A Bayesian Hierarchical Model for Speech Enhancement with Time-Varying Audio Channel

Yaron Laufer, *Student Member, IEEE*, and Sharon Gannot, *Senior Member, IEEE*

***Abstract*—We present a fully Bayesian hierarchical approach for multichannel speech enhancement with time-varying audio channel. 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 acoustic channel. Furthermore, we assume a Wishart prior for the noise precision matrix. We derive a variational Expectation-Maximization (VEM) algorithm which uses a variant of multichannel Wiener filter (MCWF) to infer the sound source and a Kalman smoother to infer the acoustic channel. It is further shown that the VEM speech estimator can be recast as a multichannel minimum variance distortionless response (MVDR) beamformer followed by a single-channel variational postfilter. The proposed algorithm was evaluated using both simulated and real room environments with several noise types and reverberation levels. Both static and dynamic scenarios are considered. In terms of speech quality, it is shown that a significant improvement is obtained with respect to the noisy signal, and that the proposed method outperforms a baseline algorithm. In terms of channel alignment and tracking ability, a superior channel estimate is demonstrated.**

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

I. Introduction

Background noise significantly degrades the quality of the speech signal and its intelligibility. Speech enhancement aims at recovering a speech source from microphone signals recorded in a noisy and reverberant environment, with a vast amount of applications as cellular phones, hearing aids, humanoid robots and autonomous systems. Personal assistants and navigation systems also take advantage of speech enhancement methods to obtain noise-robust automatic speech recognition. Speech enhancement algorithms have therefore attracted a great deal of interest in recent years.

In this paper, we address the challenging scenario of *time-varying* audio channel, which arises when the acoustic impulse responses (AIRs) between the speaker and the microphones change over time. Time-varying acoustic channel can describe moving speaker, moving microphones, or other environmental changes.

Array beamforming is a common multichannel method that exploits spatial diversity, i.e. the different acoustic properties between channels, for enhancing desired source while suppressing sounds from other directions. In this work, the steering vector used for beamforming is the relative transfer function (RTF) [1]. Design of beamformers requires the estimation of several model parameters. For instance, multichannel Wiener filter (MCWF) beamformers require the RTF of the speaker, its respective variance (or alternatively, the speaker's entire covariance matrix), and the covariance matrix of the noise [2]. Numerous methods exist for estimating these parameters. Some approaches utilize the speech presence probability (SPP), by first determining the time-frequency bins dominated by either speech or noise, and then estimating independently the model parameters [3].

In this context, the problem of RTF identification can be tackled with various approaches. In [1], [4], the nonstationarity of speech signal is exploited. A subspace-based approach was proposed in [5], where a generalized eigenvalue decomposition (GEVD) of the covariance matrix was used to extract an RTF estimator. However, many realistic scenarios are dynamic. The application of standard beamforming techniques to dynamic scenarios is not an easy task, since the need to track changes in the RTF enforces short time frames, which in turn reduces the number of data points for the RTF estimation. A modified version of the nonstationarity-based system identification algorithm was introduced in [6], based on hypothesis testing. A weighted least-squares (WLS) estimator, which incorporates the SPP decisions, was proposed in [7].

Rather than estimating independently the parameters as suggested by the SPP-based methods, other methods jointly estimate all model parameters by optimizing some criterion, such as maximum likelihood (ML) or maximum a posteriori (MAP) criteria [8]–[10]. When the resulting optimization problem cannot be solved directly, a common solution is the Expectation-Maximization (EM) algorithm [11], [12], which 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 are very useful, since they allow to define complex structured models. These models use a multi-level modeling to capture important dependencies among parameters. The resulting complex structure is of particular interest when the parameter of interest is not related directly to the observation but rather to another parameter, which itself is related to the observation. However, to apply the EM algorithm we must know the posterior distribution, which might be intractable in complex Bayesian models. To alleviate this drawback, one may resort to approximate methods, which are broadly divided into two classes, namely stochastic approximations and deterministic approximations [13]. Stochastic techniques are based on numerical sampling methods, such as Markov Chain Monte-Carlo (MCMC). Given large computational resource, they can generate exact results. In practice, however, sampling methods

Yaron Laufer and Sharon Gannot are with the Faculty of Engineering, Bar-Ilan University, Ramat-Gan, 5290002, Israel (e-mail: yaron_laufer@walla.com, Sharon.Gannot@biu.ac.il).

can be computationally expensive, often limiting their use. The deterministic approximation methods employ an analytical approximation to the posterior distribution by assuming a specific form. These methods are highly efficient and thus allow Bayesian techniques to be used in large-scale applications. A common method in this family is the variational Expectation-Maximization (VEM) [13]–[15]. Rather than estimating a single value for the model parameters, this approach estimates their posterior distributions. As the inference process relies on the entire posterior probability density function (PDF) instead of point estimates, this method has the potential to obtain estimators that are more robust and less sensitive to local maxima and overfitting, compared to the ML and MAP estimators [2], [15].

In [16], the problem of blind adaptive beamforming is addressed in the short-time Fourier transform (STFT) domain, using a VEM approach. The speech signal is modelled as a Gaussian process and the RTF is modelled as a first-order Markov model. By modelling the speech signal and the channel as latent variables, their posterior distributions are jointly estimated in the E-step. However, the considered model is not fully Bayesian, since the variance of the speaker and covariance matrix of the noise are modelled as deterministic unknown parameters, rather than latent random variables. These parameters are estimated in the M-step using point estimators. In addition, the RTF was estimated using the casual *Kalman filter*, i.e. using only past and present observations.

The Bayesian approach has also been pursued in the context of speech dereverberation in [17], [18] and in the context of source separation, e.g. in [19], [20].

In this paper, we extend the probabilistic model proposed in [16] towards a 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 complex Wishart prior for the noise precision matrix. This way, the precisions are also modelled as latent random variables, for which posterior distributions are inferred. The inference of the hidden variables is carried out using a VEM algorithm. Inspired by the decomposition of the multichannel minimum mean square error (MMSE) estimator of a single speech signal, namely the MCWF, into a multichannel minimum variance distortionless response (MVDR) beamformer followed by a subsequent single-channel Wiener postfilter [21], [22], we show that the VEM speech estimator has an analogous decomposition. Similarly to the MCWF, it includes an MVDR beamformer as an initial stage. However, the single-channel Wiener postfilter, which aims to minimize the residual noise at the MVDR output, is substituted by a variational postfilter, which also takes into account the uncertainty in the RTF estimate, and weights accordingly the single-channel at the MVDR output.

Considering the channel coefficients as random latent variables, may have higher modelling capacity than considering them as deterministic and unknown variables [23]. Therefore, it was proposed in the context of source separation [19] to model the time-varying audio channel as a set of hidden random variables, which are temporally related through a first-order linear dynamical system (LDS) [13], with a reduced set of parameters. We adopt this latent Bayesian model for the underlying filtering, and derive a *Kalman smoother* to infer the time-varying RTF. This way we exploit all available data to estimate the RTF.

In order to simultaneously infer all hidden variables, our VEM algorithm consists of alternating the following steps: (i) inferring the instantaneous speech signal by means of Wiener filtering, (ii) inferring the RTF sequence with Kalman smoother, (iii) inferring the speech precision, (iii) inferring the noise precision matrix, and (iv) updating rules for the hyperparameters.

The current paper is an essential extension of [24]. In [24], the noise is assumed to be a spatially homogeneous and spherically diffuse sound field, with a time-invariant covariance matrix modelled by a known spatial coherence matrix multiplied by an unknown noise power. Hence, the inverse power of the noise is treated as a latent random variable, modelled with the Gamma distribution. Here we generalize the model by omitting the assumptions about the spatial structure of the noise covariance. The entire precision matrix of the noise is therefore treated as a latent variable, modelled with the *complex* Wishart distribution. By allowing complex-valued precision matrices, our model accounts for any spatial structure of the noise field. We also provide here a comprehensive description of the assumed model, and derive the proposed VEM algorithm in detail. We describe the full mathematical derivation that was omitted in [24]. We also extend the preliminary performance evaluation, presented in [24], using variety of speech signals and noise types. In addition, we examine the performance on real-life recordings in static as well as dynamic scenarios.

The remainder of the paper is organized as follows. Section II introduces some notations and preliminary notes. Section III presents the problem formulation, and describes the probabilistic model. Section IV describes the proposed VEM algorithm. Section V demonstrates the performance of the proposed algorithm by an extensive experimental study based on both simulated data and real recordings. The paper is concluded in Section VI.

II. Notation and Preliminaries

In our notation, scalars are denoted with regular lowercase letters, vectors are denoted with bold lowercase letters and matrices are denoted with bold uppercase letters. The superscripts $\top$ and H describe transposition and Hermitian transposition, respectively. $|\cdot|$ denotes the determinant of a matrix, $\mathrm{tr}\{\cdot\}$ denotes the trace operator and $\overset{c}{=}$ denotes equality up to an additive constant. For a random vector $\mathbf{a}$, a multivariate proper complex Gaussian distribution is given by [25]–[27]:

$$\mathcal{N}_c(\mathbf{a};\boldsymbol{\mu}_a,\boldsymbol{\Phi}_a) = \frac{1}{|\pi\boldsymbol{\Phi}_a|}\exp\left[-\left(\mathbf{a}-\boldsymbol{\mu}_a\right)^{\mathrm{H}}\boldsymbol{\Phi}_a^{-1}\left(\mathbf{a}-\boldsymbol{\mu}_a\right)\right], \tag{1}$$

where $\boldsymbol{\mu}_a$ is the mean vector and $\boldsymbol{\Phi}_a$ is an Hermitian positive definite complex covariance matrix. The inverse covariance matrix, $\boldsymbol{\Phi}_a^{-1}$, is the precision matrix. A Gamma distribution for a non-negative random variable λ with shape and rate

parameters $a, b > 0$ is given by [13]:

$$\mathrm{Gam}(\lambda; a, b) = \begin{cases} \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda) & \lambda > 0 \\ 0 & \lambda \leq 0 \end{cases}, \quad (2)$$

where $\Gamma(\cdot)$ is the gamma function defined by

$$\Gamma(x) \equiv \int_0^\infty u^{x-1} \exp(-u) du. \quad (3)$$

The Gamma distribution has the following properties [13]:

$$\mathbb{E}[\lambda] = \frac{a}{b}, \quad (4a)$$

$$\mathbb{E}[\ln \lambda] = \Psi(a) - \ln(b), \quad (4b)$$

where $\Psi(\cdot)$ is the digamma function, defined by

$$\Psi(a) \equiv \frac{d}{da} \ln \Gamma(a). \quad (5)$$

A complex Wishart distribution for a random $N \times N$ Hermitian positive definite matrix $\boldsymbol{T}_c$ is given by [26]:

$$\mathcal{W}_c(\boldsymbol{T}_c; \boldsymbol{W}_c, \nu_c) = B_c(\boldsymbol{W}_c, \nu_c) |\boldsymbol{T}_c|^{\nu_c - N} \exp\Big[-\mathrm{tr}\big(\boldsymbol{W}_c^{-1} \boldsymbol{T}_c\big)\Big], \quad (6)$$

with

$$B_c(\boldsymbol{W}_c, \nu_c) = |\boldsymbol{W}_c|^{-\nu_c} \left[\pi^{N(N-1)/2} \prod_{i=1}^{N} \Gamma\big(\nu_c + 1 - i\big) \right]^{-1}, \quad (7)$$

where $\boldsymbol{W}_c$ is the Hermitian positive definite scale matrix, and ν_c is the number of degrees of freedom. The complex Wishart distribution has the following properties (see [28]):

$$\mathbb{E}[\boldsymbol{T}_c] = \nu_c \boldsymbol{W}_c, \quad (8a)$$

$$\mathbb{E}[\ln |\boldsymbol{T}_c|] = \sum_{i=1}^{N} \Psi(\nu_c + 1 - i) + \ln |\boldsymbol{W}_c|. \quad (8b)$$

III. Problem Formulation

A. Signal Model

Consider a speech signal received by N microphones, in a noisy and reverberant acoustic environment. We work with the STFT representation of the measured signals. Let $k \in [1, K]$ denote the frequency bin index, and $\ell \in [1, L]$ denote the time frame index. The N-channel observation signal $\mathbf{x}(\ell, k) = [x_1(\ell, k), \cdots, x_N(\ell, k)]^\top$ is given by

$$\mathbf{x}(\ell, k) = s(\ell, k)\mathbf{a}(\ell, k) + \mathbf{u}(\ell, k), \quad (9)$$

where $s(\ell, k)$ is the echoic speech signal as received by the first microphone (designated as a reference microphone) and $\mathbf{a}(\ell, k) = [1, a_2(\ell, k), \cdots, a_N(\ell, k)]^\top$ is the RTF vector. By taking a reference microphone and normalizing the acoustic transfer function (ATF) to construct the RTF, we circumvent the problem of gain ambiguity between the source signal and the ATF [1]. The ambient noise, which may originate from both microphone responses and from environmental sources, is denoted by $\mathbf{u}(\ell, k) = [u_1(\ell, k), \cdots, u_N(\ell, k)]^\top$.

