ESTIMATION OF THE NUMBER OF SOURCES IN MEASURED SPEECH MIXTURES WITH COLLAPSED GIBBS SAMPLING

Yang Sun, Yang Xian, Pengming Feng, Jonathon A. Chambers, Syed Mohsen Naqvi

Intelligent Sensing and Communications Group,
School of Engineering, Newcastle University, NE1 7RU, UK
{y.sun29, y.xian2, p.feng2, jonathon.chambers, mohsen.naqvi}@newcastle.ac.uk

ABSTRACT

In blind source separation (BSS), the number of sources present in the measured speech mixtures is unknown. The focus of this work is therefore to automatically estimate the number of sources from binaural speech mixtures. Collapsed Gibbs sampling (CGS), a Markov chain Monte Carlo (MCMC) technique, is used to obtain samples from the joint distribution of the speech mixtures. Then the Chinese Restaurant Process (CRP) within the framework of the Dirichlet Process (DP) is exploited to cluster samples into different components to finally estimate the number of speakers. The accuracy of the proposed method, under different reverberant environments, is evaluated with real binaural room impulse responses (BRIRs) and speech signals from the TIMIT database. The experimental results confirm the accuracy and robustness of the proposed method.

Index Terms— Blind source separation, Collapsed Gibbs sampling, Chinese restaurant process, Dirichlet process

1. INTRODUCTION

The signal processing community exploits BSS to attempt to solve the machine cocktail party problem (CPP) [1]. In most of the BSS approaches, not only are speech mixtures acquired by microphones, but also the number of speakers in the mixtures is assumed as a prior knowledge [2]. This assumption limits the BSS application in real scenarios, particularly in the underdetermined cases, where the number of sources is greater than the number of sensors.

In order to overcome this limitation, some approaches have been proposed to determine the number of sources for BSS [3–8]. In [3, 4], multimodal (audio-video) information is exploited and the video modality is used to determine the number of speakers and assist the source separation process. However, in some cases, only the speech mixtures are available, which limits the application of the above methods.

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) Grant number EP/K014307 and the MOD University Defence Research Collaboration in Signal Processing.

Nonparametric Bayesian refers to a class of techniques that allows parameter dimension to depend on the data samples [9]. In [5, 6], full Bayesian inference was assumed over all model parameters to calculate the number of sources. The variational expectation maximization (VEM) algorithm was applied to count the number of active sources in a speech mixture. However, in these methods, the maximum possible number of sources and full set of parameters need to be initialized. In recent work [7], the DP was utilized to determine the optimal number of mixture components. In this method, however, the speech mixtures were generated in a non-reverberant environment and the parameters of binaural cues were used.

In the proposed method, the acquired speech mixtures of the left and right channels are transformed to the frequency domain and a Gaussian Mixture Model (GMM) is used to determine the number of sources. Moreover, the CGS method, from a class of convenient MCMC algorithms, is used with the DP [10] to obtain the samples from the joint distribution. In this sampling method, only the latent parameters related to the observed data and hyperparamters of the prior distribution are exploited [11]. Then, the CRP is exploited to cluster samples into different components and obtain the number of sources [12]. Hence, the proposed method can be used to estimate the number of sources from the speech mixtures which are generated in the reverberant environment. The computational cost of this approach is also relatively low.

The remainder of the paper is organized as follows, in Section 2, the model of the binaural speech mixtures and DP are described. In Section 3, the proposed method based on DP is presented; experimental results are shown in Section 4 to confirm the accuracy of the proposed approach. Finally, conclusions are drawn in Section 5.

2. MODEL AND DIRICHLET PROCESS DESCRIPTION

The aim of solving underdetermined BSS is to mimic human auditory perception [13]. Therefore, the microphones are named as left and right channels. Assuming the speech sources are $s(t)$, and $l(t)$ and $r(t)$ are signals acquired by the left and right microphones, respectively. According to [14],

978-1-5386-1663-5/17/$31.00 ©2017 IEEE

the signals are represented as $l(t) = s(t) * h_l(t) + n_l(t)$ and $r(t) = s(t) * h_r(t) + n_r(t)$, where $h_l(t)$ and $h_r(t)$ are the impulse responses of the left and right channels, respectively. The $n_l(t)$ and $n_r(t)$ are the additional noises of left and right channels.

In the frequency domain, the Fourier transform of the left and right channels are $L(\omega, t)$ and $R(\omega, t)$, respectively [14]. The interaural spectrogram is a kind of sparse representation [15], which can be obtained by the ratio of $L(\omega, t)$ to $R(\omega, t)$. The interaural spectrogram is expressed as:

$$\frac{L(\omega,t)}{R(\omega,t)} = e^{-j\omega(\tau_l-\tau_r)}H(\omega)N(\omega,t) \tag{1}$$

where $N(\omega, t) = N_l(\omega, t)/N_r(\omega, t)$ represents the Fourier transform of the noise and $H(\omega) = \mathcal{F}\{h_l(t)\}/\mathcal{F}\{h_r(t)\}$ is the ratio of Fourier transforms of the impulse responses [14].

