms of further parameters, are very useful since they allow to define structured models, with dependencies among parameters. However, to apply the EM algorithm we must know the posterior distribution, which might be intractable in complex Bayesian models. The variational methodology [11–15] alleviates this drawback of the EM procedure, by approximating the posterior distribution. This approach is more robust and less sensitive to local maxima and overfitting [1, 13], as the inference process relies on the entire probability mass rather than just point estimates. Furthermore, it allows incorporation of a priori statistical belief.

As proposed in [15], we model in this paper the speech signal as a Guassian distributed and the RTF by means of first-order Markov model, to account for time-varying channel. By modeling the speech signal and the channel as latent variables, their posterior distribution are jointly estimated in the E-step. However, the authors do not adopt a fully Bayesian model, since the covariances of the source and the ambient noise are still treated as unknown deterministic parameters, for which point estimates are computed in the M-step. In addition, the RTF was estimated using the casual Kalman filter, i.e. using only past and present observations. In this paper, we propose to extend the probabilistic model towards the fully Bayesian model. We introduce a hierarchical model which is based on a Gaussian prior for the speech signal and a Gamma hyperprior for the speech precision. Furthermore, we assume a Gamma prior for the noise precision. This way, the precisions are also modeled as latent random variables, for which posterior distributions are inferred in the E-step. We derive a VEM algorithm in which a *Kalman smoother* is used for the inference of the RTF, as proposed in [16]. This way we exploit all the available data to estimate the RTF.[1]

In [17, 18], the authors show that the MCWF estimator of a single speech signal can be decomposed into two stages. First, a multichannel MVDR beamformer that exploits the spatial diversity to preserve a distortionless response while minimizing the output noise power; and second, a subsequent single-channel Wiener postfilter that reduces the residual noise at the output of the first stage by using the covariance of the speech and the residual noise. Inspired by this decomposition, we show that the VEM speech estimator can also be decomposed into an MVDR beamformer, followed by a variational postfilter, which takes into account the uncertainty in RTF estimate.

2. MULTICHANNEL ACOUSTIC STATE-SPACE MODEL

2.1. Signal Model

The model is formulated in the short-time Fourier transform (STFT) domain, where $k \in [1, K]$ denotes the frequency band, and $\ell \in [1, L]$ denotes the time frame. Consider a speech signal

[1]The authors thank Dr. Dionyssos Kounades-Bastian from INRIA, Grenoble, France for his constructive comments.

received by N microphones, in a noisy and reverberant acoustic environment. The N-channel observation signal $\mathbf{x}(\ell,k) = [X_1(\ell,k),\cdots,X_N(\ell,k)]^T$ writes

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

where $S(\ell,k)$ is the echoic speech signal as received by the first microphone (that was arbitrary chosen as the reference microphone), modeled as a zero-mean Gaussian [19] with time-varying precision $p(S(\ell,k)|\tau(\ell,k)) = \mathcal{N}_c(S(\ell,k);0,\tau^{-1}(\ell,k))$, $\mathbf{a}(\ell,k) = [1,A_2(\ell,k),\cdots,A_N(\ell,k)]^T$ is the RTF vector and $\mathbf{u}(\ell,k) = [U_1(\ell,k),\cdots,U_N(\ell,k)]^T$ is a zero-mean multivariate Gaussian ambient noise with $p(\mathbf{u}(\ell,k)|\mathbf{\Phi}_u(k)) = \mathcal{N}_c(\mathbf{u}(\ell,k);\mathbf{0},\mathbf{\Phi}_u(k))$. The noise covariance matrix is assumed to be time-invariant, modeled by a full-rank coherence matrix multiplied by the noise power $\mathbf{\Phi}_u(k) = \beta^{-1}(k)\mathbf{\Gamma}(k)$, where $\beta(k)$ is the time-invariant inverse noise power and $\mathbf{\Gamma}(k)$ is the time-invariant spatial coherence matrix. In this work (as in [20]), a prior knowledge about the spatial characteristics of the noise is assumed. However, the power of the noise is an unknown parameter. We assume that $\mathbf{\Gamma}(k)$ can be modeled using a spatially homogeneous and spherically isotropic sound field [21, 22] plus diagonal loading: $\Gamma_{ij}(k) = \text{sinc}\left(\frac{2\pi f_s k}{K}\frac{d_{ij}}{c}\right) + \epsilon\delta_{i-j}$, where $\text{sinc}(x) = \sin(x)/x$, d_{ij} is the inter-distance between microphones i and j, and ϵ is the level of diagonal loading. The conditional data distribution is therefore given by $p(\mathbf{x}(\ell,k)|S(\ell,k),\mathbf{a}(\ell,k),\beta(k)) = \mathcal{N}_c(\mathbf{x}(\ell,k);S(\ell,k)\mathbf{a}(\ell,k),\beta^{-1}(k)\mathbf{\Gamma}(k))$.

