Variational Inference for DOA Estimation in Reverberant Conditions

Yosef Soussana and Sharon Gannot
Faculty of Engineering, *Bar-Ilan University*, Ramat-Gan, Israel
{sharon.gannot,yosef.sussany}@biu.ac.il

***Abstract*—A concurrent speaker direction of arrival (DOA) estimator in a reverberant environment is presented. The reverberation phenomenon, if not properly addressed, is known to degrade the performance of DOA estimators. In this paper, we investigate a variational Bayesian (VB) inference framework for clustering time-frequency (TF) bins to candidate angles. The received microphone signals are modelled as a sum of anechoic speech and the reverberation component. Our model relies on Gaussian prior for the speech signal and Gamma prior for the speech precision. The noise covariance matrix is modelled by a time-invariant full-rank coherence matrix multiplied by time-varying gain with Gamma prior as well. The benefits of the presented model are verified in a simulation study using measured room impulse responses.**

***Index Terms*—DOA estimation, Variational Bayes inference, Variational Expectation-Maximization**

I. Introduction

Estimation of the DOAs of sound sources is a fundamental problem in signal processing for several decades. In acoustic enclosures, the signal acquired by the microphone array is usually distorted by reverberation, a result of the multiple sound reflections on the room facets and objects, as well as contaminated by noise. These phenomena, if not treated properly, may cause a severe degradation in the estimation accuracy of the DOA. DOA is often estimated by employing time difference of arrival (TDOA) estimator as a preliminary stage. A classical approach for estimating the TDOA between two observed speech signals is by cross-correlation, most commonly its normalized version, the generalized cross correlation (GCC) [1] with phase transform (PHAT) normalization. A generalization of the GCC-PHAT to microphone arrays in far-field scenarios is the steered response power (SRP)-PHAT [2].

The coexistence of multiple speakers in the same (reverberant) environment may further degrade the performance. Hence, the problem of multiple DOA estimation has been the focus of recent research efforts. In the model-based EM source separation and localization (MESSL) approach [3], [4], the authors use the Expectation-Maximization (EM) procedure for Mixture of Gaussians (MoG) clustering of the TDOAs of multiple speakers, considering the two microphone (binaural) case, and using the W-disjoint property of speech signals [5]. In the E-step, a time-frequency mask, associating each TF bin to a specific Gaussian, is estimated. In the M-step the mixture weights are estimated, using the number of associations of TF bins. The final DOA estimates are selected by choosing the most probable Gaussians. A generalization of [4] is presented in [6], by estimating the speakers coordinates rather than their DOAs. This approach uses the phase ratios between the signals received by microphone pairs. Additionally, efficient tracking algorithms for multiple moving speakers are presented. Another MoG-based approach directly uses the speech spectrogram with explicit modelling of the reverberation properties [7]. While the mean of the Gaussians is zero, the covariance-matrices of the Gaussians comprise an explicit spatial modelling of the speech and the reverberation [8]. Variant of this work is proposed in [9] with special treatment to the additive noise.

In the current work, we consider VB methodology, to apply a grid-based MoG clustering directly to the short-time Fourier transform (STFT) of the microphone signals. We introduce a Bayesian modelling with appropriate conjugate priors to the all problem parameters, including explicit treatment of the reverberation phenomenon. Iterative variational expectation-maximization (VEM) algorithm is then developed for estimating the model parameters. In the E-step the posterior distributions of the inferred parameters are calculated given the observations, and in the M-step the hyper-parameters are computed given these posterior distributions. A Bayesian variant of the MESSL technique was already studied in the literature [10]. In this approach, VB is used to cluster TF bins by associating their interaural phase differences (IPDs) with a candidate TDOA, in order to preform separation. The model does not take into account the reverberation phenomenon.