Given the observed STFTs $\{\mathbf{x}(\ell, k)\}_{\ell,k=1}^{L,K}$, we are interested in estimating the speech signal $\{s(\ell, k)\}_{\ell,k=1}^{L,K}$. This requires the estimation of the RTF and the characteristics of both the speech and noise.

B. Probabilistic Model

The speech STFT coefficients are assumed to follow a zero-mean proper (i.e. circularly-symmetric) complex Gaussian distribution with time-varying precision $\tau(\ell, k)$, and they are all statistically independent frame- and frequency-wise. Hence, the PDF writes:

$$p\big(s(\ell, k) | \tau(\ell, k)\big) = \mathcal{N}_c\big(s(\ell, k); 0, \tau^{-1}(\ell, k)\big). \quad (10)$$

The noise is modelled as a zero-mean proper complex multivariate Gaussian, given by

$$p\big(\mathbf{u}(\ell, k) | \boldsymbol{T}(k)\big) = \mathcal{N}_c\big(\mathbf{u}(\ell, k); \mathbf{0}, \boldsymbol{T}^{-1}(k)\big), \quad (11)$$

where $\boldsymbol{T}(k)$ denotes the noise precision matrix, which is assumed to be time-invariant. The conditional data distribution is therefore given by

$$p\big(\mathbf{x}(\ell, k) | s(\ell, k), \mathbf{a}(\ell, k), \boldsymbol{T}(k)\big) = \mathcal{N}_c\big(\mathbf{x}(\ell, k); s(\ell, k)\mathbf{a}(\ell, k), \boldsymbol{T}^{-1}(k)\big). \quad (12)$$

The RTF vector may be time-varying if the speaker or the microphone array are moving around the enclosure. Thus it should be described with a dynamical model. We model the RTF as a set of hidden variables and parametrize its temporal progression as a random walk [19]. Accordingly, it is assumed that for each frequency k, the RTF sequence $\{\mathbf{a}(\ell, k)\}_{\ell=1}^{L}$ is governed by a first-order LDS. The initial and transition distributions are both assumed to follow complex Gaussian distributions. Specifically, this writes

$$p\big(\mathbf{a}(1, k)\big) = \mathcal{N}_c\big(\mathbf{a}(1, k); \boldsymbol{\mu}_a(k), \boldsymbol{\Phi}_a(k)\big), \quad (13)$$

$$p\big(\mathbf{a}(\ell, k) | \mathbf{a}(\ell - 1, k)\big) = \mathcal{N}_c\big(\mathbf{a}(\ell, k); \mathbf{a}(\ell - 1, k), \boldsymbol{\Phi}_a(k)\big), \quad (14)$$

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

C. 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 about the model. For the exponential family of distributions, choosing a conjugate form of prior leads to a posterior distribution with the same functional form, and therefore results with a simplified Bayesian analysis. In this framework, it is often more convenient to work with the precision rather than the covariance [13]. Our hierarchical generative model is therefore established by introducing priors for the precisions of the speaker and the noise. For a Gaussian vector, the conjugate prior distribution of the precision matrix is the Wishart distribution [13]. Since the speech precision $\tau(\ell, k)$ is a scalar, we use the degenerated form of the Wishart distribution, namely the Gamma distribution:

$$p\big(\tau(\ell, k)\big) = \mathrm{Gam}\big(\tau(\ell, k); a(\ell, k), b_0(\ell, k)\big), \quad (15)$$

where $a_0(\ell, k), b_0(\ell, k)$ are hyperparameters. Each time-frequency (TF) bin is modelled 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

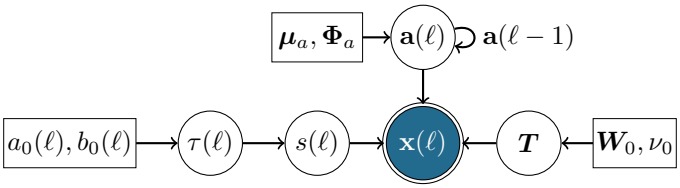


Fig. 1: Graphical model. Double circles represent observations, circles represent latent variables and rectangles represent deterministic parameters (Frequency index is omitted).

the precision $\tau(\ell,k)$, the true prior distribution of the speech signal turns to be a Student's t-distribution [13]. 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 [29], [30], which have sparse distribution in the TF domain. For the time-invariant precision matrix of the ambient noise $\boldsymbol{T}(k)$, we assume a complex Wishart distribution as prior:

$$p\big(\boldsymbol{T}(k)\big) = \mathcal{W}_c\big(\boldsymbol{T}(k); \boldsymbol{W}_0(k), \nu_0(k)\big). \tag{16}$$

In general, the noise precision matrix might be complex-valued, hence a better fit can be obtained by modelling it with a complex distribution. The complete graphical model of the proposed hierarchical model is shown in Fig. 1.

IV. VARIATIONAL BAYESIAN ALGORITHM

In this section, we develop a VEM algorithm which jointly infer the latent variables and estimate the model parameters. We begin with the general problem of Bayesian inference, then we introduce the VEM framework. Then we describe our E-step, composed of an E-s step for the speech signal, E-$\mathbf{a}$ step for the RTF sequence, E-τ step for the speech precision and E-$\boldsymbol{T}$ step for the noise precision matrix. Finally, the M-step is presented.

A. Variational Inference

Consider a probabilistic model in which we denote all of the observed data as $\mathcal{X}$ and all of the latent variables as $\mathcal{H}$. The joint distribution $p(\mathcal{X},\mathcal{H};\theta)$ is governed by a set of parameters denoted θ. Bayesian inference aims to infer the latent variables according to their posterior mean (PM), using an ML estimation of the model parameters. To this end, it is required to compute the posterior distribution of the hidden variables $p(\mathcal{H}|\mathcal{X};\theta) = p(\mathcal{X},\mathcal{H};\theta)/p(\mathcal{X};\theta)$. The EM algorithm is a general technique for finding ML solutions for probabilistic models having latent variables [11]. However, the EM algorithm requires knowledge of the posterior in order to maximize the likelihood function iteratively. A serious shortcoming of this methodology is that for complex Bayesian models this posterior is not available, and thus the EM algorithm is not applicable.

The VEM approach provides approximate solutions for Bayesian inference problems. This method relaxes the limiting requirements of the EM algorithm and thus is applicable to complex problems. In the variational inference procedure, the exact posterior distribution over the hidden variables is substituted by an approximate one $q(\mathcal{H}) \approx p(\mathcal{H}|\mathcal{X};\theta)$. In order to achieve tractable distributions, we must restrict the family of distribution $q(\mathcal{H})$. A particular form that has been used with success is the factorized one, called *mean field theory* [14], [31]. According to this approximation, we assume that $q(\mathcal{H})$ can be factorized over some partition of the hidden variables:

$$q(\mathcal{H}) = \prod_{\mathcal{H}_i \subset \mathcal{H}} q(\mathcal{H}_i). \tag{17}$$

The variational approach is based on the following decomposition of the log-likelihood function. For any PDF $q(\mathcal{H})$ defined over the latent variables, the following decomposition holds:

$$\ln p(\mathcal{X};\theta) = F(q(\mathcal{H});\theta) + \mathrm{KL}\big(q(\mathcal{H})||p(\mathcal{H}|\mathcal{X};\theta)\big), \tag{18}$$

with

$$F(q(\mathcal{H});\theta) = \int q(\mathcal{H}) \ln \frac{p(\mathcal{X},\mathcal{H};\theta)}{q(\mathcal{H})} d\mathcal{H}, \tag{19}$$

and

$$\mathrm{KL}(q(\mathcal{H})||p(\mathcal{H}|\mathcal{X};\theta)) = -\int q(\mathcal{H}) \ln \frac{p(\mathcal{H}|\mathcal{X};\theta)}{q(\mathcal{H})} d\mathcal{H}, \tag{20}$$

where $\mathrm{KL}(q\big(\mathcal{H})||p(\mathcal{H}|\mathcal{X};\theta)\big)$ is the Kullback-Leibler (KL) divergence between $q(\mathcal{H})$ and the posterior distribution $p(\mathcal{H}|\mathcal{X};\theta)$. Since the KL divergence is always non-negative, it follows that $F(q(\mathcal{H});\theta)$ is a lower bound on the log-likelihood, with equality if and only if $q(\mathcal{H}) = p(\mathcal{H}|\mathcal{X};\theta)$.

We wish to find $q(\mathcal{H})$ of the form of (17) that maximizes the lower bound $F(q(\mathcal{H});\theta)$. The maximization of the lower bound consists in iterating the following two steps until convergence. In the E-step we compute

$$q^*(\mathcal{H}) = \underset{q(\mathcal{H})}{\operatorname{argmax}}\ F(q(\mathcal{H});\theta_{\text{old}}), \tag{21}$$

where θ_{old} is the current value of the parameter vector. For solving (21), it is well known that the optimal marginal posterior distribution of a subset $\mathcal{H}_i \subseteq \mathcal{H}$ is given by [13]

$$\ln q^*(\mathcal{H}_i) = \mathbb{E}_{q(\mathcal{H}/\mathcal{H}_i)}[\ln p(\mathcal{X},\mathcal{H};\theta)] + \text{const}, \tag{22}$$

where $q(\mathcal{H}/\mathcal{H}_i)$ denotes the approximated joint posterior distribution of all latent variables, excluding the subset $\mathcal{H}_i$. The additive constant in (22) can be obtained through normalization. Accordingly, $q(\mathcal{H})$ is inferred by alternating over each subset $\mathcal{H}_i \subset \mathcal{H}$. Once we have the posterior distributions of all the variables in $\mathcal{H}$, $F(q^*(\mathcal{H});\theta)$ is maximized with respect to the parameters in the M-Step. To this end, $F(q(\mathcal{H});\theta)$ can be further decomposed as

$$F(q(\mathcal{H});\theta) = \mathbb{E}_{q(\mathcal{H})}\left[\ln p(\mathcal{X},\mathcal{H};\theta)\right] - \mathbb{E}_{q(\mathcal{H})}\left[\ln q(\mathcal{H})\right], \tag{23}$$

where the first term is the expected log-likelihood of the complete-data (denoted by $\mathcal{L}(\theta)$) and the second term is the entropy of the distribution $q(\mathcal{H})$, which is independent of θ. Using (23) the maximization in the M-step writes

$$\theta_{\text{new}} = \underset{\theta}{\operatorname{argmax}}\ \mathbb{E}_{q^*(\mathcal{H})}\left[\ln p(\mathcal{X},\mathcal{H};\theta)\right]. \tag{24}$$

B. Variational EM for Speech Enhancement

In this work, the observations set is $\mathcal{X} = \{\mathbf{x}(\ell,k)\}_{\ell,k=1}^{L,K}$, the set of hidden variables consists of

$$\mathcal{H} = \Big\{s(\ell,k), \mathbf{a}(\ell,k), \tau(\ell,k), \boldsymbol{T}(k)\Big\}_{\ell,k=1}^{L,K},$$

and the parameter set consists of

$$\theta = \Big\{a_0(\ell,k), b_0(\ell,k), \boldsymbol{W}_0(k), \nu_0(k), \boldsymbol{\mu}_a(k), \boldsymbol{\Phi}_a(k)\Big\}_{\ell,k=1}^{L,K}.$$