To account for time-varying channel, we model the RTF as a set of temporally-linked continuous latent variables, parameterized with a first-order linear dynamical system (LDS), as in [16]:

$$p(\mathbf{a}(1,k)) = \mathcal{N}_c(\mathbf{a}(1,k);\boldsymbol{\mu}_a(k),\mathbf{\Phi}_a(k)), \tag{2}$$
$$p(\mathbf{a}(\ell,k)|\mathbf{a}(\ell-1,k)) = \mathcal{N}_c(\mathbf{a}(\ell,k);\mathbf{a}(\ell-1,k),\mathbf{\Phi}_a(k)), \tag{3}$$

where $\boldsymbol{\mu}_a(k) \in \mathbb{C}^N$ is the prior mean and $\mathbf{\Phi}_a(k) \in \mathbb{C}^{N\times N}$ is the evolution covariance matrix. For brevity, $\mathbf{a}(1{:}L,k) = \{\mathbf{a}(\ell,k)\}_{\ell=1}^L$ denotes the entire sequence of RTFs at frequency k. Also, let $\mathcal{X} = \{\mathbf{x}(\ell,k)\}_{\ell,k=1}^{L,K}$ denotes the set of observations.

2.2. Conjugate Priors

Our hierarchical model is established by introducing priors for the precision of the speaker and the noise. The conjugate prior for the precision of a univariate Gaussian is the Gamma probability [12]. Hence, the prior for the speech precision is given by:

$$p(\tau(\ell,k)) = \text{Gam}(\tau(\ell,k);a_0(\ell,k),b_0(\ell,k)). \tag{4}$$

Each time-frequency (TF) bin is modeled with distinct shape and rate, to allow the flexibility of modeling local characteristics of the speech signal. This choice of two-level hierarchical prior can be further justified by the fact that by marginalizing over the precision $\tau(\ell,k)$, the true prior distribution of the speech signal turns to be a Student's t-distribution [12]. The heavy tailed Student-t prior favors sparse models, i.e. models with a few nonzero parameters, and thus was proposed to model speech coefficients [23], which have sparse distribution in the TF domain. Similarly, we assume a Gamma distribution as prior for the time-invariant noise precision:

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

The graphical model of the proposed probabilistic model is shown in Fig. 1.

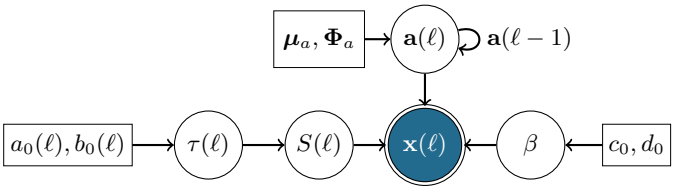

Fig. 1: Graphical model (Frequency index is omitted).

