Underdetermined Model-Based Blind Source Separation of Reverberant Speech Mixtures using Spatial Cues in a Variational Bayesian Framework

Victor Popa[*,*], Wenwu Wang[†], Atiyeh Alinaghi[†]

[*]*Faculty of Electronics, Telecommunications and Information Technology, University Politehnica of Bucharest, Romania*
Email: victor_popa_2004@yahoo.com
[†]*Department of Electronic Engineering (FEPS), University of Surrey, Guildford GU2 7XH, U.K.*
Emails: w.wang@surrey.ac.uk, a.alinaghi@surrey.ac.uk

Abstract. In this paper, we propose a new method for underdetermined blind source separation of reverberant speech mixtures by classifying each time-frequency (T-F) point of the mixtures according to a combined variational Bayesian model of spatial cues, under sparse signal representation assumption. We model the T-F observations by a variational mixture of circularly-symmetric complex-Gaussians. The spatial cues, e.g. interaural level difference (ILD), interaural phase difference (IPD) and mixing vector cues, are modelled by a variational mixture of Gaussians. We then establish appropriate conjugate prior distributions for the parameters of all the mixtures to create a variational Bayesian framework. Using the Bayesian approach we then iteratively estimate the hyper-parameters for the prior distributions by optimizing the variational posterior distribution. The main advantage of this approach is that no prior knowledge of the number of sources is needed, and it will be automatically determined by the algorithm. The proposed approach does not suffer from overfitting problem, as opposed to the Expectation-Maximization (EM) algorithm, therefore it is not sensitive to initializations.

1. Introduction

Separating unknown source signals from their speech mixtures without the knowledge of the mixing channels is a problem known as blind (speech) source separation (BSS). The separation process becomes increasingly challenging when the mixtures are presented with noise and room reverberation. Early solutions for the BSS problem often assumed that the number of source signals is smaller than (overdetermined BSS) or equal to (determined BSS) the number of mixtures. Independent component analysis (ICA) [1] has been a popular choice of solutions for this case. If the number of sources is greater than the number of sensors, i.e. the so-called underdetermined BSS, the problem becomes ill-posed, and the solution to this problem will have to rely on extra constraints/assumptions imposed on the separation process. One such idea is to assume that speech signals are sparse in the time-frequency (T-F) domain and only one source is dominant in each T-F point of the mixture, under the asumption of W-disjoint orthogonality [2]. As a result, the mixtures can be transformed to the T-F domain using e.g. short-time Fourier transform (STFT) and each time-frequency point of these mixtures are then clustered into a specific source.

In this work we consider the separation of speech sources from two mixtures under reverberant conditions, mimicking the aspects of binaural hearing in human auditory perception. Hence, it is natural to employ the spatial cues in the separation process, such as the interaural level difference (ILD), the interaural phase difference (IPD), and the mixing vectors as examined in [3,4,5], where separation is achieved by modeling the various observations as Gaussian mixtures and applying the Expectation-Maximization (EM) algorithm to obtain the model parameters. However, there are some limitations with the EM algorithm. First, it needs the number of sources known a priori. Second, if the algorithm is not properly initialized it can lead to the overfitting problem and give poor separation results.

In order to solve the above problems with the EM algorithm, a variational Bayesian approach has been used in some previous work [6,7]. In this case the algorithm does not suffer from poor overfitting and it will automatically detect the number of sources. We examined a similar approach by integrating the spatial cues in the separation process. In this paper we model the T-F point as a mixture of complex-Gaussian distribution, similar to [7] and we also model the IPD and ILD as a mixture of Gaussians [4,5]. In this Bayesian framework, we establish proper conjugate prior distributions on the parameters of the model.

The remainder of the paper is organised as follows. In Section 2, we describe the observation set and the pre-processing performed in order to improve the separation process. Section 3 describes the model and the calculations for the updates of the hyper-parameters, needed in the final algorithm described in Section 4. We then test the proposed algorithm by showing the performance in detection of the number of sources and present a series of results along with our future work.

2. The observation set

Binaural recordings are composed of two signals corresponding to the left and right ear, $l(n)$ and $r(n)$. Each channel is

[*]The work was undertaken during Victor Popa's visit at the Centre for Vision, Speech and Signal Processing, University of Surrey, UK.

a mixture of the original source signal convolved with the impulse response from the source to the senzor, filtered with reverberant noise:

$$\begin{aligned} l(n) &= \sum_{i=1}^{I} s_i(n) * h_{il}(n) * n_l(n) \\ r(n) &= \sum_{i=1}^{I} s_i(n) * h_{ir}(n) * n_r(n) \end{aligned} \tag{1}$$

where I is the number of sources, which in our case is unknown. We will assume a high number of sources and we will see that the variational Bayesian algoritm will determine automatically the number of sources, by subsequently dropping the components that are not present in the mixture. $h_{il}(n)$ and $h_{ir}(n)$ are the impulse responses associated with the head related transfer functions (HRTF) for the left and right ear, and $n_l(n)$ and $n_r(n)$ are the convolutive noise associated with each ear.