The complete-data distribution writes

$$\begin{aligned} p(\mathcal{X},\mathcal{H};\theta) &= p(\mathcal{X}|\mathcal{H};\theta)p(\mathcal{H};\theta) \\ &= \prod_{\ell,k=1}^{L,K}\Big[p\big(\mathbf{x}(\ell,k)|s(\ell,k),\mathbf{a}(\ell,k),\boldsymbol{T}(k)\big) \\ &\qquad \times p\big(s(\ell,k)|\tau(\ell,k)\big)p\big(\tau(\ell,k);a_0(\ell,k),b_0(\ell,k)\big)\Big] \\ &\quad \times \prod_{k=1}^{K}\Big[p\big(\boldsymbol{T}(k);\boldsymbol{W}_0(k),\nu_0(k)\big)p\big(\mathbf{a}(1,k);\boldsymbol{\mu}_a(k),\boldsymbol{\Phi}_a(k)\big) \\ &\qquad \times \prod_{\ell=2}^{L} p\big(\mathbf{a}(\ell,k);\mathbf{a}(\ell-1,k),\boldsymbol{\Phi}_a(k)\big)\Big]. \end{aligned} \tag{25}$$

However, the likelihood $p(\mathcal{X};\theta) = \int p(\mathcal{X},\mathcal{H};\theta)d\mathcal{H}$ cannot be computed analytically from (25), hence the posterior $p(\mathcal{H}|\mathcal{X};\theta)$ cannot be expressed in closed-form and exact inference becomes intractable. We resort to the approximate variational inference methodology, which circumvents this difficulty by approximating the posterior. According to the factorized approximation (17), we assume that the speech signal, RTF, speech precision and the noise precision matrix are conditionally independent given the observations. Hence, the approximate posterior distribution has the following factorized form:

$$\begin{aligned} q(\mathcal{H}) = &\prod_{\ell,k=1}^{L,K}\Big[q\big(s(\ell,k)\big)q\big(\tau(\ell,k)\big)\Big] \\ &\times \prod_{k=1}^{K}\Big[q\big(\boldsymbol{T}(k)\big)q\big(\mathbf{a}(1{:}L,k)\big)\Big]. \end{aligned} \tag{26}$$

For the sake of brevity, the frequency bin index k is henceforth omitted.

C. E-s Step

The approximate posterior distribution of the source can be computed from (25) and (22) by recognizing the terms of $\ln q(s(\ell))$ that are function of $s(\ell)$:

$$\begin{aligned} \ln q(s(\ell)) &\stackrel{c}{=} \mathbb{E}_{q(\mathbf{a}(\ell))q(\tau(\ell))q(\boldsymbol{T})}\Big[\ln p\big(\mathbf{x}(\ell)|s(\ell),\mathbf{a}(\ell),\boldsymbol{T}\big) \\ &\qquad + \ln p\big(s(\ell)|\tau(\ell)\big)\Big] \\ &\stackrel{c}{=} -|s(\ell)|^2\Big\{\mathbb{E}_{q(\mathbf{a}(\ell))q(\boldsymbol{T})}\big[\mathbf{a}^{\mathrm{H}}(\ell)\boldsymbol{T}\mathbf{a}(\ell)\big] + \hat{\tau}(\ell)\Big\} \\ &\qquad + s^*(\ell)\hat{\mathbf{a}}^{\mathrm{H}}(\ell)\hat{\boldsymbol{T}}\mathbf{x}(\ell) + s(\ell)\mathbf{x}^{\mathrm{H}}(\ell)\hat{\boldsymbol{T}}\hat{\mathbf{a}}(\ell), \end{aligned} \tag{27}$$

where $\hat{\mathbf{a}}(\ell), \boldsymbol{\Sigma}_a(\ell), \hat{\boldsymbol{T}}$ and $\hat{\tau}(\ell)$ are posterior statistics that will be defined in Sections IV-D-IV-F. The remaining expectation can be simplified by the cyclic property of the trace operation:

$$\begin{aligned} \mathbb{E}_{q(\mathbf{a}(\ell))q(\boldsymbol{T})}\big[\mathbf{a}^{\mathrm{H}}(\ell)\boldsymbol{T}\mathbf{a}(\ell)\big] &= \mathrm{tr}\Big\{\mathbb{E}_{q(\mathbf{a}(\ell))q(\boldsymbol{T})}\big[\mathbf{a}(\ell)\mathbf{a}^{\mathrm{H}}(\ell)\boldsymbol{T}\big)\big]\Big\} \\ &= \mathrm{tr}\Big\{\big[\hat{\mathbf{a}}(\ell)\hat{\mathbf{a}}^{\mathrm{H}}(\ell) + \boldsymbol{\Sigma}_a(\ell)\big]\hat{\boldsymbol{T}}\Big\} \\ &= \hat{\mathbf{a}}^{\mathrm{H}}(\ell)\hat{\boldsymbol{T}}\hat{\mathbf{a}}(\ell) + \mathrm{tr}\big[\boldsymbol{\Sigma}_a(\ell)\hat{\boldsymbol{T}}\big]. \end{aligned} \tag{28}$$

Eq. (27) is an incomplete quadratic form in $s(\ell)$, thus represents a Gaussian distribution:

$$q\left(s(\ell)\right) = \mathcal{N}_c\big(s(\ell);\hat{s}(\ell),\Sigma_s(\ell)\big), \tag{29}$$

with mean $\hat{s}(\ell)$ and variance $\Sigma_s(\ell)$ given by:

$$\Sigma_s(\ell) = \Big\{\hat{\mathbf{a}}^{\mathrm{H}}(\ell)\hat{\boldsymbol{T}}\hat{\mathbf{a}}(\ell) + \mathrm{tr}\big[\boldsymbol{\Sigma}_a(\ell)\hat{\boldsymbol{T}}\big] + \hat{\tau}(\ell)\Big\}^{-1}, \tag{30}$$

$$\hat{s}(\ell) = \frac{\hat{\mathbf{a}}^{\mathrm{H}}(\ell)\hat{\boldsymbol{T}}\mathbf{x}(\ell)}{\hat{\mathbf{a}}^{\mathrm{H}}(\ell)\hat{\boldsymbol{T}}\hat{\mathbf{a}}(\ell) + \mathrm{tr}\big[\boldsymbol{\Sigma}_a(\ell)\hat{\boldsymbol{T}}\big] + \hat{\tau}(\ell)}. \tag{31}$$

The speech signal can be estimated by the PM, namely $\hat{s}(\ell)$, with the variance $\Sigma_s(\ell)$. This speech estimator resembles the form of the MCWF [32], [33]:

$$\hat{s}_{\mathrm{MCWF}}(\ell) = \frac{\hat{\mathbf{a}}^{\mathrm{H}}(\ell)\hat{\boldsymbol{T}}\mathbf{x}(\ell)}{\hat{\mathbf{a}}^{\mathrm{H}}(\ell)\hat{\boldsymbol{T}}\hat{\mathbf{a}}(\ell) + \hat{\tau}(\ell)}, \tag{32}$$

except the term $\mathrm{tr}\big[\boldsymbol{\Sigma}_a(\ell)\hat{\boldsymbol{T}}\big]$. The MCWF is the optimal beamformer in the sense of minimizing the MMSE of the speech signal. Furthermore, it can be decomposed into a multichannel MVDR beamformer followed by a single-channel Wiener postfilter [21], [22], [33]:

$$\hat{s}_{\mathrm{MCWF}}(\ell) = \underbrace{\frac{\hat{\mathbf{a}}^{\mathrm{H}}(\ell)\hat{\boldsymbol{T}}\hat{\mathbf{a}}(\ell)}{\hat{\mathbf{a}}^{\mathrm{H}}(\ell)\hat{\boldsymbol{T}}\hat{\mathbf{a}}(\ell) + \hat{\tau}(\ell)}}_{H_W(\ell)} \times \underbrace{\frac{\hat{\mathbf{a}}^{\mathrm{H}}(\ell)\hat{\boldsymbol{T}}}{\hat{\mathbf{a}}^{\mathrm{H}}(\ell)\hat{\boldsymbol{T}}\hat{\mathbf{a}}(\ell)}}_{\mathbf{w}_{\mathrm{MVDR}}^{\mathrm{H}}(\ell)}\mathbf{x}(\ell), \tag{33}$$

where $H_W(\ell)$ is the single-channel Wiener filter at the output of the MVDR. In a similar way to the MCWF decomposition, $\hat{s}(\ell)$ can be decomposed as

$$\hat{s}(\ell) = \underbrace{\frac{\hat{\mathbf{a}}^{\mathrm{H}}(\ell)\hat{\boldsymbol{T}}\hat{\mathbf{a}}(\ell)}{\hat{\mathbf{a}}^{\mathrm{H}}(\ell)\hat{\boldsymbol{T}}\hat{\mathbf{a}}(\ell) + \mathrm{tr}\big[\boldsymbol{\Sigma}_a(\ell)\hat{\boldsymbol{T}}\big] + \hat{\tau}(\ell)}}_{H(\ell)} \times \underbrace{\frac{\hat{\mathbf{a}}^{\mathrm{H}}(\ell)\hat{\boldsymbol{T}}}{\hat{\mathbf{a}}^{\mathrm{H}}(\ell)\hat{\boldsymbol{T}}\hat{\mathbf{a}}(\ell)}}_{\mathbf{w}_{\mathrm{MVDR}}^{\mathrm{H}}(\ell)}\mathbf{x}(\ell) \tag{34}$$

where $H_W(\ell)$ is replaced by a single-channel variational post-filter, $H(\ell)$, having the following role. 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 its uncertainty. In contrast, the VEM estimator treats the RTF as a latent variable and considers its posterior distribution, which captures the uncertainty regarding the parameter estimate. Therefore, it includes $H(\ell)$ as a postfilter that takes into account that uncertainty level, expressed by $\boldsymbol{\Sigma}_a(\ell)$, and weights accordingly the single-channel at the MVDR output. When $\boldsymbol{\Sigma}_a(\ell) \to \mathbf{0}$, $\hat{s}(\ell)$ reduces to $\hat{s}_{\mathrm{MCWF}}(\ell)$.