3. VEM FOR BEAMFORMING

In this work, the set of hidden variables consists of $\mathcal{H} = \{S(\ell,k),\mathbf{a}(\ell,k),\tau(\ell,k),\beta(k)\}_{\ell,k=1}^{L,K}$. The parameter set consists of $\theta = \{a_0(\ell,k),b_0(\ell,k),c_0(k),d_0(k),\boldsymbol{\mu}_a(k),\mathbf{\Phi}_a(k)\}_{\ell,k=1}^{L,K}$. Bayesian inference requires the computation of the posterior distribution $p(\mathcal{H}|\mathcal{X};\theta) = \frac{p(\mathcal{X},\mathcal{H};\theta)}{p(\mathcal{X};\theta)}$. Using Bayes' rule, the complete-data distribution writes

$$\begin{aligned} p(\mathcal{X},\mathcal{H};\theta) = \prod_{\ell,k=1}^{L,K} \Big(& p\Big(\mathbf{x}(\ell,k)|S(\ell,k),\mathbf{a}(\ell,k),\beta(k)\Big) \\ & \times p\Big(S(\ell,k)|\tau(\ell,k)\Big)p\Big(\tau(\ell,k);a_0(\ell,k),b_0(\ell,k)\Big)\Big) \\ & \times \prod_{k=1}^{K}\Big(p\Big(\beta(k);c_0(k),d_0(k)\Big)p\Big(\mathbf{a}(1,k);\boldsymbol{\mu}_a(k),\mathbf{\Phi}_a(k)\Big) \\ & \times \prod_{\ell=2}^{L} p\Big(\mathbf{a}(\ell,k);\mathbf{a}(\ell-1,k),\mathbf{\Phi}_a(k)\Big)\Big). \end{aligned} \tag{6}$$

However, the likelihood $p(\mathcal{X};\theta) = \int p(\mathcal{X},\mathcal{H};\theta)d\mathcal{H}$ cannot be computed analytically from (6), hence the posterior $p(\mathcal{H}|\mathcal{X};\theta)$ cannot be expressed in closed-form and exact inference becomes intractable. To tackle this problem, we resort to the approximate variational inference methodology, which circumvents this difficulty by approximating the posterior $q(\mathcal{H}) \approx p(\mathcal{H}|\mathcal{X};\theta)$. According to the *mean field theory* assumption [24, 25], we assume that the speech signal, RTF, speech precision and noise precision are conditionally independent given the observations. Therefore, the approximate posterior distribution naturally factorizes as:

$$q(\mathcal{H}) = \prod_{\ell,k=1}^{L,K}\Big(q(S(\ell,k))q(\tau(\ell,k))\Big)\prod_{k=1}^{K}\Big(q(\beta(k))q(\mathbf{a}(1{:}L,k))\Big). \tag{7}$$

Given a factorization of $q(\mathcal{H})$ over a partition of latent variables, the optimal marginal posterior distribution of a subset $\mathcal{H}_0 \subseteq \mathcal{H}$ can be computed [12] in the E-Step with:

$$\ln q(\mathcal{H}_0) = \mathbb{E}_{q(\mathcal{H}/\mathcal{H}_0)}[\ln p(\mathcal{X},\mathcal{H};\theta)] + \text{const}, \tag{8}$$

where $q(\mathcal{H}/\mathcal{H}_0)$ is the approximation of the joint posterior distribution of all hidden variables, except the subset $\mathcal{H}_0$. Subsequently, $q(\mathcal{H})$ can be inferred for each $\mathcal{H}_0 \subset \mathcal{H}$. Once we have the posterior distributions of the variables in $\mathcal{H}$, the expected complete-data log-likelihood $\mathcal{L}(\theta) = \mathbb{E}_{q(\mathcal{H})}[\ln p(\mathcal{X},\mathcal{H};\theta)]$ is maximized with respect to the parameters in the M-Step. A detailed derivation of the algorithm is omitted due to space constraints. In the following, the frequency index k is omitted for brevity whenever possible.

3.1. E-S Step

The approximate posterior distribution of the source is obtained from (6) and (8) by keeping only the terms that depend on $S(\ell)$:

$$\ln q(S(\ell)) = \mathbb{E}_{q(\mathbf{a}(\ell))q(\tau(\ell))q(\beta)}[\ln p(\mathbf{x}(\ell)|S(\ell),\mathbf{a}(\ell),\beta) + \ln p(S(\ell)|\tau(\ell))] + \text{const}, \quad (9)$$

which can be shown to be a Gaussian distribution: $q\left(S(\ell)\right) = \mathcal{N}_c\left(S(\ell);\hat{S}(\ell),\Sigma_s(\ell)\right)$, with

