A BAYESIAN HIERARCHICAL MIXTURE OF GAUSSIAN MODEL FOR MULTI-SPEAKER DOA ESTIMATION AND SEPARATION

Yaron Laufer and Sharon Gannot

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

ABSTRACT

In this paper we propose a fully Bayesian hierarchical model for multi-speaker direction of arrival (DoA) estimation and separation in noisy environments, utilizing the W-disjoint orthogonality property of the speech sources. Our probabilistic approach employs a mixture of Gaussians formulation with centroids associated with a grid of candidate speakers' DoAs. The hierarchical Bayesian model is established by attributing priors to the various parameters. We then derive a variational Expectation-Maximization algorithm that estimates the DoAs by selecting the most probable candidates, and separates the speakers using a variant of the multichannel Wiener filter that takes into account the responsibility of each candidate in describing the received data. The proposed algorithm is evaluated using real room impulse responses from a freely-available database, in terms of both DoA estimates accuracy and separation scores. It is shown that the proposed method outperforms competing methods.

***Index Terms*—** Audio source separation, DoA estimation, variational EM, Mixture of Gaussians, W-disjoint orthogonality.

1. INTRODUCTION

Sound source separation and direction of arrival (DoA) estimation of multiple speakers from a mixture signal are fundamental problems in the field of audio signal processing. A common challenge arises in the presence of noise, which may degrade both the accuracy of the DoA estimates as well as the quality of the separated speakers.

Common DoA estimation methods are multiple signal classification (MUSIC) [1] and steered-response power phase transform (SRP-PHAT) [2]. However, these techniques are prone to performance degradation when multiple speakers co-exist in the same environment. A widely-used approach for estimating the DoAs of concurrent speakers relies on the W-disjoint orthogonality (WDO) assumption [3], i.e. the sparsity of speech sources in the time-frequency (TF) domain. Typically, a TF mask is constructed using the Expectation-Maximization (EM) algorithm, associating each TF bin to a single candidate angle, and then DoA estimates can be selected as the most probable candidates [4].

In [5], a model-based EM source separation and localization (MESSL) algorithm is proposed for a stereo mixture. Based on interaural phase difference (IPD) and interaural level difference (ILD) features, each TF bin is associated with both a candidate speaker and a candidate DoA. The resulting TF masks are used to separate the individual sound sources. Note that additive noise is not explicitly modeled with these features, thus motivating to directly cluster the raw observations [6–11]. In [6–8], the association procedure consists of two steps. First, each TF bin is associated with a candidate speaker, and then the speaker is associated with a candidate angle. An extension to the case of moving sound sources is presented in [9]. In [10, 11], the association procedure is simplified to include only DoA candidates. An explicit modelling of the noise power spectral density (PSD) is included, where the noise PSD matrix is either assumed to be known [10], or included in the estimation procedure [11]. As opposed to [6–9] which adopt a variational Bayesian approach for the inference of the parameters, in [10, 11] the various model parameters are treated as deterministic unknown parameters, for which point estimates are computed. Note that the separation methods in [5–7] apply the TF masks directly to the multichannel mixture, whereas in [8–10] the TF mask is applied to the output of a beamformer, which aims at reducing the ambient noise.

In the Bayesian framework, model parameters are viewed as random variables having a prior probability density function (PDF), rather than deterministic unknown parameters. This approach allows us to include prior knowledge and to explore uncertainty in the model. As the inference is carried out based on the entire PDF rather than point estimates, the obtained estimators are more robust and less sensitive to local maxima [12]. Hierarchical Bayesian models are very useful, since they use a multi-level modeling to capture important dependencies among parameters. However, the posterior distribution might be intractable in complex Bayesian models, and thus the EM algorithm cannot be applied. The variational approach [12–14] circumvents this difficulty by approximating the posterior. Recently, several works utilized the variational EM (VEM) method for speech enhancement [15–17], speech dereverberation [18], and audio source separation [6–9, 19].

In this work, we extend the probabilistic model proposed in [10,11] towards a fully Bayesian model. We introduce a hierarchical mixture of Gaussians (MoG) model in a Bayesian setting, by placing Gaussian priors over the speech signals and Gamma hyperpriors over the speech precisions. The noise is modelled as a spatially homogeneous sound field, with a known spatial coherence matrix and an unknown precision. A Gamma prior is attributed to the noise precision. A latent activity indicator is introduced for the assignment of TF bins to candidate DoAs, with its mixture weights having a Dirichlet distribution. The posterior distributions of the various latent variables are inferred through a VEM algorithm, leading to a probability map over the candidate angles, from which we select the most probable candidates as the DoA estimates. The separated speakers are then obtained as the posterior speech estimates associated with the dominant candidates.