The recordings are transformed to the T-F domain using STFT to obtain the spectrogram observations $\mathbf{x}(\omega,t) = [L(\omega,t), R(\omega,t)]^T$. Previous work [1,3,5,7] suggested that by normalizing and pre-whitening the observation the results will be improved. Therefore we will normalize the amplitude of the observations as follows:

$$\mathbf{x}(\omega,t) \leftarrow \frac{\mathbf{x}(\omega,t)}{\sqrt{|L(\omega,t)|^2 + |R(\omega,t)|^2}} \tag{2}$$

The pre-whitening is done by multiplying the normalized observations with the whitening matrix $\mathbf{x}(\omega,t) \leftarrow \mathbf{W}\mathbf{x}(\omega,t)$, where $\mathbf{W} = \sqrt{\mathbf{A}}\mathbf{G}^H$. The values of $\mathbf{A}$ and $\mathbf{G}$ are determined from the eigenvalue decomposition of the correlation matrix $\mathrm{E}\left[\mathbf{x}\mathbf{x}^T\right] = \mathbf{G}\mathbf{A}\mathbf{G}^H$. The normalizing procedure (2) is repeated after whitening.

To determine the IPD and ILD values we calculate the ratio between the STFT of the left and right channels, expressed in terms of the phase and amplitude:

$$\frac{L(\omega,t)}{R(\omega,t)} = 10^{\alpha(\omega,t)/20} e^{j\phi(\omega,t)} \tag{3}$$

In equation (3), $\alpha(\omega,t)$ are the ILD values, and $\phi(\omega,t)$ are the IPD values. Previous work has shown that the ILD can be uniquely associated with a particular source, but the IPD produces abiguity due to phase wrapping. In this article we use the method proposed by Mandel et. al. [4] to solve the phase ambiguity problem.

3. Model description and parameter estimation

In the mixture model, as in [5], we combine three cues, the T-F observations, the ILD values and the IPD values, denoted by $\mathbf{x}(\omega,t)$, $\alpha(\omega,t)$ and $\phi(\omega,t)$, respectively, where the total number of time frames is T and the total number of frequency channels is Ω. The total number of initial sources is denoted by I.

We model the spectrogram observations by a mixture of circularly-symmetric complex-Gaussians [7] with the mixing coefficients $\gamma_x = \{\gamma_{x,i}\}$. We then consider a latent variable or indication vector, $\mathbf{z} = \{z_i\}$, which is a 1-by-I binary vector and has the value 1 if the observation belongs to component i and 0 otherwise. Since we consider the sparsity assumption, the vector will have only one value of 1 and the rest are 0s, for any observation. In a particular frequency channel, all the observations are i.i.d., therefore we will omit the frequency index hereafter. Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ denote the observation set in a particular frequency channel and $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_T\}$ the latent variables associated with each observation. The conditional distribution of $\mathbf{Z}$ given the mixing coefficients will be:

$$p(\mathbf{Z}|\gamma_x) = \prod_{t=1}^{T}\prod_{i=1}^{I} \gamma_{x,i}^{z_{ti}} \tag{4}$$

We express the conditional distribution of $\mathbf{X}$ given the latent variables $\mathbf{Z}$, the mean $\mu_x = \{\mu_{x,i}\}$ and the precision $\lambda_x = \{\lambda_{x,i}\}$ of the complex Gaussians as follows:

$$p(\mathbf{X}|\mathbf{Z},\mu_x,\lambda_x) = \prod_{t=1}^{T}\prod_{i=1}^{I} \mathcal{N}_c(\mathbf{x}_t|\mu_{x,i},\lambda_{x,i}^{-1})^{z_{ti}} \tag{5}$$

$$\mathcal{N}_c(\mathbf{x}_t|\mu_{x,i},\lambda_{x,i}^{-1}) = \frac{1}{\left(\pi\lambda_{x,i}^{-1}\right)^2} e^{-\lambda_{x,i}||\mathbf{x}_t - (\mu_{x,i}^H \mathbf{x}_t)\mu_{x,i}||^2} \tag{6}$$

We model the ILD observations by a mixture of Gaussians, considering the set of observations $\alpha = \{\alpha_1, \ldots, \alpha_T\}$ in a particular frequency channel.