$$\hat{S}(\ell) = \frac{\hat{\mathbf{a}}^H(\ell)\hat{\beta}\mathbf{\Gamma}^{-1}\mathbf{x}(\ell)}{\hat{\mathbf{a}}^H(\ell)\hat{\beta}\mathbf{\Gamma}^{-1}\hat{\mathbf{a}}(\ell) + \text{tr}(\mathbf{\Sigma}_a(\ell)\hat{\beta}\mathbf{\Gamma}^{-1}) + \hat{\tau}(\ell)}, \quad (10)$$

$$\Sigma_s(\ell) = \left(\hat{\mathbf{a}}^H(\ell)\hat{\beta}\mathbf{\Gamma}^{-1}\hat{\mathbf{a}}(\ell) + \text{tr}(\mathbf{\Sigma}_a(\ell)\hat{\beta}\mathbf{\Gamma}^{-1}) + \hat{\tau}(\ell)\right)^{-1}, \quad (11)$$

where $\hat{\mathbf{a}}(\ell), \mathbf{\Sigma}_a(\ell), \hat{\beta}, \hat{\tau}(\ell)$ are posterior statistics that will be defined in Sections 3.2-3.4. The speech signal can be estimated by the posterior mean, namely $\hat{S}(\ell)$, with the variance $\Sigma_s(\ell)$. This speech estimator resembles the form of the MCWF [26, 27]:

$$\hat{S}_{\text{MCWF}}(\ell) = \frac{\hat{\mathbf{a}}^H(\ell)\hat{\beta}\mathbf{\Gamma}^{-1}\mathbf{x}(\ell)}{\hat{\mathbf{a}}^H(\ell)\hat{\beta}\mathbf{\Gamma}^{-1}\hat{\mathbf{a}}(\ell) + \hat{\tau}(\ell)}, \quad (12)$$

except the term $\text{tr}(\mathbf{\Sigma}_a(\ell)\hat{\beta}\mathbf{\Gamma}^{-1})$. In a similar way to the MCWF decomposition [17, 18], $\hat{S}(\ell)$ can be decomposed as

$$\hat{S}(\ell) = \underbrace{\frac{\hat{\mathbf{a}}^H(\ell)\hat{\beta}\mathbf{\Gamma}^{-1}\hat{\mathbf{a}}(\ell)}{\hat{\mathbf{a}}^H(\ell)\hat{\beta}\mathbf{\Gamma}^{-1}\hat{\mathbf{a}}(\ell) + \text{tr}(\mathbf{\Sigma}_a(\ell)\hat{\beta}\mathbf{\Gamma}^{-1}) + \hat{\tau}(\ell)}}_{H(\ell)} \times \underbrace{\frac{\hat{\mathbf{a}}^H(\ell)\mathbf{\Gamma}^{-1}}{\hat{\mathbf{a}}^H(\ell)\mathbf{\Gamma}^{-1}\hat{\mathbf{a}}(\ell)}}_{\mathbf{w}^H_{\text{MVDR}}(\ell)}\mathbf{x}(\ell). \quad (13)$$

Due to the RTF uncertainty, the speech signal at the output of the MVDR stage may be distorted. The MCWF treats the RTF estimator $\hat{\mathbf{a}}(\ell)$ as a point estimator, and thus ignores this uncertainty. In contrast, the VEM estimator treats the RTF as a latent variable and considers its posterior distribution, which captures the uncertainty about the parameter estimate. Therefore, it includes $H(\ell)$ as a postfilter that takes into account the uncertainty level, expressed by $\mathbf{\Sigma}_a(\ell)$, and weights accordingly the single channel at the MVDR output. When $\mathbf{\Sigma}_a(\ell) \to \mathbf{0}$, $\hat{S}(\ell)$ reduces to $\hat{S}_{\text{MCWF}}(\ell)$.