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

978-1-7281-6662-9/20/$31.00 ©2020 IEEE

2. PROBLEM FORMULATION

2.1. Signal Model

Consider a mixture of J speakers received by N microphones, in a noisy acoustic environment. We work in the short-time Fourier transform (STFT) domain, where $k \in [1, K]$ denotes the frequency band, and $\ell \in [1, L]$ denotes the time frame. The N-channel measurement signal $\mathbf{z}(\ell, k) = [z_1(\ell, k), \cdots, z_N(\ell, k)]^\top$ writes

$$\mathbf{z}(\ell, k) = \sum_{j=1}^{J} s_j(\ell, k)\mathbf{g}_j(k) + \mathbf{u}(\ell, k), \tag{1}$$

where $s_j(\ell, k)$ is the jth speech signal as received by the first microphone (that was arbitrarily chosen as the reference microphone), $\mathbf{g}_j(k) = [1, g_{j,2}(k), \cdots, g_{j,N}(k)]^\top$ is the relative direct-path transfer function (RDTF) vector associated with the jth speaker and $\mathbf{u}(\ell, k) = [u_1(\ell, k), \cdots, u_N(\ell, k)]^\top$ denotes the additive noise. The nth element of $\mathbf{g}_j$ is given by $g_{j,n}(k) = \exp\left(-j\frac{2\pi k}{K} f_s \tau_{j,n}\right)$, where f_s denotes the sampling frequency and $\tau_{j,n}$ is the time difference of arrival (TDOA) of the jth speaker between microphone n and the reference microphone, i.e. $\tau_{j,n} = \frac{d_n \cos(\theta_j)}{c}$, where θ_j is the DoA of the jth speaker, d_n is the distance between the nth microphone and the reference microphone and c is the sound velocity.

2.2. Probabilistic Model

We model the observations using a MoG model with M components. The centroid of each Gaussian is associated with an optional source DoA from a predefined grid of candidate angles $\{\theta_1, \cdots, \theta_M\}$, as in [10, 11]. By utilizing the disjoint activity of the speakers in the STFT domain [3], each TF bin can be associated with a single active source that impinges the array from a particular angle in the grid.

Let $\mathbf{g}_m$ and s_m denote, respectively, the RDTF and the speech signal associated with a candidate speaker located at the mth angle (as opposed to $\mathbf{g}_j$ and s_j in (1) which are defined per actual source). The candidate speech signals are modelled as independent zero-mean Gaussian random variables, $p\big(s_m(\ell, k)|\tau_m(\ell, k)\big) = \mathcal{N}_c\big(s_m(\ell, k); 0, \tau_m^{-1}(\ell, k)\big)$, where τ_m denotes the precision of the mth candidate. For brevity, we denote $\mathbf{s}(\ell, k) = [s_1(\ell, k), \cdots, s_M(\ell, k)]^\top$. The noise is modelled as a stationary, zero-mean multivariate Gaussian process with $p\big(\mathbf{u}(\ell, k)|\mathbf{\Phi}_\mathbf{u}(k)\big) = \mathcal{N}_c\big(\mathbf{u}(\ell, k); \mathbf{0}, \mathbf{\Phi}_\mathbf{u}(k)\big)$. The noise is assumed to be spatially homogeneous, i.e. $\mathbf{\Phi}_\mathbf{u}(k) = \beta^{-1}(k)\mathbf{\Gamma}(k)$, where $\beta(k)$ is the inverse noise power and $\mathbf{\Gamma}(k)$ is a spatial coherence matrix, assumed to be known.

The mixture distribution is formulated in terms of discrete latent variables, indicating the assignment of each TF bin to a specific source angle. To this end, we introduce an M-dimensional binary random variable for each TF bin $\mathbf{x}(\ell, k) = [x_1(\ell, k), \cdots, x_M(\ell, k)]^\top$, in which a particular element $x_m(\ell, k)$ equals 1 (indicating the active source DoA) while the rest are zeros. Thus, the conditional data distribution writes

$$p\big(\mathbf{z}(\ell, k)|\mathbf{s}(\ell, k), \mathbf{x}(\ell, k), \beta(k)\big) = \prod_{m=1}^{M} \mathcal{N}_c\big(\mathbf{z}(\ell, k); s_m(\ell, k)\mathbf{g}_m(k), \beta^{-1}(k)\mathbf{\Gamma}(k)\big)^{x_m(\ell, k)}. \tag{2}$$