According to the assumption of W-disjoint orthogonality (W-DO) [16], at most one source is active at each time-frequency (T-F) point in the interaural spectrogram. Therefore, the mixture distribution of T-F points can be used to cluster and determine the optimal number of speakers.

We assume D is the number of channels, N is the number of frequency bins in the mixture and K is the number of time frames in the spectrogram. The proposed approach is a model-based clustering method, which assumes the parameters of data points are generated by a mixture model. Then, the number of components in the mixture is calculated by using the latent parameters after DP with the CGS [17].

Hence, in the Dirichlet process mixture model (DPMM), considering each data point x_i is generated from a distribution defined by parameter θ_i and $\boldsymbol{\theta} = \{\theta_1, ..., \theta_N\}$. By using DP, the model can be described as:

$$G \sim DP(\alpha, H) \tag{2}$$

$$\theta_i \sim G \tag{3}$$

$$x_i \sim F(\theta_i) \tag{4}$$

where α is the concentration parameter to control the dispersion of the new distribution G, H is the basement distribution and θ_i is generated from the new discrete distribution G. F is a distribution, which generates x_i. Each observation is based on an independent parameter θ_i. Because of the nonuniqueness of parameter sets and G is discrete, there is always a non-zero probability of two samples colliding [18].

From the Bayesian rule [19], the relation between parameters and observed data point is:

$$p(\theta_i|x_i) \propto p(x_i|\theta_i) \times p(\theta_i) \tag{5}$$

where $p(\theta_i|x_i)$ is the posterior probability, $p(x_i|\theta_i)$ is the likelihood function and $p(\theta_i)$ is the prior probability [19]. The method can be implemented easily based on the CGS with models based on conjugate prior distributions [20]. For every observed data point x_i, there is a distribution and a parameter set θ_i which is generated from the ith distribution. From (2), (3) and (4), $\boldsymbol{\theta}$ must have some identical factors, which indicates these parameters come from the same distribution. Therefore, the latent variables are associated with these parameters and represent the clusters of the observed data. Based on the description above, the graphical model is shown in Figure 1:

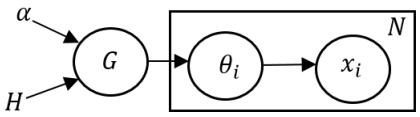

Fig. 1: The graphical model of the proposed DP. The number of input data points is defined as N.

3. COLLAPSED GIBBS SAMPLING AND DATA CLUSTERING

For the observed data point x_N, its cluster assignment z_N is considered as the latent variable. By using the same process as in [19, 21], the conjugate prior is exploited, which is assumed as the Normal Inverse Wishart distribution in the proposed method.

The predictive distribution of $p(x_N|x_{1:N-1})$ is used in the CRP, where $x_{1:N-1} = \{x_1, ..., x_{N-1}\}$. Assume $x_{1:N-1}$ and x_N are generated from the same distribution and ω is the parameter set of this distribution. It can be obtained as:

$$\begin{aligned} p(x_N|x_{1:N-1}) &= \int_\omega p(x_N, \omega|x_{1:N-1})d\omega \\ &= \int_\omega p(x_N|\omega)p(\omega|x_{1:N-1})d\omega \end{aligned} \tag{6}$$

From (6), the predictive distribution of $p(\theta_N|\theta_{1:N-1})$ can be expressed similarly :

$$p(\theta_N|\theta_{1:N-1}) = \int_G p(\theta_N|G)p(G|\theta_{1:N-1})dG \tag{7}$$

where $\theta_{1:N-1} = \{\theta_1, ..., \theta_{N-1}\}$.

Therefore, with the setting of the DP, the latent variable z_N is utilized to cope with the clustering problem and map the parameters sets into a number of clusters, the expression is:

$$p(z_N = m|z_{1:N-1}) = \frac{p(z_N = m, z_{1:N-1})}{p(z_{1:N-1})} \tag{8}$$

where $z_{1:N-1} = \{z_1, ..., z_{N-1}\}$ and $z_N = m$ means the latent variable z_N belongs to component m. The predictive distribution (8) is the expression of the probability that the new data point x_N belongs to the component m.