3.2. E-a Step

The joint posterior of the RTF sequence is similarly obtained by keeping only the terms that depend on $\mathbf{a}(1{:}L)$:

$$\ln q(\mathbf{a}(1{:}L)) = \sum_{\ell=1}^{L}\mathbb{E}_{q(S(\ell))q(\beta)}\left[\ln p(\mathbf{x}(\ell)|S(\ell),\mathbf{a}(\ell),\beta)\right] + \ln p(\mathbf{a}(1{:}L)) + \text{const}. \quad (14)$$

It can be shown that the first term reduces to a Gaussian distribution, and thus the posterior distribution is an LDS along the frames, as in [16]. Hence, the marginal posterior distribution of each frame is also a Gaussian distribution, $q(\mathbf{a}(\ell)) = \mathcal{N}_c(\mathbf{a}(\ell);\hat{\mathbf{a}}(\ell),\mathbf{\Sigma}_a(\ell))$, which can be recursively calculated using the *Kalman smoother* [12]. The RTF can therefore be estimated by the posterior mean, namely $\hat{\mathbf{a}}(\ell)$, with uncertainty $\mathbf{\Sigma}_a(\ell)$. The pair-wise joint posterior distribution of two successive frames will be required to update $\mathbf{\Phi}_a$ in Section 3.5. Marginalizing out all other frames in (14) results with a Gaussian distribution, $q(\mathbf{a}(\ell),\mathbf{a}(\ell-1)) = \mathcal{N}_c\left(\left[\mathbf{a}(\ell)^T,\mathbf{a}(\ell-1)^T\right]^T;\mathbf{a}_\xi(\ell),\mathbf{\Sigma}_\xi(\ell)\right)$. For the sake of clarity, note that $\hat{\mathbf{a}}(\ell) \in \mathbb{C}^N, \mathbf{\Sigma}_a(\ell) \in \mathbb{C}^{N\times N}, \mathbf{a}_\xi(\ell) \in \mathbb{C}^{2N}$ and $\mathbf{\Sigma}_\xi(\ell) \in \mathbb{C}^{2N\times 2N}$ are the mean and the covariance of the marginal and pair-wise posterior distributions, respectively. The second-order joint posterior moment is defined as $\boldsymbol{Q}(\ell) = \mathbf{\Sigma}_\xi(\ell) + \mathbf{a}_\xi(\ell)\mathbf{a}_\xi^H(\ell)$.

3.3. E-τ Step

Using (6) and (8), the posterior distribution of the speech precision writes:

$$\ln q(\tau(\ell)) = \mathbb{E}_{q(S(\ell))}[\ln p(S(\ell)|\tau(\ell))] + \ln p(\tau(\ell);a_0(\ell),b_0(\ell)) + \text{const}, \quad (15)$$

which can be shown to be a Gamma distribution: $q(\tau(\ell)) = \text{Gam}(\tau(\ell);a_p(\ell),b_p(\ell))$, with

$$a_p(\ell) = a_0(\ell) + 1\ ,\ b_p(\ell) = b_0(\ell) + \widehat{|S(\ell)|^2}. \quad (16)$$

We therefore obtain the posterior mean estimate for the source precision as:

$$\hat{\tau}(\ell) = \frac{a_p(\ell)}{b_p(\ell)} = \frac{a_0(\ell)+1}{b_0(\ell) + \widehat{|S(\ell)|^2}}, \quad (17)$$

where $a_0(\ell), b_0(\ell)$, the prior parameters, are updated in the M-Step. It should be noted that treating the speech precision as an unknown deterministic parameter as in [15], leads to the following point estimator $\hat{\tau}_D(\ell) = 1/\widehat{|S(\ell)|^2}$. It is instructive to relate the variational solution to the deterministic one. To do this, consider the marginal case where the parameters are fixed to very small values, i.e. $a_0(\ell) = b_0(\ell) = 0$, in which the VEM posterior estimator coincides with the deterministic estimator. This equivalence can be explained by the fact that a *non-informative prior* is obtained for the Gamma distribution as the special case $a_0(\ell) = b_0(\ell) = 0$, since it corresponds to the limit of an infinitely broad prior [12].