II. Problem Formulation

A. Signal model

The proposed model is formulated in the STFT domain, where $k = 0, \dots, K-1$ denotes the frequency band and $t = 0, \dots, T-1$ denotes the time frame. Consider an arbitrary scenario with J unknown speakers and N microphones in noiseless environment. Let $z_n(t,k)$ be the signals received by the nth microphone, with $n = 1, \dots, N$:

$$z_n(t,k) = \sum_{i=1}^{J} g_{j,n}(k) s_j(t,k) + r_{j,n} \tag{1}$$

where $s_j(t,k)$ denotes the anechoic speech signal of the jth speaker, which is modelled as a zero-mean Gaussian process with time-varying precision

$$p(S_j(t,k)|\tau_j(t,k)) = \mathcal{N}_c(S_j(t,k); 0, \tau_j^{-1}(t,k)),$$

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

$g_{j,n}(k)$ denotes the relative direct transfer function (RDTF) from the jth speaker to the nth microphone w.r.t. the reference microphone (the first microphone is arbitrarily chosen), and $r_{j,n}$ denotes the reverberation tail associated with the jth speaker. Concatenating all channels in a vector we obtain:

$$\mathbf{z}(t,k) = \sum_{i=1}^{J} \mathbf{g}_j(k) s_j(t,k) + \mathbf{r}_j(t,k)$$

where

$$\begin{aligned} \mathbf{z}(t,k) &= [z_1(t,k), \ldots, z_N(t,k)]^T \\ \mathbf{g}_j(k) &= [g_{j,1}(t,k), \ldots, g_{j,N}(t,k)]^T \\ \mathbf{r}_j(t,k) &= [r_{j,1}(t,k), \ldots, r_{j,N}(t,k)]^T. \end{aligned}$$

For a linear array and under far-field regime, the RDTF can be expressed as $g_{j,n}(k) = \exp(-i\frac{2\pi}{K}\frac{\xi_{j,n}}{T_s})$, where $i = \sqrt{-1}$, T_s denotes the sampling time and $\xi_{j,n}$ denotes the TDOA, associated with the jth speaker, between the nth microphone and the reference microphone. The TDOA can be expressed as $\xi_{j,n} = \frac{d_n \cos \vartheta_j}{c}$, where c is the sound velocity, d_n is the distance between the nth microphone and the reference microphone, and ϑ_j is the angle of arrival of the jth speaker.

B. Statistical Model

The core idea of the proposed method, inspired by [4], is to define a set of candidate angles and to determine the probability of the measured source signals to impinge the array from the candidate angles. Note that this strategy will yield an indirect estimate of the DOA, by selecting the proper peaks of the estimated probability map. The various speakers are assumed to exhibit a disjoint activity in the STFT domain [5]. Therefore, by means of clustering, each TF bin of $\mathbf{z}(t,k)$ can be associated with only a single active source. Hence, based on the above arguments, the observations can be described in the following probabilistic MoG description:

$$\mathbf{z}(t,k) \sim \sum_{m=1}^{M} \psi_m \mathcal{N}_c(\mathbf{z}(t,k); s(m,t,k)\mathbf{g}_m(k); \boldsymbol{\phi}_r(m,t,k)) \tag{2}$$

where ψ_m is the probability of a speaker to impinge the array from a candidate angle ϑ_m, with M being the total number of candidate angles. $s(m,t,k)$ is now the speech signal, similarly to II-A, but with an association to candidate angle ϑ_m. Note, that by our model, the J sources can be located in any of the M angles, assuming $M > J$. The RDTF $\mathbf{g}_m(k)$ is accordingly defined as the candidate source DOA and ϑ_m its respective angle. Note the differences between $\mathbf{g}_m(k)$ and ϑ_m with $m = 1, \ldots, M$, which are defined per candidate angle, and the corresponding quantities $\mathbf{g}_j(k)$ and ϑ_j with $j = 1, \ldots, J$, which are defined per actual source.

We further assume that the spatial covariance matrix of the reverberation term $\boldsymbol{\phi}_r(m,t,k)$ can be decomposed to a time-invariant spatial coherence matrix multiplied by a time-varying power. This can be justified if the source-microphone constellation is fixed. It is common to assume that the reverberation is spatially homogeneous and spherically isotropic sound-field, hence $\boldsymbol{\phi}_{\boldsymbol{r}}(m,t,k)) = \beta^{-1}(m,t,k)\boldsymbol{\Gamma}(k)$, with $\Gamma_{ij}(k) =\text{sinc}\left(\frac{2\pi k}{K}\frac{d_{i,j}}{T_s c}\right) + \varepsilon\delta(i-j)$, where $\text{sinc}(x) = \frac{\sin(x)}{x}$, $d_{i,j}$ is the inter-distance between the array elements and ε is a known diagonal loading representing uncorrelated spatial noise (see e.g. [11]). The inverse of the *unknown* reverberation power $\beta(m,t,k)$ is time-varying and associated with a speaker located at candidate angle m.