Because the DP is based on an infinite mixture model, in order to simplify the expression and reduce cost of calculation, we assume the number of clusters is C, thus (8) is expressed as:

$$\frac{p(z_N = m, z_{1:N-1})}{p(z_{1:N-1})}$$
$$= \frac{\int_{\boldsymbol{p}} p(z_N = m, z_{1:N-1}|\boldsymbol{p})p(\boldsymbol{p})d(\boldsymbol{p})}{\int_{\boldsymbol{p}} p(z_{1:N-1}|\boldsymbol{p})p(\boldsymbol{p})d(\boldsymbol{p})}$$
$$= \frac{\int_{\boldsymbol{p}} p(z_N = m, z_{1:N-1}|\boldsymbol{p})Dir(\frac{\alpha}{C}, ..., \frac{\alpha}{C})d(\boldsymbol{p})}{\int_{\boldsymbol{p}} p(z_{1:N-1}|\boldsymbol{p})Dir(\frac{\alpha}{C}, ..., \frac{\alpha}{C})d(\boldsymbol{p})} \quad (9)$$

where $\boldsymbol{p} = \{p_1, ..., p_C\}$, the proportional coefficient vector for components, Dir represents the Dirichlet distribution.

In [18] :

$$\int_{\boldsymbol{p}} p(n_1, ..., n_C|\boldsymbol{p})p(\boldsymbol{p}|\alpha_1, ..., \alpha_C)d(\boldsymbol{p})$$
$$= \frac{n!}{n_1!, ..., n_C!} \cdot \frac{\Gamma(\sum_{i=1}^{C} \alpha_i)}{\prod_{i=1}^{C} \Gamma(\alpha_i)} \cdot \frac{\prod_{i=1}^{C} \Gamma(\alpha_i + n_i)}{\Gamma(\sum_{i=1}^{C} \alpha_i + n)} \quad (10)$$

where $n_1, ..., n_C$ are the numbers of data points belonging to different components, which satisfy a multinomial distribution. The Gamma function satisfies $\Gamma(X) = (X-1) \cdot \Gamma(X-1)$. Therefore, (10) can be expressed as:

$$\frac{\Gamma(n_{m,N-1} + \frac{\alpha}{C} + 1) \prod_{l=1}^{C} \Gamma(n_{l,N-1})}{(N + \alpha - 1)\Gamma(n + \alpha - 1)} \times \frac{\Gamma(N + \alpha - 1)}{\prod_{l=1}^{C} \Gamma(n_{l,N-1})}$$
$$= \frac{n_{m,N-1}}{(N + \alpha - 1)} \quad (11)$$

where $n_{m,N-1}$ is the number of variables in $z_{1:N-1}$, which belongs to component m, $n_{l,N-1}$ is the number of variables in $z_{1:N-1}$ belonging to component l and N is the total number of samples. For C components, the sum of (11) is:

$$p(z_N = m|z_{1:N-1}, \alpha) = \frac{\sum_{m=1}^{C} n_{m,N-1}}{(N + \alpha - 1)} = \frac{N-1}{(N + \alpha - 1)} \quad (12)$$

Hence, (12) is the probability of new data belonging to the existing components. Because (13) is a probability distribution, the summation should be one. The difference value between one and (13) is the probability of the new data belonging to a new component, which is expressed as:

$$p(z_N = C_{new}|z_{1:N-1}, \alpha) = \frac{\alpha}{(N + \alpha - 1)} \quad (13)$$

After the procedure of CRP is illustrated, the key point is estimation of the predictive distribution. In the proposed method, the Gaussian Inverse Wishart distribution [19] is selected as the prior distribution with parameter set $\boldsymbol{\phi}$, where $\boldsymbol{\phi} = \{\beta_0, \mu_0, \nu_0, \mathbf{S}_0\}$.

By using the CGS, the cluster parameters are integrated out and the latent variable is the only one that needs to be sampled [20]. Therefore, the equation of posterior prediction can be expressed as [11]:

$$p(z_N = m|z_{1:N-1}, x_{1:N}, \alpha, \boldsymbol{\phi}) \quad (14)$$
$$\propto p(z_N = m|z_{1:N-1}, \alpha)p(x_{1:N}|z_{1:N-1}, z_N = m, \boldsymbol{\phi})$$

Therefore, in the proposed method, the mean and variance parameters of each component are no longer needed in the process. In (14), $p(z_N = m|z_{1:N-1}, \alpha)$ can be known from (12) or (13), which is called the prior probability.