$$p(\mathbf{Z}|\gamma_\alpha) = \prod_{t=1}^{T}\prod_{i=1}^{I} \gamma_{\alpha,i}^{z_{ti}} \tag{7}$$

$$p(\alpha|\mathbf{Z},\mu_\alpha,\lambda_\alpha) = \prod_{t=1}^{T}\prod_{i=1}^{I} \mathcal{N}(\alpha_t|\mu_{\alpha,i},\lambda_{\alpha,i}^{-1})^{z_{ti}} \tag{8}$$

$$\mathcal{N}(\alpha_t|\mu_{\alpha,i},\lambda_{\alpha,i}^{-1}) = \frac{1}{\sqrt{2\pi\lambda_{\alpha,i}^{-1}}} e^{-\frac{\lambda_{\alpha,i}}{2}(\alpha_t - \mu_{\alpha,i})^2} \tag{9}$$

To model the IPD we use the approach presented in [4], by mapping in a top-down fashion a set of discrete values of ITDs to their corresponding IPDs, thereby solving the problem of spatial aliasing for most frequencies. We determine the phase residuals at each T-F point:

$$\hat{\phi}(\omega,t;\tau) = \arg\left(e^{j\phi(\omega,t)} e^{-j\omega\tau(\omega)}\right) \tag{10}$$

For each T-F point we choose a fixed number of equally spaced τ time delays between -15 and 15 samples. After determining the residuals for these values, we choose I most dominant τ delays from the PHAT histogram [8]. We model the I selected phase residuals by a mixture of Gaussians with the mixing coefficients $\gamma_{\hat{\phi},i}$ which represents the probability of a phase residual comming from source i.

Considering the observation set for a frequency channel, $\hat{\Phi} = \{\hat{\phi}_{ti}\}$, the mixing coefficients and the latent variables $\mathbf{Z}$, we can express the conditional probability distributions:

$$p\left(\mathbf{Z}|\gamma_{\hat{\phi}}\right) = \prod_{t=1}^{T}\prod_{i=1}^{I}\gamma_{\hat{\phi},i}^{z_{ti}} \tag{11}$$

$$p(\hat{\Phi}|\mathbf{Y},\mu_{\hat{\phi}},\lambda_{\hat{\phi}}) = \prod_{t=1}^{T}\prod_{i=1}^{I}\mathcal{N}\left(\hat{\phi}_{ti}|\mu_{\hat{\phi},i},\lambda_{\hat{\phi},i}^{-1}\right)^{z_{ti}} \tag{12}$$

where $\mathcal{N}\left(\hat{\phi}_{ti}|\mu_{\hat{\phi},i},\lambda_{\hat{\phi},i}^{-1}\right)$ is a Gaussian distribution, similar to (9).

By combining the distributions of the three observations given the latent variables and the paramaters, we obtain one mixing coefficient for all the mixtures, denoted by $\gamma = \{\gamma_i\}$ where $\gamma_i = \gamma_{\mathbf{x},i}\gamma_{\alpha,i}\gamma_{\hat{\phi},i}$.

In the next step we establish proper conjugate priors for each of the parameters of the distributions. The mixing coefficients are modeled with a Dirichlet distribution:

$$p(\gamma) = Dir(\gamma|\mathbf{a}_0) = B(\mathbf{a}_0)\prod_{i=1}^{I}\gamma_{t\tau}^{a_0-1} \tag{13}$$

where $\mathbf{a}_0$ is a vector of I equal values a_0 which represent the concentration parameters. They are equal because we assume that every component has the same probability. If the concentration parameters are not equal and they form the vector $\mathbf{a}$ then we define $B(\mathbf{a}) = \frac{\prod_{i=1}^{I}\Gamma(a_i)}{\Gamma(\sum_{i=1}^{I}a_i)}$.

For the mean and precision of the Gaussian distributions we choose Gauss-Gamma conjugate priors as follows:

$$\begin{aligned} &p(\mu_x,\lambda_x) = p(\mu_x|\lambda_x)p(\lambda_x) = \\ &\prod_{i=1}^{I}\mathcal{N}_c\left(\mu_{x,i}|\mathbf{m}_{x,0},(\lambda_{x,i}\beta_{x,0}\mathbf{I})^{-1}\right)\mathcal{G}(\lambda_{x,i}|b_{x,0},c_{x,0}) \end{aligned} \tag{14}$$