D. E-$\mathbf{a}$ Step

The joint posterior distribution of the RTF sequence is obtained by identifying the terms that are function of $\mathbf{a}(1{:}L)$:

$$\ln q(\mathbf{a}(1{:}L)) \stackrel{c}{=} \sum_{\ell=1}^{L} \mathbb{E}_{q(s(\ell))q(\boldsymbol{T})}\Big[\ln p\big(\mathbf{x}(\ell)|s(\ell),\mathbf{a}(\ell),\boldsymbol{T}\big)\Big] + \ln p\big(\mathbf{a}(1{:}L)\big). \tag{35}$$

Following the same lines as in [19], the first term reduces to a Gaussian distribution $\mathcal{N}_c\big(\boldsymbol{\mu}_L(\ell);\mathbf{a}(\ell),\boldsymbol{\Phi}_L(\ell)\big)$ with $\boldsymbol{\Phi}_L(\ell) = \Big(\widehat{|s(\ell)|^2}\hat{\boldsymbol{T}}\Big)^{-1}$ and $\boldsymbol{\mu}_L(\ell) = \frac{\hat{s}^*(\ell)}{\widehat{|s(\ell)|^2}}\mathbf{x}(\ell)$. Thus we obtain:

$$q\big(\mathbf{a}(1{:}L)\big) = p\big(\mathbf{a}(1{:}L)\big)\prod_{\ell=1}^{L}\mathcal{N}_c\big(\boldsymbol{\mu}_L(\ell);\mathbf{a}(\ell),\boldsymbol{\Phi}_L(\ell)\big). \tag{36}$$

We identify (36) as a first-order LDS [13] over the latent states $\{\mathbf{a}(\ell)\}_{\ell=1}^{L}$, where (13) is the initial distribution, (14) is the transition distribution and $\mathcal{N}_c(\boldsymbol{\mu}_L(\ell);\mathbf{a}(\ell),\boldsymbol{\Phi}_L(\ell))$ denotes the emission distribution. Hence, the marginal posterior distribution of each frame is also a Gaussian distribution, which can be recursively calculated using the *Kalman smoother* [13]. This is a recursive algorithm that consists of a forward pass and a backward pass. Forward inference over the states is used to estimate a distribution over $\mathbf{a}(\ell)$ using all observations up to time ℓ, whereas backward inference uses future observations from time $\ell+1$ up to L. The two passes are afterwards combined to give the posterior marginal distribution of the state variable $\mathbf{a}(\ell)$ given all observations:

$$q(\mathbf{a}(\ell)) = \mathcal{N}_c\big(\mathbf{a}(\ell);\hat{\mathbf{a}}(\ell),\boldsymbol{\Sigma}_a(\ell)\big), \tag{37}$$

with

$$\boldsymbol{\Sigma}_a(\ell) = \Big(\boldsymbol{\Phi}_f(\ell)^{-1} + \boldsymbol{\Phi}_\beta(\ell)^{-1}\Big)^{-1} \tag{38}$$

$$\hat{\mathbf{a}}(\ell) = \boldsymbol{\Sigma}_a(\ell)\Big(\boldsymbol{\Phi}_f(\ell)^{-1}\mathbf{a}_f(\ell) + \boldsymbol{\Phi}_\beta(\ell)^{-1}\mathbf{a}_\beta(\ell)\Big). \tag{39}$$

$\boldsymbol{\Phi}_f(\ell),\mathbf{a}_f(\ell)$ are provided by the forward pass, and $\boldsymbol{\Phi}_\beta(\ell),\mathbf{a}_\beta(\ell)$ are provided by the backward pass, both detailed in the Appendix. The RTF can therefore be estimated by the *smoothed* PM, namely $\hat{\mathbf{a}}(\ell)\in\mathbb{C}^N$, with uncertainty $\boldsymbol{\Sigma}_a(\ell)\in\mathbb{C}^{N\times N}$. Note that in [16] the RTF was inferred using the forward pass alone, i.e. $\mathbf{a}(\ell)$ was estimated given only causal observations up to time ℓ. We propose to find the marginal for $\mathbf{a}(\ell)$ based on all observations up to time L, i.e. comprise future as well as past observations. Although this cannot be used for real-time prediction, it plays a key role in learning the model.

Another quantity of interest is the pair-wise joint posterior distribution of two successive frames $\mathbf{a}(\ell)$ and $\mathbf{a}(\ell-1)$, which will be required to update $\boldsymbol{\Phi}_a$ in Section IV-G6. Marginalizing out all other frames in (36) results with a Gaussian distribution, $q\big(\mathbf{a}(\ell),\mathbf{a}(\ell-1)\big) = \mathcal{N}_c\big(\left[\mathbf{a}(\ell)^\top,\mathbf{a}(\ell-1)^\top\right]^\top;\mathbf{a}_\xi(\ell),\boldsymbol{\Sigma}_\xi(\ell)\big)$, where the mean vector $\mathbf{a}_\xi(\ell)\in\mathbb{C}^{2N}$ and the covariance matrix $\boldsymbol{\Sigma}_\xi(\ell)\in\mathbb{C}^{2N\times 2N}$ are given in the Appendix. The second-order joint posterior moment is defined as

$$\boldsymbol{Q}(\ell) = \boldsymbol{\Sigma}_\xi(\ell) + \mathbf{a}_\xi(\ell)\mathbf{a}_\xi^{\mathrm{H}}(\ell). \tag{40}$$

E. E-τ Step

Using (25) and (22), the posterior distribution of the speech precision writes:

$$\begin{aligned}\ln q(\tau(\ell)) &\stackrel{c}{=} \mathbb{E}_{q(s(\ell))}\Big[\ln p\big(s(\ell)|\tau(\ell)\big)\Big] + \ln p\big(\tau(\ell);a_0(\ell),b_0(\ell)\big)\\ &\stackrel{c}{=} a_0(\ell)\ln\tau(\ell) - \tau(\ell)\Big(b_0(\ell) + \widehat{|s(\ell)|^2}\Big).\end{aligned} \tag{41}$$

This is an exponent of a Gamma distribution:

$$q\left(\tau(\ell)\right) = \mathrm{Gam}\big(\tau(\ell);a_p(\ell),b_p(\ell)\big), \tag{42}$$

with parameters $a_p(\ell)$ and $b_p(\ell)$, given by

$$a_p(\ell) = a_0(\ell) + 1, \tag{43}$$

$$b_p(\ell) = b_0(\ell) + \widehat{|s(\ell)|^2}. \tag{44}$$

Thus, the PM estimate for the source precision writes

$$\hat{\tau}(\ell) = \frac{a_p(\ell)}{b_p(\ell)} = \frac{a_0(\ell)+1}{b_0(\ell) + \widehat{|s(\ell)|^2}}, \tag{45}$$

where the hyperparameters $a_0(\ell), b_0(\ell)$ are updated in the M-Step. It should be noted that treating the speech precision as an unknown deterministic parameter as in [16], 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 [13].

F. E-$\boldsymbol{T}$ Step

Similarly, the posterior distribution of the noise precision is given by:

$$\begin{aligned}\ln q(\boldsymbol{T}) &\stackrel{c}{=} \sum_{\ell=1}^{L}\mathbb{E}_{q(s(\ell))q(\mathbf{a}(\ell))}\Big[\ln p\big(\mathbf{x}(\ell)|s(\ell),\mathbf{a}(\ell),\boldsymbol{T}\big)\Big] + \ln p(\boldsymbol{T};\boldsymbol{W}_0,\nu_0)\\ &\stackrel{c}{=} -\mathrm{tr}\Bigg(\boldsymbol{T}\Big\{\boldsymbol{W}_0^{-1} + \sum_{\ell=1}^{L}\mathbb{E}_{q(s(\ell))q(\mathbf{a}(\ell))}\Big[\big(\mathbf{x}(\ell)-s(\ell)\mathbf{a}(\ell)\big)\\ &\quad\times\big(\mathbf{x}(\ell)-s(\ell)\mathbf{a}(\ell)\big)^{\mathrm{H}}\Big]\Big\}\Bigg) + (\nu_0 - N + L)\ln|\boldsymbol{T}|.\end{aligned} \tag{46}$$

This is an exponent of a complex Wishart distribution:

$$q(\boldsymbol{T}) = \mathcal{W}(\boldsymbol{T};\boldsymbol{W}_p,v_p), \tag{47}$$

with parameters $\boldsymbol{W}_p$ and v_p, given by

$$\begin{aligned}\boldsymbol{W}_p^{-1} = \boldsymbol{W}_0^{-1} + \sum_{\ell=1}^{L}\Big[&\mathbf{x}(\ell)\mathbf{x}^{\mathrm{H}}(\ell) - \mathbf{x}(\ell)\hat{\mathbf{a}}^{\mathrm{H}}(\ell)\hat{s}^*(\ell)\\ &- \hat{s}(\ell)\hat{\mathbf{a}}(\ell)\mathbf{x}^{\mathrm{H}}(\ell) + \widehat{|s(\ell)|^2}\Big(\hat{\mathbf{a}}(\ell)\hat{\mathbf{a}}^{\mathrm{H}}(\ell) + \boldsymbol{\Sigma}_a(\ell)\Big)\Big],\end{aligned} \tag{48}$$

$$v_p = \nu_0 + L. \tag{49}$$

Thus, the PM estimate for the noise precision matrix writes

$$\hat{\boldsymbol{T}} = v_p \boldsymbol{W}_p = (\nu_0 + L)\boldsymbol{W}_p. \tag{50}$$

Treating the precision matrix of the noise as an unknown deterministic parameter as in [16], leads to the following point estimator:

$$\hat{\boldsymbol{T}}_D^{-1} = \frac{1}{L}\sum_{\ell=1}^{L}\Big[\mathbf{x}(\ell)\mathbf{x}^{\mathrm{H}}(\ell) - \mathbf{x}(\ell)\hat{\mathbf{a}}^{\mathrm{H}}(\ell)\hat{s}^*(\ell) - \hat{s}(\ell)\hat{\mathbf{a}}(\ell)\mathbf{x}^{\mathrm{H}}(\ell) + \widehat{|s(\ell)|^2}\Big(\hat{\mathbf{a}}(\ell)\hat{\mathbf{a}}^{\mathrm{H}}(\ell) + \boldsymbol{\Sigma}_a(\ell)\Big)\Big]. \tag{51}$$

Note that $\boldsymbol{W}_p^{-1} = \boldsymbol{W}_0^{-1} + L\hat{\boldsymbol{T}}_D^{-1}$, hence (50) becomes $\hat{\boldsymbol{T}} = (\nu_0 + L)\left(\boldsymbol{W}_0^{-1} + L\hat{\boldsymbol{T}}_D^{-1}\right)^{-1}$. For the Wishart distribution, a *non-informative prior* is obtained as the special case $\nu_0 \to 0$ and $\boldsymbol{W}_0 = \frac{1}{\lambda}\boldsymbol{I}$ where $\lambda \to 0$. In this case we obtain $\hat{\boldsymbol{T}} = \hat{\boldsymbol{T}}_D$. However, the least informative *proper* Wishart prior is obtained by setting $\nu_0 = N$, leading to $\hat{\boldsymbol{T}} = \left(\frac{N}{L} + 1\right)\hat{\boldsymbol{T}}_D$.

The approximate posterior distributions in (29), (37), (42) and (47) are then iteratively updated until convergence, since they depend on the statistics of each other.

G. M-Step

In order to estimate the parameters, we maximize the expected log-likelihood of the complete-data $\mathcal{L}(\theta) = \mathbb{E}_{q(\mathcal{H})}\left[\ln p(\mathcal{X}, \mathcal{H}; \theta)\right]$ with respect to each parameter. For our model, $\mathcal{L}(\theta)$ is given by (52), at the bottom of the page.

1) M-$a_0(\ell)$ Step: Differentiating $\mathcal{L}(\theta)$ w.r.t. $a_0(\ell)$, we obtain

$$\frac{\partial \mathcal{L}(\theta)}{\partial a_0(\ell)} = \frac{\partial}{\partial a_0(\ell)}\mathbb{E}_{q(\tau(\ell))}\Big[\ln p\big(\tau(\ell); a_0(\ell), b_0(\ell)\big)\Big] = -\Psi(a_0(\ell)) + \ln b_0(\ell) + \mathbb{E}_{q(\tau(\ell))}\Big[\ln\big(\tau(\ell)\big)\Big], \tag{53}$$

where $\Psi(\cdot)$ is the digamma function, defined in (5). Using (4b) and comparing the derivative to zero yields:

$$a_0(\ell) = \Psi^{-1}\left[\Psi\big(a_p(\ell)\big) + \ln\frac{b_0(\ell)}{b_p(\ell)}\right]. \tag{54}$$

2) M-$b_0(\ell)$ Step: Differentiating $\mathcal{L}(\theta)$ w.r.t. $b_0(\ell)$, we obtain

$$\frac{\partial \mathcal{L}(\theta)}{\partial b_0(\ell)} = \frac{a_0(\ell)}{b_0(\ell)} - \mathbb{E}_{q(\tau(\ell))}\big[\tau(\ell)\big]. \tag{55}$$

Using (4a) and comparing to zero yields:

$$b_0(\ell) = \frac{a_0(\ell)}{a_p(\ell)} b_p(\ell). \tag{56}$$

3) M-ν_0 Step: Differentiating $\mathcal{L}(\theta)$ w.r.t. ν_0, we obtain

$$\frac{\partial \mathcal{L}(\theta)}{\partial \nu_0} = \frac{\partial}{\partial \nu_0}\mathbb{E}_{q(\boldsymbol{T})}\Big[\ln p\big(\boldsymbol{T}; \boldsymbol{W}_0, \nu_0\big)\Big] = -\ln|\boldsymbol{W}_0| - \sum_{i=1}^{N}\Psi\left(\nu_0 + 1 - i\right) + \mathbb{E}_{q(\boldsymbol{T})}\big[\ln|\boldsymbol{T}|\big]. \tag{57}$$

Using (8b) and comparing to zero yields the following equation:

$$\sum_{i=1}^{N}\Psi\left(\nu_0 + 1 - i\right) = \sum_{i=1}^{N}\Psi\left(\nu_p + 1 - i\right) + \ln\frac{|\boldsymbol{W}_p|}{|\boldsymbol{W}_0|}. \tag{58}$$

Since this equation cannot be solved in a closed-form for ν_0, we apply the iterative method of Newton to find a maximum [34]:

$$\nu_0^{(m+1)} = \nu_0^{(m)} - \frac{d\left(\nu_0^{(m)}\right)}{h\left(\nu_0^{(m)}\right)}, \tag{59}$$

where $d(\nu_0)$ and $h(\nu_0)$ are the first and second-order derivatives of $\mathcal{L}(\theta)$ w.r.t. ν_0:

$$d(\nu_0) \equiv \frac{\partial \mathcal{L}(\theta)}{\partial \nu_0}\ ,\ h(\nu_0) \equiv \frac{\partial^2 \mathcal{L}(\theta)}{\partial \nu_0^2}. \tag{60}$$

The derivatives in our case are given by

$$d(\nu_0) = \sum_{i=1}^{N}\Big[\Psi\big(\nu_p + 1 - i\big) - \Psi\big(\nu_0 + 1 - i\big)\Big] + \ln\frac{|\boldsymbol{W}_p|}{|\boldsymbol{W}_0|}, \tag{61}$$

$$h(\nu_0) = -\sum_{i=1}^{N}\Psi'\big(\nu_0 + 1 - i\big), \tag{62}$$

where $\Psi'(\cdot)$ is the trigamma function.