3.4. E-β Step

Similarly, the posterior distribution of the noise precision writes:

$$\ln q(\beta) = \sum_{\ell=1}^{L}\mathbb{E}_{q(S(\ell))q(\mathbf{a}(\ell))}\left[\ln p(\mathbf{x}(\ell)|S(\ell),\mathbf{a}(\ell),\beta)\right] + \ln p(\beta;c_0,d_0) + \text{const}, \quad (18)$$

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

$$c_p = c_0 + NL, \quad (19)$$

$$d_p = d_0 + \sum_{\ell=1}^{L}\Bigg(\mathbf{x}^H(\ell)\mathbf{\Gamma}^{-1}\mathbf{x}(\ell) - \mathbf{x}^H(\ell)\mathbf{\Gamma}^{-1}\hat{S}(\ell)\hat{\mathbf{a}}(\ell) - \hat{\mathbf{a}}^H(\ell)\hat{S}^*(\ell)\mathbf{\Gamma}^{-1}\mathbf{x}(\ell) + \widehat{|S(\ell)|^2}\left(\hat{\mathbf{a}}^H(\ell)\mathbf{\Gamma}^{-1}\hat{\mathbf{a}}(\ell) + \text{tr}(\mathbf{\Sigma}_a(\ell)\mathbf{\Gamma}^{-1})\right)\Bigg). \quad (20)$$

We therefore obtain the posterior mean estimate for the noise precision as:

$$\hat{\beta} = \frac{c_p}{d_p} = \frac{c_0+NL}{d_p}. \quad (21)$$

Treating the inverse noise power as an unknown deterministic parameter leads to the following point estimator:

$$\hat{\beta}_D^{-1} = \frac{1}{NL}\sum_{\ell=1}^{L}\Bigg(\mathbf{x}^H(\ell)\mathbf{\Gamma}^{-1}\mathbf{x}(\ell) - \mathbf{x}^H(\ell)\mathbf{\Gamma}^{-1}\hat{S}(\ell)\hat{\mathbf{a}}(\ell) - \hat{\mathbf{a}}^H(\ell)\hat{S}^*(\ell)\mathbf{\Gamma}^{-1}\mathbf{x}(\ell) + \widehat{|S(\ell)|^2}\Big(\hat{\mathbf{a}}^H(\ell)\mathbf{\Gamma}^{-1}\hat{\mathbf{a}}(\ell) + \text{tr}\left(\mathbf{\Sigma}_a(\ell)\mathbf{\Gamma}^{-1}\right)\Big)\Bigg). \quad (22)$$

Note that $d_p = d_0 + NL\hat{\beta}_D^{-1}$, hence (21) becomes $\hat{\beta} = \frac{c_0+NL}{d_0+NL\hat{\beta}_D^{-1}}$. Letting c_0, d_0 approach zero, $\hat{\beta} = \hat{\beta}_D$.

3.5. M Step

Once we have the posterior distributions of the hidden variables, the expected complete-data log-likelihood $\mathcal{L}(\theta) = \mathbb{E}_{q(\mathcal{H})}\left[\ln p(\mathcal{X}, \mathcal{H}; \theta)\right]$ is maximized with respect to the prior parameters. We obtain the following closed-form expressions:

$$a_0(\ell) = \Psi^{-1}\left(\Psi(a_p(\ell)) + \ln\frac{b_0(\ell)}{b_p(\ell)}\right),\ b_0(\ell) = \frac{a_0(\ell)}{a_p(\ell)}b_p(\ell),$$
$$c_0 = \Psi^{-1}\left(\Psi(c_p) + \ln\frac{d_0}{d_p}\right), \quad d_0 = \frac{c_0}{c_p}d_p,$$
$$\boldsymbol{\mu}_a = \hat{\mathbf{a}}(1),$$
$$\mathbf{\Phi}_a = \frac{1}{L}\Bigg(\mathbf{\Sigma}_a(1) + \sum_{\ell=2}^{L}\Big(\boldsymbol{Q}_{11}(\ell) - \boldsymbol{Q}_{12}(\ell) - \boldsymbol{Q}_{21}(\ell) + \boldsymbol{Q}_{22}(\ell)\Big)\Bigg), \quad (23)$$