$$\begin{aligned} &p(\mu_\alpha,\lambda_\alpha) = p(\mu_\alpha|\lambda_\alpha)p(\lambda_\alpha) = \\ &\prod_{i=1}^{I}\mathcal{N}\left(\mu_{\alpha,i}|m_{\alpha,0},(\lambda_{\alpha,i}\beta_{\alpha,0})^{-1}\right)\mathcal{G}(\lambda_{\alpha,i}|b_{\alpha,0},c_{\alpha,0}) \end{aligned} \tag{15}$$

$$\begin{aligned} &p(\mu_{\hat{\phi}},\lambda_{\hat{\phi}}) = p(\mu_{\hat{\phi}}|\lambda_{\hat{\phi}})p(\lambda_{\hat{\phi}}) = \\ &\prod_{i=1}^{I}\mathcal{N}\left(\mu_{\hat{\phi},i}|m_{\hat{\phi},0},(\lambda_{\hat{\phi},i}\beta_{\hat{\phi},0})^{-1}\right)\mathcal{G}(\lambda_{\hat{\phi},i}|b_{\hat{\phi},0},c_{\hat{\phi},0}) \end{aligned} \tag{16}$$

Here, $\mathcal{N}_c\left(\mu_{x,i}|\mathbf{m}_{x,0},(\lambda_{x,i}\beta_{x,0}\mathbf{I})^{-1}\right)$ is a complex Gaussian distribution with mean $\mathbf{m}_{x,0}$ and precision $\lambda_{x,i}\beta_{x,0}\mathbf{I}$, of the form:

$$\begin{aligned} &\mathcal{N}_c\left(\mu_{x,i}|\mathbf{m}_{x,0},(\lambda_{x,i}\beta_{x,0}\mathbf{I})^{-1}\right) = \\ &\frac{1}{\left(\pi\left(\lambda_{x,i}\beta_{x,0}\right)^{-1}\right)^2}e^{-\lambda_{x,i}(\mu_{x,i}-\mathbf{m}_{x,0})^H\beta_{x,0}\mathbf{I}(\mu_{x,i}-\mathbf{m}_{x,0})} \end{aligned} \tag{17}$$

The Gamma distributions with the shape parameter b_0 and rate parameter c_0 are given by:

$$\mathcal{G}(\lambda_i|b_0,c_0) = \frac{1}{\Gamma(b_0)}c_0^{b_0}\lambda_i^{b_0-1}e^{-c_0\lambda_i} \tag{18}$$

Now that we have all the observations and parameter distributions, we can express the joint distribution of the observations, considering all the latent variables $\mathbf{Z}$ and parameters $\Theta = \left\{\gamma,\mu_x,\lambda_x,\mu_\alpha,\lambda_\alpha,\mu_{\hat{\phi}},\lambda_{\hat{\phi}}\right\}$.

$$\begin{aligned} p(\mathbf{X},\alpha,\hat{\Phi},\mathbf{Z},\Theta) =& p(\mathbf{X}|\mathbf{Z},\mu_x,\lambda_x)p(\mu_x|\lambda_x)p(\lambda_x) \\ & p(\alpha|\mathbf{Z},\mu_\alpha,\lambda_\alpha)p(\mu_\alpha|\lambda_\alpha)p(\lambda_\alpha) \\ & p(\hat{\Phi}|\mathbf{Z},\mu_{\hat{\phi}},\lambda_{\hat{\phi}})p(\mu_{\hat{\phi}}|\lambda_{\hat{\phi}})p(\lambda_{\hat{\phi}}) \\ & p(\mathbf{Z}|\gamma)p(\gamma) \end{aligned} \tag{19}$$

The variational posterior distribution of the latent variables and component parameters can be expressed as:

$$q(\mathbf{Z},\Theta) = q(\mathbf{Z})q(\Theta) \tag{20}$$

In order to optimize the variational posterior distribution $q(\mathbf{Z},\Theta)$ we have to optimize the posterior distribution of the latent variables $q(\mathbf{Z})$ and parameter distributions $q(\Theta)$. We can aproximate the log of the optimized distribution $q^*(\mathbf{Z})$ by [9]:

$$\ln q^*(\mathbf{Z}) = \mathrm{E}_\Theta\left[\ln p(\mathbf{X},\alpha,\hat{\Phi},\mathbf{Z},\Theta)\right] + const \tag{21}$$

Also we can express this for the optimum parameter distributions:

$$\ln q^*(\Theta) = \mathrm{E}_\mathbf{Z}\left[\ln p(\mathbf{X},\alpha,\hat{\Phi},\mathbf{Z},\Theta)\right] + const \tag{22}$$