The PDF of the latent variable $\mathbf{x}(\ell, k)$ is a categorical distribution, i.e. $p\left(\mathbf{x}(\ell, k)|\boldsymbol{\psi}\right) = \prod_{m=1}^{M} \psi_m^{x_m(\ell, k)}$, where $\boldsymbol{\psi} = [\psi_1, \cdots, \psi_M]^\top$ is the vector of mixture coefficients, defined by $\psi_m \triangleq p\big(x_m(\ell, k) = 1\big)$, and $\sum_{m=1}^{M} \psi_m = 1$. It can be shown that by multiplying (2) with $p\left(\mathbf{x}(\ell, k)|\boldsymbol{\psi}\right)$ and then marginalizing over $\mathbf{x}(\ell, k)$, a standard MoG formulation [14] is obtained:

$$p\big(\mathbf{z}(\ell, k)|\mathbf{s}(\ell, k), \beta(k), \boldsymbol{\psi}\big) = \sum_{m=1}^{M} \psi_m \mathcal{N}_c\big(\mathbf{z}(\ell, k); s_m(\ell, k)\mathbf{g}_m(k), \beta^{-1}(k)\mathbf{\Gamma}(k)\big). \tag{3}$$

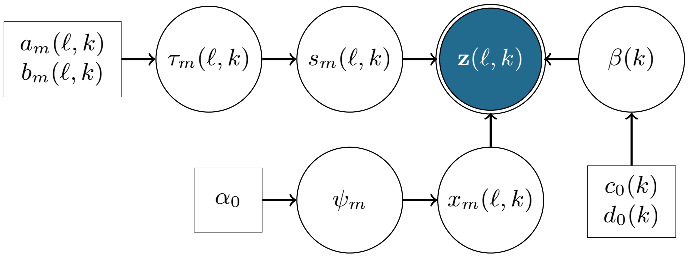

Fig. 1: Graphical model.

2.3. Conjugate Priors

Our hierarchical model is established by introducing priors over all model parameters, namely: i) The precisions of the speech candidates; ii) The precision of the noise; and iii) The mixture weights. The conjugate prior for the precision of a univariate Gaussian is the Gamma distribution [14]. Hence, the prior for the speech precision of the mth candidate writes:

$$p\big(\tau_m(\ell, k)\big) = \text{Gam}\big(\tau_m(\ell, k); a_m(\ell, k), b_m(\ell, k)\big). \tag{4}$$

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

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

For the weights of a categorical distribution, the conjugate prior is the Dirichlet distribution:

$$p\big(\boldsymbol{\psi}\big) = \text{Dir}\big(\boldsymbol{\psi}; \alpha_0\big). \tag{6}$$

The proposed hierarchical model is illustrated in Fig. 1. The definitions of the various probability distributions appear in the Appendix.

3. VEM ALGORITHM

In this work, $\mathcal{Z} = \{\mathbf{z}(\ell, k)\}_{\ell,k=1}^{L,K}$ denotes the set of observations, the set of hidden variables consists of $\mathcal{H} = \big\{s_m(\ell, k), \tau_m(\ell, k), \beta(k), x_m(\ell, k), \psi_m\big\}_{\ell,k,m=1}^{L,K,M}$, and the parameter set consists of $\Theta = \big\{a_m(\ell, k), b_m(\ell, k), c_0(k), d_0(k), \alpha_0\big\}_{\ell,k,m=1}^{L,K,M}$. Bayesian inference of latent variables requires the computation of the posterior distribution $p(\mathcal{H}|\mathcal{Z}; \Theta) = \frac{p(\mathcal{Z}, \mathcal{H}; \Theta)}{p(\mathcal{Z}; \Theta)}$. Based on the probabilistic assumptions of Section 2, the complete-data distribution writes

$$p(\mathcal{Z},\mathcal{H};\Theta) = \prod_{\ell,k,m=1}^{L,K,M} \Big[p\big(\mathbf{z}(\ell,k)|s_m(\ell,k),\mathbf{g}_m(\ell,k),\beta(k)\big)^{x_m(\ell,k)}$$
$$\times\, p\big(s_m(\ell,k)|\tau_m(\ell,k)\big)p\big(\tau_m(\ell,k);a_m(\ell,k),b_m(\ell,k)\big)\Big]$$
$$\times \prod_{\ell,k=1}^{L,K}\Big[p\big(\mathbf{x}(\ell,k)|\boldsymbol{\psi}\big)\Big]p(\boldsymbol{\psi};\alpha_0)\prod_{k=1}^{K}p\big(\beta(k);c_0(k),d_0(k)\big). \quad (7)$$

Due to the complex form of (7), $p(\mathcal{Z};\Theta) = \int p(\mathcal{Z},\mathcal{H};\Theta)d\mathcal{H}$ cannot be computed analytically and thus exact inference becomes intractable. We therefore propose a variational inference procedure, which approximates the posterior, $q(\mathcal{H}) \approx p(\mathcal{H}|\mathcal{Z};\Theta)$. According to the *mean field theory* [13, 20], we assume that the speech signals, speech precisions, noise precision, activity indicators and mixture weights are conditionally independent given the observations. Therefore, the approximate posterior distribution factorizes as:

$$q(\mathcal{H}) = \prod_{\ell,k,m=1}^{L,K,M}\Big[q\big(s_m(\ell,k)\big)q\big(\tau_m(\ell,k)\big)\Big]$$
$$\times\prod_{\ell,k=1}^{L,K}\Big[q\big(\mathbf{x}(\ell,k)\big)\Big]q(\boldsymbol{\psi})\prod_{k=1}^{K}q\big(\beta(k)\big). \quad (8)$$