4) M-$\boldsymbol{W}_0$ Step: Differentiating $\mathcal{L}(\theta)$ w.r.t. $\boldsymbol{W}_0$, we obtain

$$\frac{\partial \mathcal{L}(\theta)}{\partial \boldsymbol{W}_0} = -\nu_0\boldsymbol{W}_0^{-1} + \boldsymbol{W}_0^{-1}\mathbb{E}_{q(\boldsymbol{T})}\big[\boldsymbol{T}\big]\boldsymbol{W}_0^{-1}. \tag{63}$$

Using (8a) and comparing to zero yields:

$$\boldsymbol{W}_0 = \frac{\nu_p}{\nu_0}\boldsymbol{W}_p. \tag{64}$$

5) M-$\boldsymbol{\mu}_a$ Step: Differentiating $\mathcal{L}(\theta)$ w.r.t. $\boldsymbol{\mu}_a$, we obtain

$$\frac{\partial \mathcal{L}(\theta)}{\partial \boldsymbol{\mu}_a{}^{\mathrm{H}}} = \frac{\partial}{\partial \boldsymbol{\mu}_a{}^{\mathrm{H}}}\mathbb{E}_{q(\mathbf{a}(1))}\Big[\ln p\big(\mathbf{a}(1); \boldsymbol{\mu}_a, \boldsymbol{\Phi}_a\big)\Big] = \boldsymbol{\Phi}_a^{-1}\Big\{\mathbb{E}_{q(\mathbf{a}(1))}\big[\mathbf{a}(1)\big] - \boldsymbol{\mu}_a\Big\}. \tag{65}$$

$$\begin{aligned}\mathcal{L}(\theta) = &\sum_{\ell,k=1}^{L,K}\mathbb{E}_{q(s(\ell,k))q(\mathbf{a}(\ell,k))q(\boldsymbol{T}(k))}\Big[\ln p\big(\mathbf{x}(\ell,k)|s(\ell,k),\mathbf{a}(\ell,k),\boldsymbol{T}(k)\big)\Big] + \sum_{\ell,k=1}^{L,K}\mathbb{E}_{q(s(\ell,k))q(\tau(\ell,k))}\Big[\ln p\big(s(\ell,k)|\tau(\ell,k)\big)\Big] \\ &+ \sum_{\ell,k=1}^{L,K}\mathbb{E}_{q(\tau(\ell,k))}\Big[\ln p\big(\tau(\ell,k); a_0(\ell,k), b_0(\ell,k)\big)\Big] + \sum_{k=1}^{K}\mathbb{E}_{q(\boldsymbol{T}(k))}\Big[\ln p\big(\boldsymbol{T}(k); \boldsymbol{W}_0(k), \nu_0(k)\big)\Big] \\ &+ \sum_{k=1}^{K}\Big\{\mathbb{E}_{q(\mathbf{a}(1,k))}\Big[\ln p\big(\mathbf{a}(1,k); \boldsymbol{\mu}_a(k), \boldsymbol{\Phi}_a(k)\big)\Big] + \sum_{\ell=2}^{L}\mathbb{E}_{q(\mathbf{a}_{\{\ell:\ell-1,k\}})}\Big[\ln p\big(\mathbf{a}(\ell,k)|\mathbf{a}(\ell-1,k); \boldsymbol{\Phi}_a(k)\big)\Big]\Big\}.\end{aligned} \tag{52}$$

Comparing the derivative to zero yields:

$$\boldsymbol{\mu}_a = \hat{\mathbf{a}}(1). \tag{66}$$

6) M-$\boldsymbol{\Phi}_a$ Step: Differentiating $\mathcal{L}(\theta)$ w.r.t. $\boldsymbol{\Phi}_a$, we obtain

$$\begin{aligned}\frac{\partial \mathcal{L}(\theta)}{\partial \boldsymbol{\Phi}_a} &= \frac{\partial}{\partial \boldsymbol{\Phi}_a}\Bigg\{\mathbb{E}_{q(\mathbf{a}(1))}\Big[\ln p\big(\mathbf{a}(1);\boldsymbol{\mu}_a,\boldsymbol{\Phi}_a\big)\Big] \\ &\quad + \sum_{\ell=2}^{L}\mathbb{E}_{q(\mathbf{a}_{\{\ell:\ell-1\}})}\Big[\ln p\big(\mathbf{a}(\ell);\mathbf{a}(\ell-1);\boldsymbol{\Phi}_a\big)\Big]\Bigg\} \\ &= \boldsymbol{\Phi}_a^{-1}\mathbb{E}_{q(\mathbf{a}(1))}\Big[\big(\mathbf{a}(1)-\boldsymbol{\mu}_a\big)\big(\mathbf{a}(1)-\boldsymbol{\mu}_a\big)^{\mathrm{H}}\Big]\boldsymbol{\Phi}_a^{-1} \\ &\quad + \boldsymbol{\Phi}_a^{-1}\sum_{\ell=2}^{L}\mathbb{E}_{q(\mathbf{a}_{\{\ell:\ell-1\}})}\Big[\big(\mathbf{a}(\ell)-\mathbf{a}(\ell-1)\big) \\ &\quad \times\big(\mathbf{a}(\ell)-\mathbf{a}(\ell-1)\big)^{\mathrm{H}}\Big]\boldsymbol{\Phi}_a^{-1} - L\boldsymbol{\Phi}_a^{-1}. \end{aligned}\tag{67}$$

Using (66) and comparing to zero yields:

$$\begin{aligned}\boldsymbol{\Phi}_a = &\frac{1}{L}\Bigg[\boldsymbol{\Sigma}_a(1) + \sum_{\ell=2}^{L}\Big(\boldsymbol{Q}_{11}(\ell) - \boldsymbol{Q}_{21}(\ell) - \boldsymbol{Q}_{12}(\ell) \\ &+ \boldsymbol{Q}_{22}(\ell)\Big)\Bigg],\end{aligned}\tag{68}$$

where the four $\boldsymbol{Q}_{np}(\ell)$, $(n,p)\in\{1,2\}$ are $N\times N$ non-overlapping subblocks of the second-order joint posterior moment $\boldsymbol{Q}(\ell)$, defined in (40). The complete VEM algorithm for speech enhancement is summarized in Algorithm 1.

Algorithm 1 VEM algorithm for speech enhancement

Input $\{\mathbf{x}(\ell)\}_{\ell=1}^{L}$.
Initialize $\hat{s}(\ell), \Sigma_s(\ell), \hat{\tau}(\ell), \hat{\boldsymbol{T}}, \theta$.
repeat
E-step
E-$\mathbf{a}$ *step (Forward-Backward Algorithm):*
Set $\boldsymbol{\Phi}_f(1)$ with (72), $\mathbf{a}_f(1)$ with (73).
for $\ell: 2$ to L
Calculate $\boldsymbol{\Phi}_f(\ell)$ with (70), $\mathbf{a}_f(\ell)$ with (71).
end
Set $\boldsymbol{\Phi}_\beta(L)=\boldsymbol{\Phi}_f(L)$, $\mathbf{a}_\beta(L)=\mathbf{a}_f(L)$.
for $\ell: L-1$ to 1
Calculate $\boldsymbol{\Phi}_i(\ell)$ with (74).
Calculate $\boldsymbol{\Phi}_\beta(\ell)$ with (75), $\mathbf{a}_\beta(\ell)$ with (76).
end
Calculate $\boldsymbol{\Sigma}_a(\ell)$ with (38), $\hat{\mathbf{a}}(\ell)$ with (39).
Calculate $\boldsymbol{\Sigma}_\xi(\ell)$ with (78), $\mathbf{a}_\xi(\ell)$ with (79).
Then Calculate $\boldsymbol{Q}(\ell)$ with (40).
E-s step: Calculate $\Sigma_s(\ell)$ with (30), $\hat{s}(\ell)$ with (31).
E-τ *step:* Calculate $a_p(\ell)$ with (43), $b_p(\ell)$ with (44).
Then Calculate $\hat{\tau}(\ell)$ with (45).
E-T step: Calculate $\boldsymbol{W}_p$ with (48), v_p with (49).
Then Calculate $\hat{\boldsymbol{T}}$ with (50).
M-step
M-$\mathbf{a}$ *step:* Update $\boldsymbol{\mu}_a$ with (66), $\boldsymbol{\Phi}_a$ with (68).
M-τ *step:* Update $a_0(\ell)$ with (54), $b_0(\ell)$ with (56).
M-T step: Update ν_0 with (59), $\boldsymbol{W}_0$ with (64).
until convergence
return the estimated source $\hat{s}(\ell)$.

V. Experimental Study

In this section, the proposed algorithm is evaluated using both simulated and real room environments. In the simulation part we provide a comprehensive performance evaluation over various reverberation times, noise types and signal to noise ratio (SNR) levels. In addition, we also report the results on real recordings, to demonstrate the applicability of the proposed method in real-life conditions. We begin by presenting the performance measures and the baseline method. Then, we describe the blind initialization method for the VEM algorithm. In the subsequent sections we detail setups and results for each series of experiments.

A. Performance Measures

In this section we describe the performance measures that are used for evaluating the quality of performance. The speech quality was assessed with two common objective measures, namely perceptual evaluation of speech quality (PESQ) [35], and log-spectral distance (LSD). Both the PESQ and the LSD were measured by comparing the speech signal as received by the first microphone, namely $s(\ell,k)$, with its PM estimate $\hat{s}(\ell,k)$, given by (31). The quality of the RTF estimate was evaluated in terms of the normalized projection misalignment (NPM) measure [36], which is a widely used error measure to evaluate an estimated impulse response, disregarding possible gain error. The NPM is computed with:

$$\text{NPM [dB]} = 20\log_{10}\left[1-\left(\frac{\mathbf{h}^{\mathrm{H}}\hat{\mathbf{h}}}{\|\mathbf{h}\|\|\hat{\mathbf{h}}\|}\right)^2\right]. \tag{69}$$

B. Baseline Method

The baseline method is [16], where the speech signal and the RTF are modelled as hidden variables, but the covariances are treated as deterministic parameters. For comparison, we also evaluated the performance obtained by our previous algorithm in [24]. In [24], the noise is assumed to be a spatially homogeneous and spherically diffuse sound field, with a time-invariant covariance matrix modelled by a known spatial coherence matrix multiplied by an unknown noise power. Assuming a known microphone array geometry, the spatial coherence between microphones i,j is given by: $\Gamma_{ij}(k)=\text{sinc}\left(\frac{2\pi f_s k}{K}\frac{d_{ij}}{c}\right)$, where f_s is the sampling rate, d_{ij} is the inter-distance between microphones i and j and c is the sound velocity. This diffuse-based method, which estimates only the power of the noise, will be referred to as Prop. (diff.). The method proposed in this paper, which estimates the entire noise precision matrix, is denoted henceforth as Prop. (mat.).