C. Conjugate Priors

In the Bayesian framework, it is common to introduce probabilistic priors over the latent variables to account for the model uncertainty. For the exponential family of distributions, choosing the so-called, *conjugate priors* leads to a posterior distribution with the same functional form as the original distribution and therefore to a simplified Bayesian analysis. In the Bayesian framework, it is often more convenient to work with precisions rather than variances [12]. We therefore first introduce prior distributions to the precision of the speaker and the reverberation signals. The conjugate prior of the precision of a univariate Gaussian is the Gamma probability [12]. Hence, the prior of the speech precision is given by:

$$p(\tau(m,t,k)) = \text{Gam}(\tau(m,t,k); b_0(m,t,k), c_0(m,t,k)). \tag{3}$$

Each TF bin in candidate angle ϑ_m is modelled with a unique shape, to allow more flexibility in the modelling of the speech characteristics. Similarly, we also assume Gamma distribution as the prior of the reverberation precision:

$$p(\beta(m,t,k)) = \text{Gam}(\beta(m,t,k); d_0(m,t,k), e_0(m,t,k)). \tag{4}$$

Finally, we choose the prior distribution for the MoG weights $\boldsymbol{\psi} = \{\psi_m\}_{m=1}^{M}$ to be the Dirichlet distribution:

$$p(\boldsymbol{\psi}) = \text{Dir}(\boldsymbol{\psi}|\boldsymbol{\alpha}) \tag{5}$$

where the hyper-parameters $\boldsymbol{\alpha} = \text{vec}_m\{\alpha_m\}$ can be interpreted as the effective prior number of observations associated with each component of the mixture.

III. Variational EM for DOA estimation

In this section, a VEM procedure for estimating the DOAs of all sources is derived. In this work, we define a hidden variable $x(m,t,k)$ to be an indicator that TF bin (t,k) is associated with a speaker located at candidate angle ϑ_m, in accordance with the W-disjoint property of speech signals. The total number of indicators is $M \times T \times K$.

Define the augmented observations and parameters for all TF bins and angles as:

$$\begin{aligned} \mathcal{Z} &= \text{vec}_{t,k}\{\mathbf{z}(t,k)\} & \mathbf{x} &= \text{vec}_{m,t,k}\{x(m,t,k)\} \\ \mathbf{s} &= \text{vec}_{m,t,k}\{s(m,t,k)\} & \boldsymbol{\tau} &= \text{vec}_{m,t,k}\{\tau(m,t,k)\} \\ \boldsymbol{\psi} &= \text{vec}_m\{\psi_m\} & \boldsymbol{\beta} &= \text{vec}_{m,t,k}\{\beta(m,t,k)\} \end{aligned}$$

Then, one can express the conditional probabilities as [12]:

$$p(\mathcal{Z}|\mathbf{x},\mathbf{s},\boldsymbol{\beta}) = \prod_{m,t,k}^{M,T,K} \cdots$$
$$\mathcal{N}_c(\mathbf{z}(t,k); s(m,t,k)\mathbf{g}_m(k), \beta^{-1}(m,t,k)\mathbf{\Gamma}(k))^{x(m,t,k)} \quad (6)$$

with

$$p(\mathbf{x}|\boldsymbol{\psi}) = \prod_{m,t,k}^{M,T,K} {\psi_m}^{x(m,t,k)} \quad (7)$$

where it is assumed that the observations for all time segments and all frequency bins are mutually independent.

Finally, define the hidden variables of the problem $\mathcal{H} = \{\mathbf{x}, \mathbf{s}, \boldsymbol{\tau}, \boldsymbol{\beta}, \boldsymbol{\psi}\}$ and its hyper-parameters:

$$\boldsymbol{\theta} = \{\alpha_m, b_0(m,t,k), c_0(m,t,k), \cdots$$
$$e_0(m,t,k), d_0(m,t,k)\}_{m,t,k=1}^{M,T,K}. \quad (8)$$