The VEM procedure consists in iterating the following two steps until convergence. In the E-Step, the approximate posterior distribution of each subset $\mathcal{H}_i \subseteq \mathcal{H}$ is computed by [14]:

$$\ln q(\mathcal{H}_i) = \mathbb{E}_{q(\mathcal{H}/\mathcal{H}_i)}[\ln p(\mathcal{Z},\mathcal{H};\Theta)] + \text{const}, \quad (9)$$

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

3.1. E-s Step

The approximate posterior PDF of the speech signal emitted from angle θ_m is obtained from (7) and (9) by keeping only the terms that depend on $s_m(\ell)$:

$$\ln q(s_m(\ell)) \propto \mathbb{E}_{q(\tau_m(\ell))q(x_m(\ell))q(\beta)}\Big[\ln p\big(\mathbf{z}(\ell)|s_m(\ell),x_m(\ell),\beta\big)$$
$$+\ln p\big(s_m(\ell)|\tau_m(\ell)\big)\Big]. \quad (10)$$

It can be shown that (10) yields a Gaussian distribution $q\left(s_m(\ell)\right) = \mathcal{N}_c\left(s_m(\ell);\hat{s}_m(\ell),\Sigma_{s,m}(\ell)\right)$, with

$$\hat{s}_m(\ell) = \frac{r_m(\ell)\mathbf{g}_m^{\mathrm{H}}\hat{\beta}\mathbf{\Gamma}^{-1}\mathbf{z}(\ell)}{r_m(\ell)\mathbf{g}_m^{\mathrm{H}}\hat{\beta}\mathbf{\Gamma}^{-1}\mathbf{g}_m+\hat{\tau}_m(\ell)}, \quad (11)$$

$$\Sigma_{s,m}(\ell) = \Big(r_m(\ell)\mathbf{g}_m^{\mathrm{H}}\hat{\beta}\mathbf{\Gamma}^{-1}\mathbf{g}_m+\hat{\tau}_m(\ell)\Big)^{-1}, \quad (12)$$

where $\hat{\tau}_m(\ell)$, $r_m(\ell)$ and $\hat{\beta}$ are posterior statistics that will be defined in the following sections. The speech signal impinging the array from θ_m can thus be estimated by the posterior mean (PM), namely $\hat{s}_m(\ell)$. This speech estimator resembles the form of the multichannel Wiener filter (MCWF) [21]:

$$\hat{s}_{\text{MCWF},m}(\ell) = \frac{\mathbf{g}_m^{\mathrm{H}}\hat{\beta}\mathbf{\Gamma}^{-1}\mathbf{z}(\ell)}{\mathbf{g}_m^{\mathrm{H}}\hat{\beta}\mathbf{\Gamma}^{-1}\mathbf{g}_m+\hat{\tau}_m(\ell)}, \quad (13)$$

besides $r_m(\ell)$. It is well known that the MCWF can be decomposed into a multichannel minimum variance distortionless response (MVDR) beamformer followed by a single-channel Wiener postfilter [22]. In a similar way, $\hat{s}_m(\ell)$ can be decomposed as

$$\hat{s}_m(\ell) = \underbrace{\frac{r_m(\ell)\mathbf{g}_m^{\mathrm{H}}\hat{\beta}\mathbf{\Gamma}^{-1}\mathbf{g}_m}{r_m(\ell)\mathbf{g}_m^{\mathrm{H}}\hat{\beta}\mathbf{\Gamma}^{-1}\mathbf{g}_m+\hat{\tau}_m(\ell)}}_{H_m(\ell)} \times \underbrace{\frac{\mathbf{g}_m^{\mathrm{H}}\mathbf{\Gamma}^{-1}}{\mathbf{g}_m^{\mathrm{H}}\mathbf{\Gamma}^{-1}\mathbf{g}_m}}_{\mathbf{w}_{\text{MVDR},m}^{\mathrm{H}}}\mathbf{z}(\ell). \quad (14)$$

As will be shown in Sec. 3.3, $r_m(\ell)$ is a soft mask, representing the *responsibility* of a speaker from the mth angle for data point $\mathbf{z}(\ell)$. Due to the unknown activity pattern of a speaker impinging the array from angle θ_m, the MVDR steered towards θ_m might enhance a speech-absent TF bin. $H_m(\ell)$ is a postfilter that takes into account the uncertainty level in the activity of the mth speaker, expressed by $r_m(\ell)$, and weights accordingly the single channel at the MVDR output. When $r_m(\ell) \to 1$, $\hat{s}_m(\ell)$ reduces to $\hat{s}_{\text{MCWF},m}(\ell)$.