Considering equation (21) we can determine that the posterior distribution has the form:

$$\ln q^*(\mathbf{Z}) = \sum_{t=1}^{T}\sum_{i=1}^{I}z_{ti}\ln\rho_{ti} + const \tag{23}$$

where $\ln\rho_{ti}$ is a notation for:

$$\begin{aligned} \ln\rho_{ti} =& \mathrm{E}_{\gamma_i}[\ln\gamma_i] - 2\ln\pi - \ln 2\pi + 2\mathrm{E}_{\lambda_{x,i}}[\ln\lambda_{x,i}] \\ &+ \frac{1}{2}\mathrm{E}_{\lambda_{\alpha,i}}[\ln\lambda_{\alpha,i}] + \frac{1}{2}\mathrm{E}_{\lambda_{\hat{\phi},i}}\left[\ln\lambda_{\hat{\phi},i}\right] \\ &- \mathrm{E}_{\mu_{x,i},\lambda_{x,i}}\left[\lambda_{x,i}\|\mathbf{x}_t - \left(\mu_{x,t}^H\mathbf{x}_t\right)\mu_{x,i}\|^2\right] \\ &- \frac{1}{2}\mathrm{E}_{\mu_{\alpha,i},\lambda_{\alpha,i}}\left[\lambda_{\alpha,i}\left(\alpha_t - \mu_{\alpha,i}\right)^2\right] \\ &- \frac{1}{2}\mathrm{E}_{\mu_{\hat{\phi},i},\lambda_{\hat{\phi},i}}\left[\lambda_{\hat{\phi},i}\left(\hat{\phi}_{ti} - \mu_{\hat{\phi},i}\right)^2\right] \end{aligned} \tag{24}$$

We observe that equation (23) is proportional to a multinomial distribution which is a consequence of choosing the proper conjugate priors on our distributions.

$$q^*(\mathbf{Z}) \propto \prod_{t=1}^{T}\prod_{i=1}^{I}\rho_{ti}^{z_{ti}} \tag{25}$$

In order to obtain a proper distribution we normalize the parameters for each time frame and determine the distribution of the latent variables:

$$r_{ti} = \frac{\rho_{ti}}{\sum_{i=1}^{I} \rho_{ti}} \tag{26}$$

$$q^*(\mathbf{Z}) = \prod_{t=1}^{T} \prod_{i=1}^{I} r_{ti}^{z_{ti}} \tag{27}$$

Since r_{ti} sum to 1 over all i at every time frame t, these represent the probability of the observation set belonging to the i^{th} component of our model, and we can determine them by using the current model parameters as presented in equation (24).

This is an expectation step, where we use the current parameter distributions to determine the responsibilities, very similar to the EM algorithm. To determine the expectations from equation (24) we use the following:

$$\begin{aligned}
&\mathrm{E}_{\gamma_i}[\ln \gamma_i] = \psi(a_i) - \psi\left(\sum_{i=1}^{I} a_i\right) \\
&\mathrm{E}_{\lambda_{x,i}}[\ln \lambda_{x,i}] = \psi(b_{x,i}) - \ln c_{x,i} \\
&\mathrm{E}_{\lambda_{\alpha,i}}[\ln \lambda_{\alpha,i}] = \psi(b_{\alpha,i}) - \ln c_{\alpha,i} \\
&\mathrm{E}_{\lambda_{\hat{\phi},i}}\left[\ln \lambda_{\hat{\phi},i}\right] = \psi\left(b_{\hat{\phi},i}\right) - \ln c_{\hat{\phi},i} \\
&\mathrm{E}_{\mu_{x,i},\lambda_{x,i}}\left[\lambda_{x,i} \| \mathbf{x}_t - \left(\mu_{x,i}^H \mathbf{x}_t\right) \mu_{x,i} \|^2\right] = \\
&\qquad \mathbf{x}_t^H \left[\frac{b_{x,i}}{c_{x,i}} \left(\mathbf{I} - \mathbf{m}_{x,i} \mathbf{m}_{x,i}{}^H\right) - \beta_{x,i}^{-1}\right] \mathbf{x}_t \\
&\mathrm{E}_{\mu_{\alpha,i},\lambda_{\alpha,i}}\left[\lambda_{\alpha,i} \left(\alpha_t - \mu_{\alpha,i}\right)^2\right] = \\
&\qquad \frac{b_{\alpha,i}}{c_{\alpha,i}} \left(\alpha_t - m_{\alpha,i}\right)^2 + \beta_{\alpha,i}^{-1} \\
&\mathrm{E}_{\mu_{\hat{\phi},i},\lambda_{\hat{\phi},i}}\left[\lambda_{\hat{\phi},i} \left(\hat{\phi}_{ti} - \mu_{\hat{\phi},i}\right)^2\right] = \\
&\qquad \frac{b_{\hat{\phi},i}}{c_{\hat{\phi},i}} \left(\hat{\phi}_{ti} - m_{\hat{\phi},i}\right)^2 + \beta_{\hat{\phi},i}^{-1}
\end{aligned} \tag{28}$$

where $\psi(\cdot)$ is the digamma function.