The VB inference necessitates the calculation of the posterior distribution $p(\mathcal{H}|\mathcal{Z};\boldsymbol{\theta}) = \frac{p(\mathcal{Z},\mathcal{H};\boldsymbol{\theta})}{p(\mathcal{Z};\boldsymbol{\theta}}$. Applying Bayes rule, the p.d.f. of the complete data can be expressed as:

$$p(\mathcal{Z},\mathcal{H};\boldsymbol{\theta}) = p(\mathcal{Z}|\mathbf{x},\boldsymbol{s},\boldsymbol{\beta})p(\mathbf{x}|\boldsymbol{\psi})p(\boldsymbol{\psi})p(\mathbf{s}|\boldsymbol{\tau})p(\boldsymbol{\tau})p(\boldsymbol{\beta}). \quad (9)$$

However, the likelihood $p(\mathcal{Z};\boldsymbol{\theta}) = \int p(\mathcal{Z},\mathcal{H};\boldsymbol{\theta})d\mathcal{H}$ cannot be evaluated in closed-form, hence neither the posterior $p(\mathcal{H}|\mathcal{Z};\boldsymbol{\theta})$ can be inferred. To alleviate this problem, we adopt the variational inference framework, in which the posterior $p(\mathcal{H}|\mathcal{Z};\boldsymbol{\theta})$ is approximated by $q(\mathcal{H})$ that can be decomposed into conditionally independent variables in accordance with the *mean field theorem* [12]:

$$q(\mathcal{H}) \approx q(\mathbf{x})q(\mathbf{s})q(\boldsymbol{\tau})q(\boldsymbol{\beta})q(\boldsymbol{\psi}). \quad (10)$$

Given the factorization of $q(\mathcal{H})$ over a partition of the hidden variables, the optimal marginal posterior distribution over a subset $\mathcal{H}_\ell \subseteq \mathcal{H}$ can be computed in the E-step by:

$$\ln q(\mathcal{H}_\ell) = \mathbb{E}_{q(\mathcal{H}/\mathcal{H}_\ell)}\left\{\ln p(\mathcal{Z},\mathcal{H};\boldsymbol{\theta})\right\} + \text{const} \quad (11)$$

where $q(\mathcal{H}/\mathcal{H}_\ell)$ is the approximation of the joint posterior distribution of all hidden variables but the subset $\mathcal{H}_\ell$. Subsequently, $q(\mathcal{H})$ can be inferred for each $\mathcal{H}_\ell \subseteq \mathcal{H}$. Once the posterior distributions of all variables in $\mathcal{H}$ are obtained, the log-likelihood of the complete data, $\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{q(\mathcal{H})}\left\{\ln p(\mathcal{Z},\mathcal{H};\boldsymbol{\theta})\right\}$, can be maximized w.r.t. the hyper-parameters in the M-step.

A detailed derivation of the algorithm is omitted due to space constraints and we only provide the final results and their interpretations. In the following, the time index t and the frequency index k are omitted for brevity whenever no ambiguity arises.

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

The approximate posterior distribution of the indicator can be computed from (9) and (11) by only keeping terms involving $\mathbf{x}$:

$$q(\mathbf{x}) \propto \mathbb{E}_{q(\boldsymbol{\psi})}\{\ln p(\mathbf{x}|\boldsymbol{\psi})\} + \mathbb{E}_{q(\mathbf{s}),q(\boldsymbol{\beta})}\{\ln p(\mathcal{Z}|\mathbf{x},\mathbf{s},\boldsymbol{\beta})\}. \quad (12)$$

Following a few mathematical transitions, the above distribution can be expressed as:

$$q(\mathbf{x}) = \prod_{m,t,k=1}^{M,T,K} r_{mtk}^{x(m,t,k)}. \quad (13)$$

It is evident that the posterior factor $q(\mathbf{x})$ takes the same functional form as the prior $p(\mathbf{x}|\boldsymbol{\psi})$, where

$$r_{mtk} = \frac{\rho_{mtk}}{\sum_{m'=1}^{M}\rho_{m'tk}}$$

and $\ln\rho_{mtk}$ is given by:

$$\ln\rho_{mtk} = \widehat{\ln\psi_m} - \hat{\beta}(m) + \left|\Gamma^{-1}\right| - N\ln(\pi) - \cdots$$
$$\left[\mathbf{z}\hat{\beta}(m)\Gamma^{-1}\mathbf{z} - \mathbf{z}^H\hat{\beta}(m)\Gamma^{-1}(k)\mathbf{g}_m\hat{s}(m) - \cdots\right.$$
$$\left.\hat{s}^H\mathbf{g}_m^H\hat{\beta}(m)\Gamma^{-1}\mathbf{z} + \widehat{|s(m)|^2}\mathbf{g}_m^H\hat{\beta}(m)\Gamma^{-1}\mathbf{g}_m\right].$$