The speech estimator of (14) will be used later for extracting the individual speakers from the noisy mixture, by selecting the most probable candidates according to the posterior distribution of the mixture weights, $q(\boldsymbol{\psi})$. Note that in [10], the speech signal associated with the mth candidate was arbitrarily estimated as

$$\hat{s}_m^{[10]}(\ell) = r_m(\ell) \times \mathbf{w}_{\text{MVDR},m}^{\mathrm{H}}\mathbf{z}(\ell). \quad (15)$$

3.2. E-τ Step

The posterior PDF of the speech precision associated with a speaker located at the mth angle writes:

$$\ln q(\tau_m(\ell)) \propto \mathbb{E}_{q(s_m(\ell))}\Big[\ln p\big(s_m(\ell)|\tau_m(\ell)\big)\Big]$$
$$+\ln p\big(\tau_m(\ell);a_m(\ell),b_m(\ell)\big), \quad (16)$$

which can be shown to be a Gamma distribution: $q(\tau_m(\ell)) = \text{Gam}(\tau_m(\ell);a_{p,m}(\ell),b_{p,m}(\ell))$, with

$$a_{p,m}(\ell) = a_m(\ell)+1\ ,\ b_{p,m}(\ell) = b_m(\ell)+\widehat{|s_m(\ell)|^2}. \quad (17)$$

As a result, the PM estimate of τ_m is given by:

$$\hat{\tau}_m(\ell) = \frac{a_{p,m}(\ell)}{b_{p,m}(\ell)} = \frac{a_m(\ell)+1}{b_m(\ell)+\widehat{|s_m(\ell)|^2}}. \quad (18)$$

3.3. E-x Step

The posterior PDF of $\mathbf{x}(\ell)$ is given by

$$\ln q(\mathbf{x}(\ell)) \propto \sum_{m=1}^{M}x_m(\ell)\mathbb{E}_{q(s_m(\ell))q(\beta)}\Big[\ln p\big(\mathbf{z}(\ell)|s_m(\ell),x_m(\ell),\beta\big)\Big]$$
$$+\mathbb{E}_{q(\boldsymbol{\psi})}\Big[\ln p\big(\mathbf{x}(\ell)|\boldsymbol{\psi}\big)\Big], \quad (19)$$

yielding a categorical distribution: $q(\mathbf{x}(\ell)) = \prod_{m=1}^{M} r_m(\ell)^{x_m(\ell)}$, with $r_m(\ell) \triangleq q\big(x_m(\ell) = 1\big) = \frac{\rho_m(\ell)}{\sum\limits_{m=1}^{M} \rho_m(\ell)}$, and

$$\begin{aligned} \rho_m(\ell) &= \exp\Big\{\mathbb{E}_{q(\boldsymbol{\psi})}\big[\ln \psi_m\big]\Big\} \\ &\times \exp\Big\{\mathbb{E}_{q(s_m(\ell))q(\beta)}\big[\ln \mathcal{N}_c\big(\mathbf{z}(\ell); s_m(\ell)\mathbf{g}_m, \beta^{-1}\boldsymbol{\Gamma}\big)\big]\Big\}. \qquad (20) \end{aligned}$$

The PM estimate of x_m is therefore $\hat{x}_m(\ell) = r_m(\ell)$, representing the responsibility of the mth candidate for data point $\mathbf{z}(\ell)$. The expression in (20) can be further simplified using (37b) in the Appendix. Note that the deterministic approach of [10, 11] yields a different expression:

$$\rho_m(\ell) = \psi_m \times \mathcal{N}_c\big(\mathbf{z}(\ell); \mathbf{0}, \tau_m^{-1}(\ell)\mathbf{g}_m\mathbf{g}_m^{\mathrm{H}} + \beta^{-1}\boldsymbol{\Gamma}\big). \qquad (21)$$

3.4. E-$\boldsymbol{\psi}$ Step

The posterior PDF of the mixture weights vector writes:

$$\ln q(\boldsymbol{\psi}) \propto \sum_{\ell,k=1}^{L,K} \mathbb{E}_{q(\mathbf{x}(\ell,k))}\Big[\ln p\big(\mathbf{x}(\ell,k)|\boldsymbol{\psi}\big)\Big] + \ln p\big(\boldsymbol{\psi}; \alpha_0\big), \qquad (22)$$

which leads to a Dirichlet distribution: $q\big(\boldsymbol{\psi}\big) = \mathrm{Dir}\big(\boldsymbol{\psi}; \boldsymbol{\alpha}_p\big)$, with $\alpha_{p,m} = \alpha_0 + \sum\limits_{\ell,k=1}^{L,K} r_m(\ell,k)$. Using (37a) in the Appendix, the PM estimate of the mixture weights is given by

$$\hat{\psi}_m = \frac{\alpha_0 + \sum\limits_{\ell,k=1}^{L,K} r_m(\ell,k)}{M\alpha_0 + LK}, \quad m = 1, \dots, M. \qquad (23)$$