The results of the MCWF in (32) with true parameters are also presented, in order to illustrate the efficiency of the proposed method. We refer to this algorithm as the *oracle* algorithm, since it knows a priori the RTF, the variance of the speaker and the covariance of the noise, information that

is not available to the fully blind VEM algorithm. For this method, the RTF and the speech variance were estimated from the clean microphone signal, while the noise covariance was estimated from the noise signal. The results obtained by this oracle method indicates the best attainable results.

C. Initialization

The proposed VEM algorithm requires the initialization of the prior parameters

$$\left\{\boldsymbol{\mu}_a(k), \boldsymbol{\Phi}_a(k), a_0(\ell,k), b_0(\ell,k), \nu_0(k), \boldsymbol{W}_0(k)\right\}_{\ell,k=1}^{L,K}.$$

as well as the posterior parameters

$$\left\{\hat{s}(\ell,k), \Sigma_s(\ell,k), a_p(\ell,k), b_p(\ell,k), \nu_p(k), \boldsymbol{W}_p(k)\right\}_{\ell,k=1}^{L,K}.$$

The proposed scheme has the great advantage of converging to the correct solution in a *fully blind* setup, where all initializations are either based on the observations or drawn randomly. The speech estimator was set to $\hat{s}(\ell,k) = x_1(\ell,k)$, i.e. the noisy signal at the first microphone, and the posterior variance was set to $\Sigma_s(\ell,k) = 0$. The prior channel parameters were initialized with $\boldsymbol{\mu}_a(k) = \mathbf{1}, \boldsymbol{\Phi}_a(k) = \epsilon\mathbf{I}$, where $\epsilon = 0.1$. The source precision parameters were initialized with $a_0(\ell,k) = 100$, $b_0(\ell,k) = 10^{-8}$, $b_p(\ell,k) = b_0(\ell,k) + |x_1(\ell,k)|^2$, and $a_p(\ell,k)$ was drawn randomly. The noise precision matrix parameters were initialized with $\nu_0(k) = 10$, $\boldsymbol{W}_0(k) = \delta\mathbf{I}$ where $\delta = 0.01$, $\boldsymbol{W}_p(k) = \left(\boldsymbol{W}_0^{-1}(k) + \sum_{\ell=1}^{L}\mathbf{x}(\ell)\mathbf{x}^{\mathrm{H}}(\ell)\right)^{-1}$ and $\nu_p(k)$ was drawn randomly. The initialization for Newton's method was $\nu_0^{(0)} = N$. Similar initialization was used for the baseline algorithm.

D. Simulations

1) Simulation Setup: To evaluate the performance of the proposed algorithm, we simulated the following static scenario. We downloaded time-invariant room impulse responses (RIRs) from the RIR database presented in [37]. The impulse responses were recorded in the $6 \times 6 \times 2.4$ m acoustic lab of the Engineering Faculty at Bar-Ilan University (BIU). The reverberation time was set to $T_{60} = 160$ ms and $T_{60} = 360$ ms by configuring 60 dedicated panels attached to the room walls, ceiling and floor. The measurements were recorded using an eight microphones uniform linear array (ULA) with microphone spacing of 8 cm. The loudspeaker was located in the end-fire direction (angle of 90°), at 1 m distance from the array. The room setup is illustrated in Fig. 2. To construct the microphone signals, we convolved the clean speech signals of five male and five female speakers from the TIMIT database [38] with the corresponding RIR, where each speaker utter sentence of $3-5$ sec long. For the additive noise, we examine three different types of noise sources: i) Speech-like Diffuse noise: An artificial diffuse noise with speech-like spectrum was created by applying the method presented in [39] on a stationary noise signal with a speech-like spectrum; ii) Babble Diffuse noise: Applying [39] on a babble noise signal from the NOISEX-92 database [40]; and iii) Directional noise: A noise source was positioned at a distance of 1 m from the array at angle of 45°, as illustrated in Fig. 2. To generate a directional noise signal we convolved a noise signal (with speech-like spectrum) with the corresponding RIR. We also added an additional white noise with a directional noise-to white noise ratio of 15 dB, in order to guarantee the invertibility of the noise covariance matrix. Finally, the microphone signals were synthesized by adding the noise signal to the reverberant speech signals with several SNR levels.

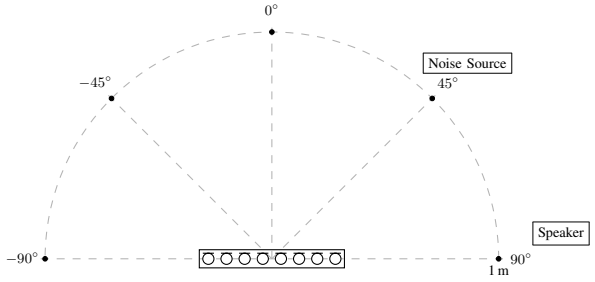


Fig. 2: Geometric setup.

The following values of the parameters were common for both simulation and real recordings. The sampling rate was set to 16 kHz, the STFT was computed with windows of 128 ms and 32 ms overlap between adjacent time frames. The number of frequency bins was 2048. The VEM algorithm was implemented using 100 iterations. For the Newton search, 5 iterations were applied.

All experiments described in this paper were run in MATLAB R2016b on a HP Compaq Elite 8300 PC, with an Intel 4-core (8 threads) i7-3770 CPU at 3.4 GHz and 16 GB RAM. In order to process a 2 sec recording at sampling rate of 16 kHz with 8 microphones, the average running time of the matrix-based proposed algorithm is 31 sec per iteration, the proposed diffuse-based algorithm [24] requires 30 sec per iteration, while the baseline method [16] requires 13 sec per iteration. The computational complexity of the proposed method is therefore about 2.4 times higher than that of the baseline method.

2) Results of Simulations: The performance measures were calculated by averaging over 10 trials with different speakers. Tables I, II, and III summarize the results for the speech-like diffuse noise case, the babble diffuse noise case and the directional noise case, respectively. NPM results are summarized in Tables IV, V and VI. The best results are highlighted in boldface. The proposed method outperforms the baseline algorithm for all noise types in all SNR levels, and for both reverberation levels. It can also be noted that in the diffuse noise case, the proposed matrix-based method is slightly superior to the diffuse-based method [24], except for the case of babble noise with SNR = 0dB. However, in the directional noise case, the wrong diffuseness assumption of [24] degrades the results significantly.

E. Real Recordings

1) Real Recordings Setup: The performance of the proposed method was also verified on real-life recordings, demonstrating the feasibility of the algorithm in real conditions. The recordings were taken in the BIU acoustic lab. The signals

TABLE I: Quality Measures for Simulations with Speech-Like Diffuse Noise

		$T_{60}=160$msec			$T_{60}=360$msec		
	Alg.\SNR	0dB	5dB	10dB	0dB	5dB	10dB
PESQ	Unprocessed	1.3	1.5	1.8	1.3	1.6	2.0
	Baseline [16]	1.8	1.9	2.0	1.8	2.0	2.0
	Prop. (diff.) [24]	2.1	2.6	3.1	2.1	2.6	3.1
	Prop. (mat.)	**2.2**	**2.9**	**3.4**	**2.4**	**3.2**	**3.6**
	Oracle MCWF	3.5	3.7	3.8	3.4	3.4	3.5
LSD	Unprocessed	10.4	7.2	4.6	10.7	7.5	4.8
	Baseline [16]	5.2	5.2	5.2	5.8	5.8	5.8
	Prop. (diff.) [24]	**3.6**	2.9	2.8	4.1	3.6	3.5
	Prop. (mat.)	**3.6**	**2.5**	**2.5**	**4.0**	**3.0**	**3.0**
	Oracle MCWF	1.7	1.2	0.9	2.4	2.0	1.8

TABLE II: Quality Measures for Simulations with Babble Diffuse Noise

		$T_{60}=160$msec			$T_{60}=360$msec		
	Alg.\SNR	0dB	5dB	10dB	0dB	5dB	10dB
PESQ	Unprocessed	1.4	1.6	1.9	1.4	1.7	2.0
	Baseline [16]	1.8	1.8	1.9	1.8	1.9	2.0
	Prop. (diff.) [24]	**2.1**	2.7	3.2	**2.2**	2.8	3.2
	Prop. (mat.)	**2.1**	**2.8**	**3.3**	**2.2**	**2.9**	**3.5**
	Oracle MCWF	3.6	3.8	3.9	3.5	3.6	3.6
LSD	Unprocessed	7.5	5.4	3.6	7.3	5.3	3.6
	Baseline [16]	5.3	5.2	5.2	5.8	5.8	5.7
	Prop. (diff.) [24]	**3.3**	2.9	2.8	**3.9**	3.6	3.5
	Prop. (mat.)	3.8	**2.8**	**2.7**	4.2	**3.3**	**3.1**
	Oracle MCWF	1.2	0.9	0.7	2.0	1.8	1.7

TABLE III: Quality Measures for Simulations with Directional Noise

		$T_{60}=160$msec			$T_{60}=360$msec		
	Alg.\SNR	0dB	5dB	10dB	0dB	5dB	10dB
PESQ	Unprocessed	1.3	1.5	1.8	1.3	1.6	2.0
	Baseline [16]	2.0	2.0	2.1	2.0	2.1	2.1
	Prop. (diff.) [24]	1.4	1.7	2.1	1.5	1.8	2.3
	Prop. (mat.)	**2.1**	**3.0**	**3.5**	**2.1**	**2.9**	**3.5**
	Oracle MCWF	2.7	3.0	3.4	2.7	3.1	3.4
LSD	Unprocessed	11.5	8.1	5.2	11.7	8.3	5.3
	Baseline [16]	5.6	5.5	5.4	6.3	6.1	6.0
	Prop. (diff) [24]	7.3	4.8	3.6	7.8	5.3	4.0
	Prop. (mat.)	**3.6**	**2.7**	**2.7**	**3.8**	**3.2**	**3.1**
	Oracle MCWF	3.3	2.5	1.8	4.0	3.2	2.4

TABLE IV: NPM Results for Simulations with Speech-Like Diffuse Noise

	$T_{60}=160$msec			$T_{60}=360$msec		
Alg.\SNR	0dB	5dB	10dB	0dB	5dB	10dB
Baseline [16]	-1.6	-2.6	-3.3	-1.6	-2.4	-2.8
Prop. (diff.) [24]	-3.6	-5.4	-6.7	-3.7	-5.4	-6.6
Prop. (mat.)	**-5.3**	**-7.5**	**-8.7**	**-5.3**	**-7.2**	**-8.0**

TABLE V: NPM Results for Simulations with Babble Diffuse Noise

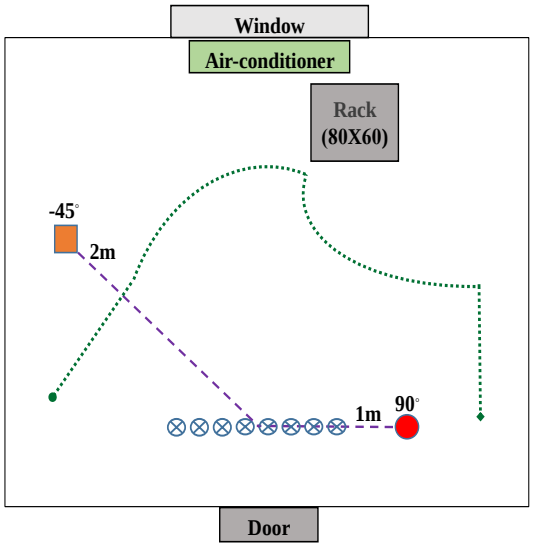
	$T_{60}=160$msec			$T_{60}=360$msec		
Alg.\SNR	0dB	5dB	10dB	0dB	5dB	10dB
Baseline [16]	-2.2	-2.8	-3.3	-2.0	-2.4	-2.6
Prop. (diff.) [24]	-5.8	-7.0	**-7.8**	-5.7	-6.7	**-7.3**
Prop. (mat.)	**-6.1**	**-7.3**	**-7.8**	**-5.9**	**-6.8**	-6.9