The variables $\hat{\beta}$, $\hat{s}$, $\widehat{|s|^2}$ and $\widehat{\ln\psi}$ are posterior expectations that will be presented in the following sections. Note that since ρ_{mtk} is given by the exponential of a real variable, r_{mtk} are always non-negative. Moreover, it can be easily verified that they sum to one, as required. Using the above distribution, the expected value of the indicator at the candidate angle m is $\mathbb{E}\{x(m,t,k)\} = r_{mtk}$, and hence r_{mtk} can be regarded as an estimated (soft) indicator.

B. $\mathbf{E}$-$\boldsymbol{\psi}$ *Step*

The posterior expectation of the mixing coefficients $\boldsymbol{\psi}$ can be determined in a similar way:

$$\ln q(\boldsymbol{\psi}) = \mathbb{E}_{q(\mathbf{x})}\{\ln p(\mathbf{x}|\boldsymbol{\psi})\} + \mathbb{E}_{q(\mathbf{x})}\{\ln p(\boldsymbol{\psi})\}. \quad (14)$$

It can be shown that the posterior $q(\boldsymbol{\psi})$ is Dirichlet distributed with a parameter α_m^{post} (associated with candidate angle m) given by:

$$\alpha_m^{\text{post}} = \alpha_m + \mathcal{R}_m \quad (15)$$

where $\mathcal{R}_m = \sum_{t,k=1}^{T,K} r_{mtk}$ is the accumulated indicator. Since r_{mtk} can be interpreted as a soft indicator, $\mathcal{R}_m$ will tend to be high if a speaker is located at angle m. In the absence of informative prior on the speakers' DOAs, the hyper-parameter α_m equally contributes to all candidate angles, hence the parameter of the posterior $q(\boldsymbol{\psi})$ will be determined by $\mathcal{R}_m$. Using this posterior, $\hat{\psi}_m$ is given by $\mathbb{E}\{\psi_m\} = \frac{\alpha_m^{\text{post}}}{\overline{\boldsymbol{\alpha}}^{\text{post}}}$ with $\overline{\boldsymbol{\alpha}}^{\text{post}} = \sum_{m=1}^{M}\alpha_m^{\text{post}}$. Note that $\hat{\psi}_m$ is a map of the MoG weights which will ultimately used to extract the number of speakers in the scene and their location. Finally, $\widehat{\ln\psi_m}$, used in (14), is given by $\mathbb{E}\{\ln\psi_m\} = \Psi(\alpha_m^{\text{post}}) - \Psi(\overline{\boldsymbol{\alpha}}^{\text{post}})$ with Ψ the diaggma function.

C. $\mathbf{E}$-$\boldsymbol{s}$ *Step*

The speech posterior is also calculated with the VB procedure, using (9) and (11):

$$\ln q(\boldsymbol{s}) = \mathbb{E}_{q(\mathbf{x})q(\boldsymbol{\beta})}\{\ln p(\mathcal{Z}|\mathbf{x},\mathbf{s},\boldsymbol{\beta})\} + \mathbb{E}_{q(\boldsymbol{\tau})}\{\ln p(\mathbf{s}|\boldsymbol{\tau})\} \quad (16)$$

which yields a product of Gaussian distributions $q(\mathbf{s}) = \prod_{m,t,k=1}^{M,T,K} \mathcal{N}_c(s(m); \hat{s}(m), \Sigma_s(m))$, with the candidate-dependent means and covariance matrices:

$$\hat{s}(m) = \frac{\mathbf{z}^H \hat{\beta}(m) \Gamma^{-1} \mathbf{g}_m r_{mtk}}{\hat{\tau}(m) + \mathbf{g}_m^H \hat{\beta} \Gamma^{-1} \mathbf{g}_m r_{mtk}} \tag{17}$$

$$\Sigma_s(m) = \left(\hat{\tau}(m) + \mathbf{g}_m^H \hat{\beta}(m) \Gamma^{-1} \mathbf{g}_m r_{mtk}\right)^{-1}. \tag{18}$$

The mean of the mth Gaussian $\hat{s}(m)$ corresponds to the multichannel Wiener filter (MCWF) [13], multiplied by the estimated indicator. If the measurement is strongly associated with the mth candidate angle, then $r_{mtk} \to 1$. If the association is weak, namely $r_{mtk} \to 0$, the minimum mean square error (MMSE) of the signal tends towards 0. Interestingly, the variance in this case tends to $\hat{\tau}^{-1}(m)$, which is the largest possible error variance in optimal filtering, rendering the contribution of this Gaussian to the product negligible.