We obtain a probability map over the candidate DoAs, $\hat{\boldsymbol{\psi}} = [\hat{\psi}_1, \cdots, \hat{\psi}_M]^\top$. The locations of the peaks in $\hat{\boldsymbol{\psi}}$, i.e. the most probable candidate angles, can be selected as the DoA estimates, and the corresponding posterior speech estimates $\hat{s}_m$ can be taken as the separated speakers. Note that if $\alpha_0 \to 0$, i.e. the Dirichlet prior is broad, then $\hat{\psi}_m \to \frac{\sum_{\ell,k=1}^{L,K} r_m(\ell,k)}{LK}$, thus coinciding with the deterministic estimator proposed in [10,11].

3.5. E-β Step

Similarly, the posterior PDF of the noise precision writes:

$$\begin{aligned} \ln q(\beta) \propto \sum_{\ell,m=1}^{L,M} &\mathbb{E}_{q(s_m(\ell))q(x_m(\ell))}\Big[\ln p\big(\mathbf{z}(\ell)|s_m(\ell), x_m(\ell), \beta\big)\Big] \\ &+ \ln p\big(\beta; c_0, d_0\big), \qquad (24) \end{aligned}$$

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

$$c_p = c_0 + NL, \qquad (25)$$

$$\begin{aligned} d_p = d_0 + \sum_{\ell,m=1}^{L,M} r_m(\ell)\Big(&\mathbf{z}^{\mathrm{H}}(\ell)\boldsymbol{\Gamma}^{-1}\mathbf{z}(\ell) \\ - 2\Re\Big\{\mathbf{z}^{\mathrm{H}}(\ell)\boldsymbol{\Gamma}^{-1}\mathbf{g}_m\hat{s}_m(\ell)\Big\} &+ |\widehat{s_m(\ell)}|^2\mathbf{g}_m^{\mathrm{H}}\boldsymbol{\Gamma}^{-1}\mathbf{g}_m\Big). \qquad (26) \end{aligned}$$

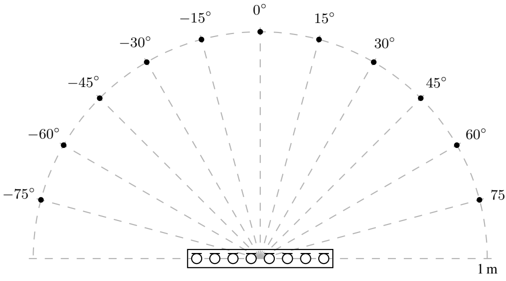

Fig. 2: Geometric setup.

Hence, the PM estimate for the noise precision is given by:

$$\hat{\beta} = \frac{c_p}{d_p}. \qquad (27)$$

3.6. M Step

The parameters are now estimated by maximizing the expected log-likelihood of the completed data $\mathcal{L}(\Theta) = \mathbb{E}_{q(\mathcal{H})}\left[\ln p(\mathcal{Z}, \mathcal{H}; \Theta)\right]$. Using standard properties of Gamma and Dirichlet distributions (see Appendix), we obtain the following updates:

$$\begin{aligned} a_m(\ell) &= \Psi^{-1}\left[\Psi(a_{p,m}(\ell)) + \ln\frac{b_m(\ell)}{b_{p,m}(\ell)}\right], \quad m = 1, \dots, M \\ b_m(\ell) &= \frac{a_m(\ell)}{a_{p,m}(\ell)} b_{p,m}(\ell), \quad m = 1, \dots, M \\ c_0 &= \Psi^{-1}\left[\Psi(c_p) + \ln\frac{d_0}{d_p}\right] \quad , \quad d_0 = \frac{c_0}{c_p}d_p, \qquad (28) \end{aligned}$$

where $\Psi(\cdot)$ is the digamma function. Differentiating $\mathcal{L}(\Theta)$ w.r.t. α_0, yields

$$\begin{aligned} \frac{\partial\mathcal{L}(\Theta)}{\partial\alpha_0} &= \frac{\partial}{\partial\alpha_0}\mathbb{E}_{q(\boldsymbol{\psi})}\Big[\ln p\big(\boldsymbol{\psi}; \alpha_0\big)\Big] \\ &= M\big(\Psi(M\alpha_0) - \Psi(\alpha_0)\big) + \sum_{m=1}^{M}\Big[\Psi(\alpha_{p,m}) - \Psi\Big(\sum_{n=1}^{M}\alpha_{p,n}\Big)\Big]. \qquad (29) \end{aligned}$$