TABLE VI: NPM Results for Simulations with Directional Noise

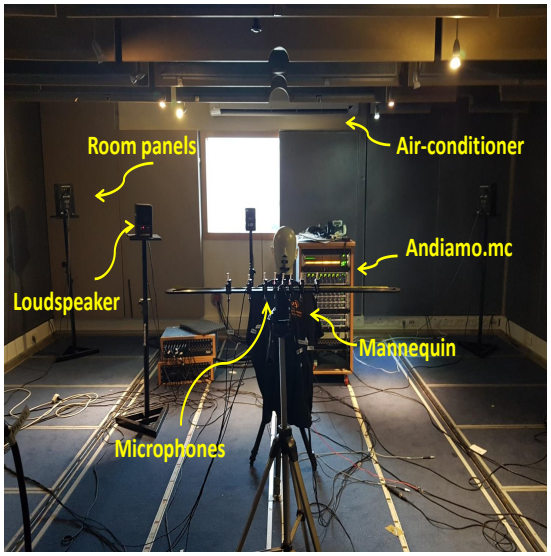
	$T_{60}=160$msec			$T_{60}=360$msec		
Alg.\SNR	0dB	5dB	10dB	0dB	5dB	10dB
Baseline [16]	-2.1	-2.9	-3.4	-1.9	-2.4	-2.7
Prop. (diff.) [24]	-1.1	-2.0	-3.7	-1.2	-2.2	-3.8
Prop. (mat.)	**-4.6**	**-7.4**	**-8.5**	**-4.8**	**-6.8**	**-7.5**

were measured by the CK32 omnidirectional microphones of AKG. Eight microphones were mounted on a metal ruler, with inter-distances of $[8, 6, 4, 3, 3, 4, 6]$ cm. For the recording we used the Hammerfall HDSPe MADI audio card of RME, where an ANDIAMO.MC acquisition panel was used as a microphone preamp and A/D converter. A Fostex 6301BX loudspeaker was utilized as a noise source. The signals were measured with 24-bits resolution and 48kHz sampling rate, and then were downsampled to 16 kHz, according to the speech signals bandwidth. The room panels were adjusted to create two reverberation levels, namely $T_{60} = 200$ ms or $T_{60} = 400$ ms. A first series of experiments was conducted with speech utterances that were drawn from the TIMIT database [38]. Real human speakers were first emulated using a Head and Torso Simulator (HATS) mannequin with built-in mouth (Type 4128C- B&K), which emulates the acoustic properties of an average human head and torso. The utterances were played in the room using the mannequin, which was located at an angle of 90°, a 1 m distance from the array. The room layout is illustrated in Fig. 3(a), and a picture of the room setup and the equipment is shown in Fig. 3(b). Ten different female and male speakers were used, each uttering $3-5$ sec long signal. Then, a second series was conducted with real-life speakers. This series consists of one male and one female speakers, which were asked to utter speech in static and dynamic scenarios. In the static scenario, the speaker was positioned 1 m in front of the array, at angle of 90°, and was asked to utter a single sentence. We also examined the capability of the proposed method to track a time-varying RTF series of a moving speaker. We devised a dynamic scenario, in which the speaker was asked to utter speech while moving on the trajectory drawn in Fig. 3(a). The speaker walked quite quickly and the entire movement lasted 15 s.

For the noise, we examined two different types of recorded noise: i) Air-conditioner noise; and ii) Directional-noise: Noise signal with a speech-like spectrum from the NOISEX-92 database was played from a loudspeaker located at a distance



(a)



(b)

Fig. 3: (a) The room sketch: microphone locations are depicted with blue 'x' marks, position of the speaker is marked by red circle, and position of noise loudspeaker is marked by orange square. The trajectory for the dynamic scenario is depicted by green dash lines. (b) A photograph of the room.

of 2 m from the array at angle of $-45°$, as illustrated in Fig. 3(a). The noise signals were recorded separately and then were added with various SNR levels to the measured speech signals.

2) Results of Static Real-Life Speakers: PESQ and LSD scores for the human speakers in the static scenario are presented in Tables VII and VIII for the air-conditioner noise case and the directional noise case, respectively. Each reported measure is the average over the two speakers. It can be seen that the proposed method outperforms the baseline method, and obtains a higher speech quality and lower distortion level for both noise types. The advantage of the proposed

TABLE VII: Quality Measures for Real-Life Static Speakers with Air-Conditioner Noise

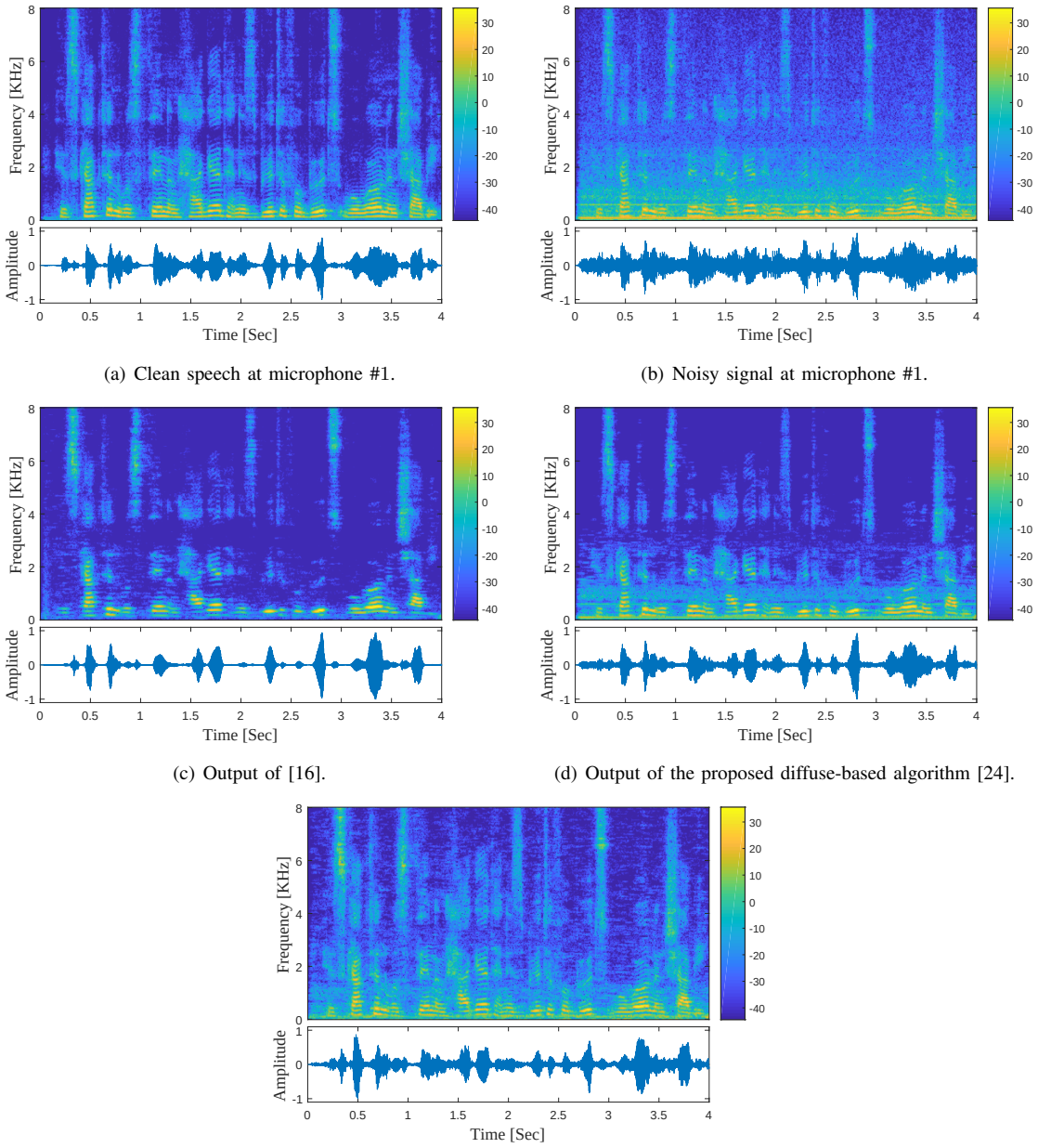
		$T_{60}=200$msec			$T_{60}=400$msec		
	Alg.\SNR	0dB	5dB	10dB	0dB	5dB	10dB
PESQ	Unprocessed	2.0	2.4	2.9	1.4	1.6	2.0
	Baseline [16]	2.4	2.4	2.4	1.8	2.0	2.0
	Prop. (diff.) [24]	2.6	3.0	**3.5**	1.7	2.2	**2.8**
	Prop. (mat.)	**2.8**	**3.3**	**3.5**	**1.9**	**2.4**	**2.8**
	Oracle MCWF	3.7	3.9	4.0	3.3	3.6	3.7
LSD	Unprocessed	5.0	3.7	2.5	6.9	5.1	3.6
	Baseline [16]	3.8	3.7	3.6	5.8	5.2	4.8
	Prop. (diff.) [24]	**2.9**	2.4	**2.2**	**4.1**	**3.2**	**2.7**
	Prop. (mat.)	**2.9**	**2.3**	**2.2**	5.1	3.7	3.2
	Oracle MCWF	1.5	1.3	1.2	2.2	1.9	1.7

TABLE VIII: Quality Measures for Real-Life Static Speakers with Directional Noise

		$T_{60}=200$msec			$T_{60}=400$msec		
	Alg.\SNR	0dB	5dB	10dB	0dB	5dB	10dB
PESQ	Unprocessed	1.5	1.8	2.2	1.4	1.7	2.1
	Baseline [16]	2.2	2.3	2.4	2.0	2.0	2.1
	Prop. (diff.) [24]	1.9	2.4	2.9	1.7	2.2	2.8
	Prop. (mat.)	**2.9**	**3.3**	**3.3**	**2.5**	**2.9**	**3.0**
	Oracle MCWF	3.2	3.5	3.7	3.3	3.5	3.7
LSD	Unprocessed	6.5	4.6	2.9	6.3	4.6	3.0
	Baseline [16]	4.0	3.9	3.8	5.1	4.9	4.7
	Prop. (diff.) [24]	3.7	2.8	**2.5**	**3.9**	**3.1**	**2.7**
	Prop. (mat.)	**3.1**	**2.6**	**2.5**	4.3	3.3	3.0
	Oracle MCWF	1.7	1.4	1.1	2.1	1.8	1.6

method is clearly demonstrated in all considered scenarios. Similar trends are obtained for both simulation and real-life recordings. It can also be noted that the proposed matrix-based method is superior to the diffuse-based method [24], especially for the directional noise case, except for the LSD measure in $T_{60} = 400$msec. Fig. 4 depicts sonogram examples of the various signals for air-conditioner noise with input SNR of 0 dB at $T_{60} = 200$ ms. Fig. 4(a) shows s, the clean speech signal as received by the first microphone. Fig. 4(b) depicts x_1, the noisy signal at the first microphone. The estimated speech signal obtained by the baseline method [16] is shown in Fig. 4(c). The output of the proposed diffuse-based method [24] is depicted in Fig. 4(d). Fig. 4(e) depicts $\hat{s}$, the estimated speech of the matrix-based proposed algorithm. We can conclude that the proposed algorithm reduces noise while maintaining low distortion. In contrast, the estimator proposed by [16] provides aggressive noise reduction, at the cost of severe speech distortion. The proposed diffuse-based method [24] produces less speech distortion than the baseline method [16]. However, it is still slightly inferior compared to the matrix-based proposed algorithm. Informal listening tests verify that the output of the proposed method is a more natural denoised speech, in comparison to the output of the baseline method [16]. Audio examples can be found in our website.[1]

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



(a) Clean speech at microphone #1.