D. $\mathbf{E}$-$\boldsymbol{\tau}$ *Step*

Using (9) and (11), the posterior distribution of the speech precision can be written as:

$$\ln q(\boldsymbol{\tau}) = \ln \mathbb{E}_{q(\mathbf{s})}\{p(\mathbf{s}|\boldsymbol{\tau})\} + \ln p(\boldsymbol{\tau}) \tag{19}$$

which can be shown to be a Gamma distribution (similar in form to the prior distribution $p(\boldsymbol{\tau})$):

$$\prod_{m,t,k=1}^{M,T,K} \mathrm{Gam}(\tau(m); b_p(m), c_p(m))$$

with

$$b_p(m) = b_0(m) + 1, \quad c_p(m) = c_0(m) + \widehat{|s(m)|}^2 \tag{20}$$

where, $\widehat{|s(m)|}^2 = \Sigma_s(m) + |\hat{s}(m)|^2$. An estimate of the speech precision for candidate angle m can be easily obtained from the posterior:

$$\hat{\tau}(m) = \frac{b_p}{c_p} = \frac{b_0(m) + 1}{c_0(m) + \widehat{|s(m)|}^2}. \tag{21}$$

where $b_0(m), c_0(m)$ are the prior hyper-parameters which will be updated in the M-step. Modelling the speech precision as a deterministic variable, a point estimate is obtained: $\hat{\tau}_{\mathrm{D}}(m) = 1/\widehat{|s(m)|}^2$ [13]. It is useful to clarify that both estimators $\hat{\tau}(m)$ and $\hat{\tau}_{\mathrm{D}}(m)$ are equivalent when the hyper-parameters tend to zero, namely $b_0(m) = c_0(m) \to 0$, which corresponds to *non-informative* prior.

E. $\mathbf{E}$-$\boldsymbol{\beta}$ *Step*

The posterior distribution of the reverberation precision can be extracted from:

$$\ln q(\boldsymbol{\beta}) = \ln \mathbb{E}_{q(\mathbf{s})q(\mathbf{x})}\{p(\mathcal{Z}|\mathbf{x}, \mathbf{s}, \boldsymbol{\beta}\} + \ln p(\boldsymbol{\beta}) \tag{22}$$

and is given by $\prod_{m,t,k=1}^{M,T,K} \mathrm{Gam}(\beta(m); d_p(m), e_p(m))$ with parameters $d_p(m) = d_0(m) + r_{mtk}$ and

$$e_p(m) = e_0(m) + r_{mtk}\big(\mathbf{z}^H \Gamma^{-1} \mathbf{z} \\ - \mathbf{z}^H \Gamma^{-1} \mathbf{g}_m \hat{s} - \hat{s}^H \mathbf{g}_m^H \Gamma^{-1} \mathbf{z} + \widehat{|s(m)|}^2 \mathbf{g}_m^H \Gamma^{-1} \mathbf{g}_m\big). \tag{23}$$

An estimate of the reverberation precision is then obtained from the posterior mean:

$$\hat{\beta}(m) = \frac{d_p}{e_p} = \frac{d_0(m) + r_{mtk}}{e_p(m)}. \tag{24}$$

Modelling the reverberation power as a deterministic parameter, the following point estimator is obtained:

$$\hat{\beta}_{\mathrm{D}}^{-1}(m) = \mathbf{z}^H \Gamma^{-1} \mathbf{z} \\ - \mathbf{z}^H \Gamma^{-1} \mathbf{g}_m \hat{s} - \hat{s}^H \mathbf{g}_m^H \Gamma^{-1} \mathbf{z} + \widehat{|s(m)|}^2 \mathbf{g}_m^H \Gamma^{-1} \mathbf{g}_m. \tag{25}$$

Note that since $e_p(m) = e_0(m) + r_{mtk}\hat{\beta}_{\mathrm{D}}^{-1}(m)$, the following relation between the Bayesian and deterministic estimates is obtained:

$$\hat{\beta}(m) = \frac{d_0(m) + r_{mtk}}{e_0(m) + r_{mtk}\hat{\beta}_{\mathrm{D}}^{-1}(m)}.$$

It is easy to verify that in the limit $e_0(m), d_0(m) \to 0$, both estimators identify, namely $\hat{\beta}(m) = \hat{\beta}_D(m)$.