Setting (29) to zero, a closed-form solution for α_0 is not available. Instead, we apply Newton's iterative method [23]:

$$\alpha_0^{(i+1)} = \alpha_0^{(i)} - \frac{d\left(\alpha_0^{(i)}\right)}{h\left(\alpha_0^{(i)}\right)}, \qquad (30)$$

where $d(\alpha_0)$ is the first-order derivative of $\mathcal{L}(\Theta)$ w.r.t. α_0, given by (29), and $h(\alpha_0)$ is the second-order derivative w.r.t. α_0:

$$h(\alpha_0) = M\big(M\Psi'(M\alpha_0) - \Psi'(\alpha_0)\big), \qquad (31)$$

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

4. PERFORMANCE EVALUATION

4.1. Simulation Setup

Room impulse responses (RIRs) were downloaded from an open-source database [24]. The database was recorded in a room with

Table 1: DoA Estimation Accuracy

	MAE [deg]		
Alg.\SNR	0dB	5dB	10dB
SRP-PHAT	17.81	13.64	13.81
EM-based	8.72	6.80	5.02
Proposed	**5.46**	**2.55**	**2.45**

dimensions $6 \times 6 \times 2.4$ m. We selected RIRs corresponding to a uniform linear array (ULA) of $N = 8$ microphones with inter-distances of 8 cm and reverberation level of $T_{60} = 0.16$ s. Our experiments consist of two concurrent speakers, located at 1 m distance from the array, at different angles in the set $\{-75^\circ, -60^\circ, \dots, 75^\circ\}$, as illustrated in Fig. 2. For the clean speech signals, we used utterances of five male and five female speakers from the TIMIT database [25]. In each experiment, utterances of 2 speakers (one male and one female) were randomly selected, and then convolved with the corresponding RIRs. An artificial diffuse noise with speech-like spectrum was generated by the method described in [26], with various signal to noise ratio (SNR) levels. The sampling rate was 16 kHz, and the STFT frame length was 64 ms with 75% overlap. The number of Gaussian candidates was set to $M = 180$, corresponding to an angular range of $[-89^\circ, 90^\circ]$ with resolution of 1°. We used the frequency band of 300–3400 Hz. For the Newton search, 5 iterations were applied. In the proposed method, the two candidate angles with the largest posterior probabilities are selected as the DoA estimates, and the corresponding posterior speech estimates (see (14)) are taken as the separated speakers. Note that when the number of sources is unknown, it can be determined by analyzing $\boldsymbol{\psi}$. However, an elaborated study of this issue is out of the scope of this paper.

4.2. Performance Measures and Competing Methods

The accuracy of the DoA estimates was assessed using the mean absolute error (MAE). The source separation performance is evaluated with two common objective measures, namely signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR) [27]. The speech quality and intelligibility is measured in terms of short-time objective intelligibility (STOI) [28]. The reported results are average measures over the 55 different possible combinations of two speakers, in the given set of angles.

The DoA estimation performance of the proposed method is compared to the following methods: (i) SRP-PHAT method [2]; and (ii) The EM-based DoA estimation method of [11]. The source separation performance is compared to the EM-based seperation procedure of [10], when used in the framework of [11], i.e. assuming that the noise power is unknown. For both the EM-based and the proposed methods, the number of iterations was fixed to 40.

4.3. Results

In Table 1, MAE results are presented for several SNR levels. SDR and SIR scores are summarized in Table 2, and STOI scores are presented in Table 3. The best results are highlighted in boldface. It is evident that the proposed method outperforms the competing methods in almost all cases, by providing more accurate DoA estimates and improved separation quality.

Table 2: Separation Scores: SDR (Left) and SIR (Right)

	SDR [dB]			SIR [dB]		
Alg.\SNR	0dB	5dB	10dB	0dB	5dB	10dB
Unprocessed	-3.09	-1.19	-0.34	0.12	0.12	0.13
EM-based	2.63	5.56	7.88	15.3	16.1	**17.1**
Proposed	**6.12**	**8.17**	**9.07**	**16.0**	**16.5**	16.5

Table 3: STOI Scores

	STOI [%]		
Alg.\SNR	0dB	5dB	10dB
Unprocessed	31.2	57.1	73.8
EM-based	67.3	83.3	92.6
Proposed	**78.2**	**92.2**	**95.1**

5. CONCLUSIONS

In this paper, we presented a Bayesian hierarchical model for multi-speaker DoA estimation and separation. The model employs a MoG formulation for the possible speakers' DoAs, utilizing the W-disjoint orthogonality (WDO) property of speech sources. A fully Bayesian approach is adopted, by placing Gamma priors over the precisions of the speakers and the noise, and a Dirichlet prior over the mixing weights of the MoG. The inference of the hidden variables is performed using a VEM algorithm. The discussion is supported by an experimental study in a room with a reverberation time of 0.16 sec and various SNR levels, demonstrating the advantage of the proposed method over competing methods.