(b) Noisy signal at microphone #1.

(c) Output of [16].

(d) Output of the proposed diffuse-based algorithm [24].

(e) Output of the matrix-based proposed algorithm.

Fig. 4: Sonograms of a real-life example, with $T_{60} = 200$ ms and SNR of 0 dB.

The results obtained for the TIMIT utterances that were played from the mannequin are pretty similar, and were omitted here due to lack of space.

3) Results of Moving Speakers: Table IX reports the PESQ and LSD scores for a moving speaker in the air-condition noise case and the directional noise case. Despite the challenging scenario, the proposed algorithm produces a high-quality estimate of the speech signal.

In order to further demonstrate the quality of the estimated time-varying RTFs, we use it to estimate the direction of arrival (DOA) of the moving speaker. We first find the peak of the relative AIR (obtained by applying the inverse FFT to the RTF). Assuming a low reverberation level, the peak location corresponds to the time difference of arrival (TDOA) w.r.t. the reference microphone, and thus can be utilized for estimating the source DOA. The tracking ability is illustrated in Fig. 5, where the estimated DOA is presented as a function of time compared to the true trajectory, which was estimated from the clean measured signal using the peak of an instantaneous least-squares (LS) estimate of the RTF.

F. Analysis of the Noise Covariance Estimation

In this section we briefly discuss the benefit of estimating the entire noise covariance matrix, rather than assuming a diffuse spatial coherence matrix and estimating only the noise power as in [24]. On the one hand, the proposed diffuse-based method [24] offers slightly lower computational complexity compared to the proposed matrix-based method, due to the

TABLE IX: Quality Measures for Real-Life Moving Speaker, $T_{60} = 200$ msec

	Noise Type	Air-conditioner			Directional		
	Alg.\SNR	0dB	5dB	10dB	0dB	5dB	10dB
PESQ	Unprocessed	1.9	2.3	3.0	1.4	1.7	2.2
	Baseline [16]	2.3	2.5	2.5	2.1	2.3	2.4
	Prop. (diff.) [24]	2.2	2.7	**3.3**	1.6	2.0	2.6
	Prop. (mat.)	**2.6**	**3.1**	**3.3**	**2.4**	**2.8**	**3.1**
	Oracle MCWF	3.9	4.1	4.3	3.0	3.4	3.7
LSD	Unprocessed	4.7	3.4	2.4	6.3	4.5	3.0
	Baseline [16]	4.6	4.2	4.0	4.8	4.6	4.3
	Prop. (diff.) [24]	3.2	2.5	2.2	4.0	3.0	2.4
	Prop. (mat.)	**2.8**	**2.3**	**2.1**	**3.3**	**2.5**	**2.3**
	Oracle MCWF	1.6	1.2	0.9	2.2	1.7	1.3

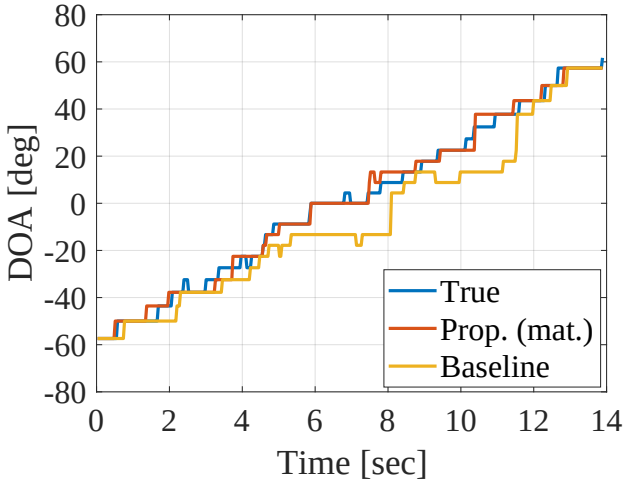


Fig. 5: Estimated DOA of a moving speaker.

reduced number of parameters to be estimated. On the other hand, from the quality measures in Tables I-IX it can be observed that the estimation of the entire noise covariance matrix yields better results than estimating only the power of the noise, especially for the directional noise case, except for the LSD in $T_{60} = 400$msec. This can be explained by the fact that the diffuse model becomes more accurate as the reverberation level increases.

VI. Conclusions

In this contribution, we presented a fully Bayesian hierarchical framework for the challenging problem of blind multichannel speech enhancement with time-varying audio channel. The model extends [16] to include the speech precision and the noise precision matrix as part of the hidden data. The set of latent variables consists of the speech signal (as received by a reference microphone), the RTF, the speech precision and the noise precision matrix. The inference of the hidden variables is performed using a VEM algorithm, leading to a variant of MCWF for estimating the source and a Kalman smoother for the time-varying acoustic channel. The speech estimator was decomposed into an MVDR beamformer followed by a variational postfilter. The proposed algorithm is fully-blind and uses only the available data. The discussion is supported by an extensive experimental study based on both simulated data and real-life recording, for static and dynamic scenarios. The proposed method outperforms the baseline method in terms of both objective speech quality scores and informal listening tests. In terms of NPM and tracking ability, the proposed algorithm offers a superior channel estimate compared to [16].

VII. Appendix

In this appendix, we detail the recursions for the forward and backward passes, as well as the pairwise statistics for the joint posterior distribution of two successive frames.

1) Forward pass: The forward distribution writes [13] $p\big(\mathbf{a}(\ell), \boldsymbol{\mu}_L(1{:}\ell)\big) = \mathcal{N}_c\big(\mathbf{a}(\ell); \mathbf{a}_f(\ell), \boldsymbol{\Phi}_f(\ell)\big)$, where

$$\boldsymbol{\Phi}_f(\ell) = \Big[\boldsymbol{\Phi}_L^{-1}(\ell) + \big(\boldsymbol{\Phi}_a + \boldsymbol{\Phi}_f(\ell-1)\big)^{-1}\Big]^{-1}, \tag{70}$$

$$\mathbf{a}_f(\ell) = \boldsymbol{\Phi}_f(\ell)\Big[\boldsymbol{\Phi}_L^{-1}(\ell)\boldsymbol{\mu}_L(\ell) + \big(\boldsymbol{\Phi}_a + \boldsymbol{\Phi}_f(\ell-1)\big)^{-1}\mathbf{a}_f(\ell-1)\Big], \tag{71}$$

with the initialization

$$\boldsymbol{\Phi}_f(1) = \big(\boldsymbol{\Phi}_L^{-1}(1) + \boldsymbol{\Phi}_a^{-1}\big)^{-1}, \tag{72}$$

$$\mathbf{a}_f(1) = \boldsymbol{\Phi}_f(1)\left(\boldsymbol{\Phi}_L^{-1}(1)\boldsymbol{\mu}_L(1) + \boldsymbol{\Phi}_a^{-1}\boldsymbol{\mu}_a\right). \tag{73}$$

2) Backward pass: The backward distribution writes $p\big(\boldsymbol{\mu}_L(\ell+1{:}L)|\mathbf{a}(\ell)\big) = \mathcal{N}_c\big(\mathbf{a}(\ell); \mathbf{a}_\beta(\ell), \boldsymbol{\Phi}_\beta(\ell)\big)$, where

$$\boldsymbol{\Phi}_i(\ell) = \Big(\boldsymbol{\Phi}_L^{-1}(\ell+1) + \boldsymbol{\Phi}_\beta^{-1}(\ell+1)\Big)^{-1}, \tag{74}$$

$$\boldsymbol{\Phi}_\beta(\ell) = \boldsymbol{\Phi}_a + \boldsymbol{\Phi}_i(\ell), \tag{75}$$

$$\mathbf{a}_\beta(\ell) = \boldsymbol{\Phi}_i(\ell)\Big(\boldsymbol{\Phi}_L^{-1}(\ell+1)\boldsymbol{\mu}_L(\ell+1) + \boldsymbol{\Phi}_\beta^{-1}(\ell+1)\mathbf{a}_\beta(\ell+1)\Big). \tag{76}$$

The backward recursion starts from the last time step L, with the last update from the forward pass as initialization: $\boldsymbol{\Phi}_\beta(L) = \boldsymbol{\Phi}_f(L)$, $\mathbf{a}_\beta(L) = \mathbf{a}_f(L)$.

3) Marginal posterior: The forward and backward probability distributions are then combined to obtain the posterior marginal distribution:

$$q\big(\mathbf{a}(\ell)\big) \propto p\big(\mathbf{a}(\ell), \boldsymbol{\mu}_L(1{:}\ell)\big)p\big(\boldsymbol{\mu}_L(\ell+1{:}L)|\mathbf{a}(\ell)\big). \tag{77}$$

4) Joint posterior distribution of two consecutive frames: Marginalizing out all other frames in (36) leads to a Gaussian distribution, with the following covariance and mean:

$$\boldsymbol{\Sigma}_\xi(\ell) = \begin{bmatrix} \boldsymbol{\Phi}_i^{-1}(\ell-1) + \boldsymbol{\Phi}_a^{-1} & -\boldsymbol{\Phi}_a^{-1} \\ -\boldsymbol{\Phi}_a^{-1} & \boldsymbol{\Phi}_f^{-1}(\ell-1) + \boldsymbol{\Phi}_a^{-1} \end{bmatrix}^{-1}, \tag{78}$$

$$\mathbf{a}_\xi(\ell) = \boldsymbol{\Sigma}_\xi(\ell)\begin{bmatrix} \boldsymbol{\Phi}_i^{-1}(\ell-1)\mathbf{a}_\beta(\ell-1) \\ \boldsymbol{\Phi}_f^{-1}(\ell-1)\mathbf{a}_f(\ell-1) \end{bmatrix}. \tag{79}$$

References

[1] 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.

[2] 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.

[3] 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.

[4] O. Shalvi and E. Weinstein, "System identification using nonstationary signals," *IEEE Tran. on Signal Processing*, vol. 44, no. 8, pp. 2055–2063, 1996.

[5] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Tran. on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1071–1086, 2009.

[6] I. Cohen, S. Gannot, and B. Berdugo, "An integrated real-time beamforming and postfiltering system for nonstationary noise environments," *EURASIP Journal on Applied Signal Processing*, vol. 2003, pp. 1064–1073, 2003.

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

[8] 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.

[9] 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.

[10] N. Q. Duong, E. Vincent, and R. Gribonval, "Under-determined 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.

[11] 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.

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

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

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

[15] 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.

[16] 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.

[17] 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.

[18] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Speech dereverberation with convolutive transfer function approximation using map and variational deconvolution approaches," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014, pp. 50–54.

[19] 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.

[20] K. Adiloğlu and E. Vincent, "Variational bayesian inference for source separation and robust feature extraction," *IEEE/ACM Tran. on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1746–1758, 2016.

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

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

[23] L. Girin and R. Badeau, "On the use of latent mixing filters in audio source separation," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2017, pp. 225–235.

[24] Y. Laufer and S. Gannot, "A Bayesian hierarchical model for speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[25] R. A. Wooding, "The multivariate distribution of complex normal variables," *Biometrika*, vol. 43, no. 1/2, pp. 212–215, 1956.

[26] N. R. Goodman, "Statistical analysis based on a certain multivariate complex gaussian distribution (an introduction)," *The Annals of mathematical statistics*, vol. 34, no. 1, pp. 152–177, 1963.

[27] F. D. Neeser and J. L. Massey, "Proper complex random processes with applications to information theory," *IEEE transactions on information theory*, vol. 39, no. 4, pp. 1293–1302, 1993.

[28] S. N. Anfinsen, A. P. Doulgeris, and T. Eltoft, "Estimation of the equivalent number of looks in polarimetric synthetic aperture radar imagery," *IEEE Tran. on Geoscience and Remote Sensing*, vol. 47, no. 11, pp. 3795–3809, 2009.

[29] 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.

[30] S. Leglaive, R. Badeau, and G. Richard, "Student's t source and mixing models for multichannel audio source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

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

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

[33] O. Schwartz, S. Gannot, and E. A. P. 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.

[34] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[35] 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.

[36] 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.

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

[38] 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," Disc, 1993.

[39] E. A. P. 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.

[40] 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.