F. $\mathbf{M}$-*Step*

Once the posterior distributions of the hidden variables are calculated, the expected complete-data log likelihood $\mathcal{L}(\theta) = \mathbb{E}_{q(\mathcal{H})}\big\{\ln p(\mathcal{Z}, \mathcal{H}; \theta)\big\}$ can maximized w.r.t. the prior parameters. Closed-form expression are available only for the hyper-parameters of the Gamma priors, i.e. the speech and noise precisions:

$$\begin{aligned} e_0(m) &= e_p(m)\frac{d_0(m)}{d_p(m)} \\ d_0(m) &= \Psi^{-1}\Big(\Psi(d_p(m) + \ln\frac{e_0(m)}{e_p(m)}\Big) \\ c_0(m) &= c_p(m)\frac{b_0(m)}{b_p(m)} \\ b_0(m) &= \Psi^{-1}\Big(\Psi(b_p(m) + \ln\frac{c_0(m)}{c_p(m)}\Big). \end{aligned} \tag{26}$$

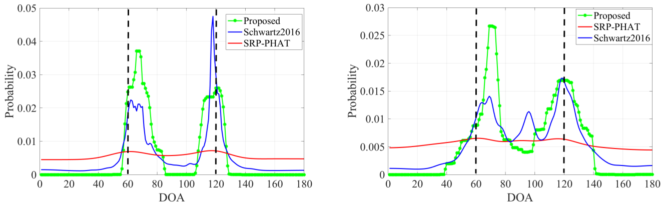
No closed-form expression for the hyper-parameters of the Dirichlet prior $\boldsymbol{\alpha}$ is available, hence we only consider in this work scenarios with non-informative prior on speakers' locations, thus choosing $\alpha_m = \frac{1}{M}$ for all candidate angles.

IV. Performance analysis

The performance of the proposed algorithm is evaluated and compared with baseline methods for the case of two concurrent speakers.

A. Simulation setup

Anechoic speech signals were convolved with room impulse responses (RIRs) downloaded from an open-source database recorded in our lab. Details about this database can be found in [14]. We selected here two reverberation levels $T_{60} = \{0.16, 0.61\}$ Sec. We further selected the loudspeakers positioned at various angles on a half-circle with a radius of 2 m. Only a subset of four microphones with inter-distances $\{3, 8, 3\}$ cm was used in our experiments. The parameters of the proposed algorithm are: sampling frequency 16 kHz,



(a) $T_{60} = 160$ ms

(b) $T_{60} = 610$ ms

Fig. 1: Probabilities vs. DOA for two speakers in two reverberation levels.

STFT frame-length 64 ms without overlap, and the number of frequency bins 1024. Two utterances, approximately 4.5 Sec long, of both male and female speakers were used. While applying the algorithm, only the frequencies below 4 kHz were used. In addition, we used only time-frequency bins for which $\mathrm{P}\{\mathbf{z}(t,k)\} > 10^{-5} \times \max \mathrm{P}\{\mathbf{z}(t,k)\}$, where $\mathrm{P}\{\mathbf{z}(t,k)\} = \frac{1}{N}\mathbf{z}^H(t,k)\mathbf{z}(t,k)$ denotes a spatially-averaged power of the (t,k)th bin. The number of iterations for each TF bin was set to 20. The angle candidates are confined to the range of $0° - 179°$ with $1°$ resolution. The diagonal loading factor ϵ was set to 0.5.

The performance of the proposed algorithm was compared with two competing algorithms, the SRP-PHAT [2], and ‘Schwartz2016 [7]. The outputs of the SRP-PHAT were normalized to sum to 1, to allow a clear comparison with the probability curves of the proposed algorithm and of ‘Schwartz2016.