The next step is to determine the updates for the hyper-parameters using (22). This is done by extracting the dependencies of the different hyper-parameters and by taking into account that $\mathrm{E}[z_{ti}] = r_{ti}$ from (27). Therefore we determine the update for the prior on the mixing coefficients:

$$a_i = \sum_{t=1}^{T} r_{ti} + a_0 \tag{29}$$

To determine the updates for the prior on the mixing vectors, we observe that a Gauss-Gamma distribution has been obtained from (22), as a consequence of choosing proper conjugate prior distributions. We determine the updates of the hyper-parameters for the mixing vectors as follows:

$$\begin{aligned}
\beta_{x,i} &= \beta_{x_0} \mathbf{I} - \sum_{t=1}^{T} r_{ti} \left(\mathbf{x}_t \mathbf{x}_t{}^H\right) \\
\mathbf{m}_{x,i} &= \beta_{x,i}{}^{-1} \beta_{x_0} \mathbf{I} \mathbf{m}_{x,0} \\
b_{x,i} &= b_{x,0} + 2 \sum_{t=1}^{T} r_{ti} \\
c_{x,i} &= c_{x,0} + \sum_{t=1}^{T} r_{ti} \left(\mathbf{x}_t{}^H \mathbf{x}_t\right) + \mathbf{m}_{x,0}{}^H \beta_{x,0} \mathbf{I} \mathbf{m}_{x,0} \\
&\quad - \mathbf{m}_{x,i}{}^H \beta_{x,i} \mathbf{m}_{x,i}
\end{aligned} \tag{30}$$

The updates of the hyper-parameters for the binaural cues can be determined by using the same approach as for the mixing vectors. Therefore the updates for the ILD parameters are:

$$\begin{aligned}
\beta_{\alpha,i} &= \beta_{\alpha_0} + \sum_{t=1}^{T} r_{ti} \\
m_{\alpha,i} &= \beta_{\alpha,i}{}^{-1} \left(\sum_{t=1}^{T} r_{ti} \alpha_t + m_{\alpha,0} \beta_{\alpha,0}\right) \\
b_{\alpha,i} &= b_{\alpha,0} + \frac{1}{2} \sum_{t=1}^{T} r_{ti} \\
c_{\alpha,i} &= c_{\alpha,0} + \frac{1}{2} \sum_{t=1}^{T} r_{ti} \alpha_t{}^2 + \frac{1}{2} m_{\alpha,0}{}^2 \beta_{\alpha,0} - \frac{1}{2} m_{\alpha,i}{}^2 \beta_{\alpha,i}
\end{aligned} \tag{31}$$

The prior updates for the IPD model can be derived from (27) and we obtain the following:

$$\begin{aligned}
\beta_{\hat{\phi},i} &= \beta_{\hat{\phi}_0} + \sum_{t=1}^{T} r_{ti} \\
m_{\hat{\phi},i} &= \beta_{\hat{\phi},i}{}^{-1} \left(\sum_{t=1}^{T} r_{ti} \hat{\phi}_{ti} + m_{\hat{\phi},0} \beta_{\hat{\phi},0}\right) \\
b_{\hat{\phi},i} &= b_{\hat{\phi},0} + \frac{1}{2} \sum_{t=1}^{T} r_{ti} \\
c_{\hat{\phi},i} &= c_{\hat{\phi},0} + \frac{1}{2} \sum_{t=1}^{T} r_{ti} \hat{\phi}_{ti}^2 + \frac{1}{2} m_{\hat{\phi},0}{}^2 \beta_{\hat{\phi},0} - \frac{1}{2} m_{\hat{\phi},i}{}^2 \beta_{\hat{\phi},i}
\end{aligned} \tag{32}$$

One important observation is that all the calculations of the hyper-parameters must be done in order, since one hyper-parameter depends on the determination of the previous one. This step is similar to the maximization step of the EM algoritm, and it optimizes the values of the hyper-parameters at each iteration.

4. Algorithm description

The algoritm begins by transforming the time-domain samples of the observations into the T-F domain via STFT, using a Hann window. The T-F points $\mathbf{x}_t$ will be normalized and prewhitened so that we eliminate the influence of speaker amplitude. Using the T-F points we determine the ILD, α_t. For the

IPDs we first use the top-down approach [4] to determine the IPD residuals, and we select the first I most probable time differences from the PHAT histogram [8], and we will only select the residuals that correspond to these time differences, $\hat{\phi}_{ti}$.