6. APPENDIX

Standard Probability Distributions

The multivariate complex Gaussian distribution is given by:

$$\mathcal{N}_c(\mathbf{a}; \boldsymbol{\mu}_a, \boldsymbol{\Phi}_a) = \frac{1}{|\pi \boldsymbol{\Phi}_a|} \exp\left[-\left(\mathbf{a} - \boldsymbol{\mu}_a\right)^{\mathrm{H}} \boldsymbol{\Phi}_a^{-1} \left(\mathbf{a} - \boldsymbol{\mu}_a\right)\right], \tag{32}$$

where $\boldsymbol{\mu}_a$ is the mean vector and $\boldsymbol{\Phi}_a$ the covariance matrix.

A Gamma distribution for a non-negative random variable λ with shape and rate parameters $a, b > 0$ is given by [14]:

$$\mathrm{Gam}(\lambda; a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp\left(-b\lambda\right), \tag{33}$$

where $\Gamma(\cdot)$ is the gamma function. The Gamma distribution has the following properties:

$$\mathbb{E}\left[\lambda\right] = \frac{a}{b} \quad ; \quad \mathbb{E}\left[\ln \lambda\right] = \Psi(a) - \ln(b), \tag{34}$$

where $\Psi(a) \equiv \frac{d}{da} \ln \Gamma(a)$ is the digamma function.

A Dirichlet distribution for a random vector $\boldsymbol{\mu} =$

$[\mu_1, \cdots, \mu_M]^\top$ with $0 \leq \mu_m \leq 1$ and $\sum_{m=1}^{M} \mu_m = 1$, writes [14]:

$$\mathrm{Dir}(\boldsymbol{\mu}; \boldsymbol{\alpha}) = C(\boldsymbol{\alpha}) \prod_{m=1}^{M} \mu_m^{\alpha_m - 1}, \tag{35}$$

where $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_M]^\top$, and

$$C(\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{m=1}^{M} \alpha_m\right)}{\prod_{m=1}^{M} \Gamma(\alpha_m)}. \tag{36}$$

The Dirichlet distribution has the following properties:

$$\mathbb{E}\left[\mu_m\right] = \frac{\alpha_m}{\sum_{m=1}^{M} \alpha_m}, \tag{37a}$$

$$\mathbb{E}\left[\ln \mu_m\right] = \Psi(\alpha_m) - \Psi\left(\sum_{n=1}^{M} \alpha_n\right). \tag{37b}$$

7. REFERENCES

[1] R. Schmidt, “Multiple emitter location and signal parameter estimation,” *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[2] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, “Microphone arrays: signal processing techniques and applications.” Springer, 2001, ch. Robust localization in reverberant rooms, pp. 157–180.

[3] 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.

[4] M. I. Mandel, D. P. Ellis, and T. Jebara, “An EM algorithm for localizing multiple sound sources in reverberant environments,” in *Advances in neural information processing systems*, 2007, pp. 953–960.

[5] M. I. Mandel, R. J. Weiss, and D. P. Ellis, “Model-based expectation-maximization source separation and localization,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 382–394, 2009.

[6] J. Taghia, N. Mohammadiha, and A. Leijon, “A variational bayes approach to the underdetermined blind source separation with automatic determination of the number of sources,” in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2012, pp. 253–256.

[7] T. Otsuka, K. Ishiguro, H. Sawada, and H. G. Okuno, “Bayesian unification of sound source localization and separation with permutation resolution,” in *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[8] H. Kameoka, M. Sato, T. Ono, N. Ono, and S. Sagayama, “Bayesian nonparametric approach to blind separation of infinitely many sparse sources,” *IEICE Transactions on Fundamentals of Electronics, Communications and Computer sciences*, vol. 96, no. 10, pp. 1928–1937, 2013.

[9] T. Higuchi, N. Takamune, T. Nakamura, and H. Kameoka, “Underdetermined blind separation and tracking of moving sources based on DOA-HMM,” in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2014, pp. 3191–3195.

[10] Y. Dorfan, O. Schwartz, B. Schwartz, E. A. Habets, and S. Gannot, “Multiple DOA estimation and blind source separation using estimation-maximization,” in *IEEE International Conference on the Science of Electrical Engineering (ICSEE)*, 2016.

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

[12] 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.

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

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

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

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

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

[18] ——, “A Bayesian hierarchical model for speech dereverberation,” in *IEEE International Conference on the Science of Electrical Engineering (ICSEE)*, 2018.

[19] ——, “A Bayesian hierarchical model for blind audio source separation,” in *28th European Signal Processing Conference (EUSIPCO)*, 2020.

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

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

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

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

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

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

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

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

[28] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.