B. DOA estimation performance

In the first set of experiments, we compared the performance of the three algorithms in reverberant environment with $T_{60} =$ 610ms, by averaging all possible two speakers' combinations in the range $15° - 165°$. Since the resolution of the database is $15°$ we have only 11 angles (altogether $11 \times 10 = 110$ different two speakers' combinations). The two DOAs with the highest probabilities were selected as the estimated DOA. The minimum absolute error (MAE) was calculated as the average of all errors between the estimated DOA and the true angle. In Table I, the MAEs for the various algorithms is presented.

Speaker\Alg.	SRP-PHAT	Schwartz2016	Proposed
Female	16.58°	6.81°	5.65°
Male	19.03°	7.46°	6.21°

TABLE I: MAE obtained the three competing algorithms. Results averaged over 110 possible pairs.

We further examine a specific scenario with two concurrent speakers, male and female, located at angles 60° and 120°. This test was repeated for two reverberation levels $T_{60} =$ 160 ms and $T_{60} = 610$ ms. The results are depicted in Fig. 1, where the weights of the MoG associated with each DOA is depicted. It is clear that the SRP-PHAT fails to localize the sources even in the low reverberation conditions. The ‘Schwartz2016’ algorithm exhibits good localization results in low reverberation conditions. However, in the high reverberation conditions, it fails to distinguish between the speakers. The proposed algorithm exhibits more robust behaviour and can clearly distinguish between the speakers in both reverberation conditions.

V. Conclusion

A DOA estimator in reverberant environment was presented. The proposed algorithm uses the VEM framework to cluster TF bins under a MoG model. It is distinguished from a previously proposed algorithm [7] by the use of a prior distribution on the candidate DOAs, as well as the speech and noise precisions, which are nuisance parameters of the problem at hand. An experimental study, using real acoustic impulse responses, demonstrates the improvement obtained by the proposed algorithm in comparison with two baseline methods [2], [7].

References

[1] C. Knapp and G. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[2] J.H. DiBiase, H.F. Silverman, and M.S. Brandstein, *Microphone arrays: signal processing techniques and applications*, chapter Robust Localization in Reverberant Rooms, pp. 157–180, Springer Verlag, 2001.

[3] M.I. Mandel, D.P.W. Ellis, and T. Jebara, “An EM algorithm for localizing multiple sound sources in reverberant environments,” *Advances in Neural Information Processing Systems*, vol. 19, pp. 953, 2007.

[4] M.I. Mandel, D.P.W. Ellis, and T. Jebara, “Model- based expectation-maximization source separation and localization,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[5] O. Yilmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[6] O. Schwartz and S. Gannot, “Speaker tracking using recursive EM algorithms,” *Transactions on Audio, Speech, and Language Processing, IEEE/ACM*, vol. 22, no. 2, pp. 392–402, 2014.

[7] O. Schwartz, Y. Dorfan, E.A.P. Habets, and S. Gannot, “Multiple DOA estimation in reverberant conditions using EM,” in *International Workshop on Acoustic Echo Cancellation and Noise Control (IWAENC), Xi'an, China*, 2016.

[8] H. Ye and R. D. DeGroat, “Maximum likelihood DOA estimation and asymptotic Cramér-Rao bounds for additive unknown colored noise,” *IEEE Transactions on Signal Processing*, vol. 43, no. 4, pp. 938–949, 1995.

[9] O. Schwartz, Y. Dorfan, M. Taseska, E.A.P. Habets, and S. Gannot, “DOA estimation in noisy environment with unknown noise power using the EM algorithm,” in *Hands-free Speech Communication and Microphone Arrays (HSCMA)*. IEEE, 2017, pp. 86–90.

[10] Z. Zohny, S.M. Naqvi, and J.A. Chambers, “Variational EM for clustering interaural phase cues in MESSL for blind source separation of speech,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2015, vol. 1, pp. I–41.

[11] A Kuklasinski, S. Doclo, S. H. Jensen, and J. Jensen, “Maximum likelihood based multi-channel isotropic reverberation reduction for hearing aids,” in *Proceedings of the 22nd European Signal Processing Conference (EUSIPCO)*, 2014, pp. 61–65.

[12] Christopher M. Bishop, *Pattern recognition and machine learning*, Information Science and Statistics. Springer-Verlag New York Inc., New York, NY, United States, 1st edition, Apr. 2006.

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

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