Once the observation are ready for processing we select a frequency channel and we begin the algoritm. First we initialize the hyper-parameters of our models. Initialization is not an issue since the framework does not suffer from overfitting, therefore we will set the initial values of the hyper-parameters to a low value so that the posterior will be influenced mainly by the observations [9].

We then initialize the responsibilities of every T-F point according to the PHAT histogram. By doing this and because of the fact that the IPD residuals are selected only for some fixed time differences, there will be no permutation misalignment between the frequency channels.

After the initialization we start the main loop consisting of two phases. In the first phase we determine the new hyper-parameter values considering the current responsibilites using equations (29), (30), (31) and (32). In the second phase we use the previously determined hyper-parameter values to determine the new responsibilities, by determining the expectations (28), needed to determine the responsibilities using (24) and (26). These two steps are cycled for a number of iterations.

To separate the sources we determine the binary masks for each source, by associating 1 to the source that has the largest responsibility. Therefore the mask for a particular source i will be $M_i(\omega,t) = 1$ if $r_{ti} \geq r_{tk}, \forall k \neq i$ and $M_i(\omega,t) = 0$ if $r_{ti} < r_{tk}, \forall k \neq i$. The reconstructed time-frequency signal can be determined by $\hat{s}_i(\omega,t) = M_i(\omega,t)\mathbf{x}(\omega,t)$. We then apply the inverse short-time Fourier transform to obtain the time domain signal.

The algorithm does not know the actual number of sources and it starts with a high and assumed one. The true number of sources can be determined by the fact that the components that do not belong to the mixture will have low responsibilities and therefore the masks will be null. This is also reflected in the values of the hyper-parameters, which will be the same as the initial ones.

5. Experimental results

For the experiments, we chose the TIMIT dataset which is composed of 6300 speech utterances spoken by 630 native American English speakers, similar to [4]. We selected 15 utterances spoken by both male and female speakers at random of the same length (about 3s) and then shortened to 2.5 s as in [5].

For the mixing process we used the binaural room impulse responses (BRIR) measured by Hummersone [10] at the University of Surrey. The measurements were made using a dummy head and torso in four different types of rooms, named A, B, C and D with the acoustical properties presented in Table 1. We selected 100 unique sets of 3 sources from the 15 utterances. The sources were placed at -60°, 0° and 60° azimuth, where zero azimuth is in front of the dummy head. The mixtures were obtained by convolving each utterance with the respective BRIR and then summing the results. We also mixed the sources under anechoic conditions, with the room number being denoted by N.

Table 1. Room acoustical properties in initial time delay gap (ITDG), direct-to-reverberant ratio (DRR) and reverberation time T_{60} [10].

Room	ITDG [ms]	DRR [dB]	T_{60} [s]
A	8.72	6.09	0.32
B	9.66	5.31	0.47
C	11.9	8.82	0.68
D	21.6	6.12	0.89

Using the generated mixtures for each room we then apply the proposed variational Bayesian algorithm starting with 5 initial sources. We use a Hann window of 1024 samples with 75% overlapping when performing the STFT.

We apply the algorithm to the mixtures and we first determine the accuracy of the number of sources estimated for each type of room. The results are presented in Table 2.

Table 2. The accuracy of the number of sources estimated by the proposed algorithm.

Room	N	A	B	C	D
Accuracy [%]	100	81	75	92	72

The number of sources is not correctly determined in some cases because at some frequency channels we do have some ambiguity in determining the IPD because of the overlapping due to phase aliasing [4] and therefore if the reverberation is high the frequency band in which this occurs will be larger.

To establish the performance of the algorithm we used the signal-to-distortion ratio (SDR), described in [11]. The determined SDRs for the experiments can be seen in Table 3. The values shown represent the mean of the SDR in each room after separation, considering the target source as the one at zero azimuth. We also determined the mean over all rooms.

Table 3. Separation results of the proposed variational Bayesian algorithm..

Room	N	A	B	C	D	Mean
SDR [dB]	8.25	7.71	4.95	7.48	3.90	6.77

The separation process yields acceptable performance, but it has the advantage that the initialization is not that strict, since we used general values to initialize the priors. The main advantage is that it can achieve this level of separation without any prior knowledge of the number of sources. As a comparison, in Figure 1, 2 and 3 we show the T-F plots of the initial source signals, the mixtures and the separated signals.