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

4. PERFORMANCE EVALUATION

4.1. Simulation Setup

We assess the performance of the proposed algorithm for a static scenario, and compare it with [15]. For the simulations, RIRs were downloaded from an open-source database recorded in our lab [28]. The room dimensions were $6 \times 6 \times 2.4$ m and the tested reverberation times were set to $T_{60} = [0.16, 0.36]$ sec. A loudspeaker was positioned at a distance of 1 m from uniform linear array (ULA) with $N = 8$ microphones, at angle 90° (endfire). The inter-distances between the microphones were 8 cm. Utterances from five female and five male speakers were drawn from the TIMIT database [29] (each sentence being 3-5 sec long), then convolved with the RIRs. For the additive noise, we used a stationary noise signal with speech-like spectrum from NOISEX-92 database [30], and applied the method described in [22] to produce diffuse noise field, with various signal to noise ratio (SNR) levels. The sampling frequency was 8 kHz, the frame length of the STFT was 128 ms with 32 ms between successive time frames. The number of VEM iterations was fixed to 15.

4.2. Performance Measures

The speech enhancement performance is evaluated in terms of two common objective measures, namely perceptual evaluation of speech quality (PESQ) [31], and log-spectral distance (LSD). Low LSD indicates a high speech quality. The quality of the RTF estimate was assessed with the normalized projection misalignment (NPM) measure [32]. All the measures were computed by averaging the results obtained using the 10 speakers.

Table 1: PESQ Scores for T_{60} of 0.16 s (Left) and 0.36 s (Right)

	$T_{60} = 0.16$sec				$T_{60} = 0.36$sec			
Alg.\SNR	−5dB	0dB	5dB	10dB	−5dB	0dB	5dB	10dB
Unprocessed	1.3	1.5	1.8	2.2	1.4	1.6	1.9	2.4
Baseline [15]	2.3	2.6	2.9	3.2	2.4	2.7	3.0	3.3
Proposed	**2.6**	**3.0**	**3.4**	**3.7**	**2.7**	**3.0**	**3.4**	**3.5**
Oracle MCWF	3.5	3.7	3.8	3.9	3.6	3.6	3.7	3.7

Table 2: LSD Results for T_{60} of 0.16 s (Left) and 0.36 s (Right)

	$T_{60} = 0.16$sec				$T_{60} = 0.36$sec			
Alg.\SNR	−5dB	0dB	5dB	10dB	−5dB	0dB	5dB	10dB
Unprocessed	12.2	9.2	6.6	4.5	12.4	9.6	7.0	4.8
Baseline [15]	4.9	4.5	4.1	3.8	5.7	5.3	4.9	4.7
Proposed	**3.8**	**3.5**	**3.5**	**3.5**	**4.6**	**4.4**	**4.3**	**4.3**
Oracle MCWF	1.9	1.4	1.0	0.8	2.7	2.3	2.1	2.0

Table 3: NPM Results for T_{60} of 0.16 s (Left) and 0.36 s (Right)

	$T_{60} = 0.16$sec				$T_{60} = 0.36$sec			
Alg.\SNR	−5dB	0dB	5dB	10dB	−5dB	0dB	5dB	10dB
Baseline [15]	-2.3	-3.0	-3.6	-4.2	-2.1	-2.7	-3.1	-3.5
Proposed	**-6.3**	**-7.9**	**-9.4**	**-9.9**	**-5.9**	**-7.3**	**-8.8**	**-9.1**

4.3. Results

PESQ and LSD scores are presented in Tables 1 and 2, respectively, for several SNR levels. NPM results are presented in Table 3. The advantage of using the proposed method is demonstrated for all SNR levels. To demonstrate the effectiveness of the proposed algorithm, the results of an MCWF with true parameters are also presented. We refer to this algorithm as the *oracle* algorithm, since it knows a priori the RTF and the covariances of the speech and the noise signals, information that is not available to the fully blind VEM algorithm. The results obtained for this oracle method can be considered as the best achievable results. Audio examples are available on our website.[2]

5. CONCLUSIONS

In this paper, we have presented a hierarchical Bayesian model for blind beamforming in a multichannel audio scenario. The model extends [15] to include the speech precision and the ambient noise precision as part of the hidden data. The inference of the hidden variables is performed using a variational EM algorithm, leading to a variant of MCWF for estimating the source and a Kalman smoother for the acoustic channel. The speech estimator was decomposed into an MVDR beamformer followed by a variational postfilter. The proposed method was tested in a room with a reverberation time of 0.16 sec and 0.36 sec for several SNR levels. In terms of the objective performance measures as well as an informal listening test, the proposed method outperforms the baseline method for the considered scenarios. In terms of NPM, the proposed algorithm offers improvement by 4 to 6 dB over the causal Kalman filter in [15].