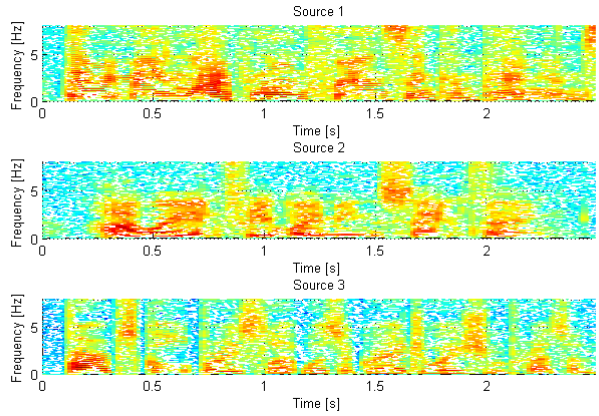

Figure 1. Spectrograms of the original sources.

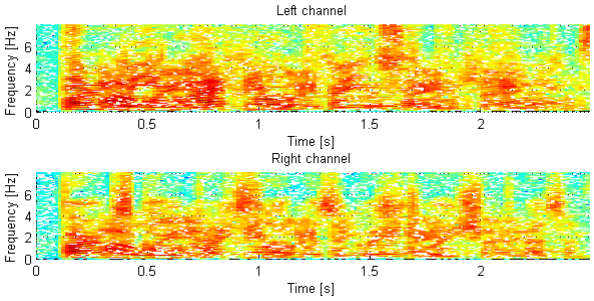

Figure 2. Spectrograms of the mixtures.

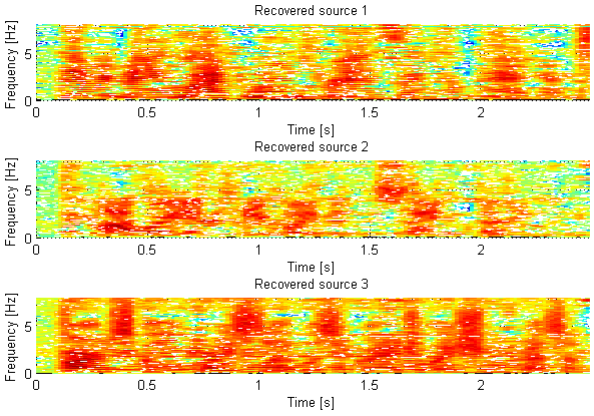

Figure 3. Spectrograms of the separated signals.

6. Conclusions and future work

A variational Bayesian framework based reverberant speech separation method has been presented. The proposed algorithm benefits from the fact that the Bayesian approach is less sensitive to improper initializations, and it also automatically determines the number of sources, which can be an advantage in real life recordings. The preliminary results suggest that the proposed method offers an acceptable performance. In future work a performance comparison will be conducted between the Bayesian approach and the EM algorithm. Further improvements will be made to the estimation of the number of sources and to the separation performance of the proposed algorithm. In addition, the convergence could be studied by determining the variational lower bound. Some studies will also be performed for assessing the influence of the angle between the source signals.

References

[1] A. Hyvärinen, J. Karhunen, E. Oja. "Independent component analysis", *J. Wiley*, New York, 2001.

[2] Ö. Yilmaz, S. Rickard. "Blind separation of speech mixtures via time-frequency masking", *IEEE Transactions on Signal Processing*, vol. 52, pp. 1830–1847, 2004.

[3] H. Sawada, S. Araki, S. Makino. "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. pp. 516–527, 2011.

[4] M. Mandel, R. Weiss, D. Ellis. "Model-based expectation-maximization source separation and localization", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 382 – 394, 2010.

[5] A. Alinaghi, W. Wang, P. Jackson. "Integrating binaural cues and blind source separation method for separating reverberant speech mixtures", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 209–212, 2011.

[6] L. Ma, A. Tsoi. "A variational Bayesian approach to number of sources estimation for multichannel blind deconvolution", *Signal, Image and Video Processing*, vol. 2, pp. 107–127, 2008.

[7] J. Taghia, N. Mohammadiha, A. Leijon. "A variational Bayes approach to the underdetermined blind source separation with automatic determination of the number of sources", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 253–256, 2012.

[8] P. Aarabi. "Self-localizing dynamic microphone arrays", *IEEE Transactions on Systems, Man, and Cybernetics - Part C*, vol. 32, pp. 474–484, 2002.

[9] C. Bishop. "Pattern recognition and machine learning", *Springer*, pp. 461–485, 2006.

[10] C. Hummersone. "A psychoacoustic engineering approach to machine sound source separation in reverberant environments", *Ph.D. thesis, Music and Sound Recording, University of Surrey, UK*, 2011.

[11] E. Vincent, R. Gribonval, C. Fevotte. "Performance measurement in blind audio source separation", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 1462–1469, 2000.