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

6. REFERENCES

[1] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Tran. on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[2] M. Souden, J. Chen, J. Benesty, and S. Affes, "Gaussian model-based multichannel speech presence probability," *IEEE Tran. on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 1072–1077, 2010.

[3] M. Taseska and E. A. Habets, "Spotforming: Spatial filtering with distributed arrays for position-selective sound acquisition," *IEEE/ACM Tran. on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1291–1304, 2016.

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

[5] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Tran. on Speech and Audio Processing*, vol. 12, no. 5, pp. 451–459, 2004.

[6] C. Févotte and J.-F. Cardoso, "Maximum likelihood approach for blind audio source separation using time-frequency Gaussian source models," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2005, pp. 78–81.

[7] E. Vincent, S. Arberet, and R. Gribonval, "Underdetermined instantaneous audio source separation via local Gaussian modeling." in *International Conference on Independent Component Analysis and Signal Separation (ICA)*, 2009, pp. 775–782.

[8] N. Q. Duong, E. Vincent, and R. Gribonval, "Underdetermined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Tran. on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[9] 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.

[10] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal processing magazine*, vol. 13, no. 6, pp. 47–60, 1996.

[11] H. Attias, J. C. Platt, A. Acero, and L. Deng, "Speech denoising and dereverberation using probabilistic models," in *Advances in neural information processing systems*, 2001, pp. 758–764.

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

[13] 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.

[14] D. Schmid, G. Enzner, S. Malik, D. Kolossa, and R. Martin, "Variational Bayesian inference for multichannel dereverberation and noise reduction," *IEEE/ACM Tran. on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 8, pp. 1320–1335, 2014.

[15] 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.

[16] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, "A variational EM algorithm for the separation of time-varying convolutive audio mixtures," *IEEE/ACM Tran. on Audio, Speech, and Language Processing*, vol. 24, no. 8, pp. 1408–1423, 2016.

[17] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone arrays*. Springer, 2001, pp. 39–60.

[18] R. Balan and J. Rosca, "Microphone array speech enhancement by Bayesian estimation of spectral amplitude and phase," in *IEEE Sensor Array and Multichannel Signal Processing Workshop*, 2002, pp. 209–213.

[19] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Tran. on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[20] O. Schwartz, Y. Dorfan, M. Taseska, E. A. Habets, and S. Gannot, "Doa estimation in noisy environment with unknown noise power using the EM algorithm," in *Hands-free Speech Communications and Microphone Arrays (HSCMA)*, 2017, pp. 86–90.

[21] N. Dal Degan and C. Prati, "Acoustic noise analysis and speech enhancement techniques for mobile radio applications," *Signal Processing*, vol. 15, no. 1, pp. 43–56, 1988.

[22] 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.

[23] C. Fevotte and S. J. Godsill, "A Bayesian approach for blind separation of sparse sources," *IEEE Tran. on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2174–2188, 2006.

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

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

[26] H. L. Van Trees, *Optimum array processing: Part IV of detection, estimation and modulation theory*. Wiley, 2002.

[27] O. Schwartz, S. Gannot, and E. A. Habets, "Multi-microphone speech dereverberation and noise reduction using relative early transfer functions," *IEEE/ACM Tran. on Audio, Speech and Language Processing*, vol. 23, no. 2, pp. 240–251, 2015.

[28] 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.

[29] 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 speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[30] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[31] I.-T. Recommendation, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[32] D. R. Morgan, J. Benesty, and M. M. Sondhi, "On the evaluation of estimated impulse responses," *IEEE Signal processing letters*, vol. 5, no. 7, pp. 174–176, 1998.