Then, the likelihood expression of the new data is expressed as :

$$\mathcal{L}_{new} = \mathcal{L}_{old} \times \frac{p(x_{N,m}|\boldsymbol{\phi})}{p(x_{1:N-1,m}|\boldsymbol{\phi})} \quad (15)$$

where $\mathcal{L}_{old}$ and $\mathcal{L}_{new}$ are the values of likelihood for the previous and current input data, respectively. $x_{1:N-1,m}$ represents the data points in $x_{1:N-1}$, which belongs to the component m. In (15), the expression for $p(x_{N,m}|\boldsymbol{\phi})$ is:

$$\int_{\mu} \int_{\Sigma} p(x_{N,m}, \mu, \Sigma|\boldsymbol{\phi})d\mu d\Sigma$$
$$= \int_{\mu} \int_{\Sigma} \prod\nolimits_{N|z_N=m} p(x_{N,m}|\mu, \Sigma)p(\mu, \Sigma|\boldsymbol{\phi})d\mu d\Sigma \quad (16)$$

where in (16), $p(x_{N,m}|\mu, \Sigma)$ is a Gaussian distribution and $p(\mu, \Sigma|\boldsymbol{\phi})$ is the Gaussian Inverse Wishart distribution.

Therefore, from [11] the new parameters are given as:

$$\mu_N = \frac{\beta_0 \mu_0 + N\overline{x}_{1:N}}{\beta_N} \quad (17)$$
$$\beta_N = \beta_0 + N \quad (18)$$
$$\nu_N = \nu_0 + N \quad (19)$$
$$\mathbf{S}_N = \mathbf{S}_0 + O + \frac{\beta_0 N}{\beta_N}(\overline{x}_{1:N} - \mu_0)(\overline{x}_{1:N} - \mu_0)^T \quad (20)$$
$$O = \sum_{i=1}^{N} (x_N - \overline{x}_{1:N})(x_N - \overline{x}_{1:N})^T \quad (21)$$

where $\overline{x}_{1:N}$ is the standard derivation of the data points. Then, by using the new parameters from (17) - (21), the posterior predictive probability $p(x_N|x_{1:N-1,m}, \boldsymbol{\phi})$ is expressed as:

$$\frac{p(x_{N,m}|\boldsymbol{\phi})}{p(x_{1:N-1,m}|\boldsymbol{\phi})}$$
$$= (\pi)^{\frac{-D}{2}} (\frac{\beta_N}{\beta_{N-1}})^{\frac{-D}{2}} \frac{|\mathbf{S}_{1:N}|^{\frac{-\nu_N}{2}}}{|\mathbf{S}_{1:N-1}|^{\frac{-\nu_{N-1}}{2}}} \frac{\Gamma(\frac{\nu_0+N}{2})}{\Gamma(\frac{\nu_0+N-D}{2})} \quad (22)$$

The prior in (14) is defined by (12) and (13), the likelihood in (15) is expressed by using (22). Therefore, the value of probability of the data point belonging to the existing component m can be expressed as:

$$\frac{n_{m,N-1}}{(N + \alpha - 1)} \times \mathcal{L}_{old} \times p(x_N|x_{1:N-1,m}, \boldsymbol{\phi}) \quad (23)$$

or a new component

$$\frac{\alpha}{(N + \alpha - 1)} \times \mathcal{L}_{old} \times p(x_N|x_{1:N-1,m}, \boldsymbol{\phi}) \quad (24)$$

Therefore, in (24), the value is for the new data cluster. If a new component is added, the number of components will increase, otherwise, the value of components remains unchanged. After all of the data points are clustered, the number of components in the mixture is confirmed. The components give different distributions which can be used to infer the number of sources in the speech mixtures.

4. EXPERIMENTAL RESULTS

In this section, the proposed method is evaluated with mixtures which are generated with real BRIRs [22]. In all of the experiments, speech signals are randomly selected from the whole of the TIMIT database [23] to generate the mixtures. Moreover, in order to confirm the proposed method is valid for all of the cases, the azimuth between sources is selected from 15° to 75° with the step size of 15° to set the physical separation as a variable. In these experiments, 20 pairs of mixtures are evaluated for each of the determined and underdetermined cases. In the proposed method, the CGS is utilized to extract data samplers.

To show the generalization ability of the proposed method, the real BRIRs from Hummersone [22] are exploited in the experiments. This dataset has four rooms with different $RT60$s, named A, B, C and D. Table 1 illustrates the parameters of these four rooms.

Table 1: Room settings for real BRIRs [22]

Room	Size	Dimension (m^3)	$RT60$ (s)
A	Medium	$5.7 \times 6.6 \times 2.3$	0.32
B	Small	$4.7 \times 4.7 \times 2.7$	0.47
C	Large	$23.5 \times 18.8 \times 4.6$	0.68
D	Medium	$8.0 \times 8.7 \times 4.3$	0.89

The experimental results for each room with different settings are shown in Figures 2 & 3. Besides, some mixtures are generated by five sources. One source is located at 0° azimuth. The remaining four sources are located symmetrically, e.g. 15° and -15°, 30° and -30°, between 15° to 75° azimuths. These experimental results are shown in Table 2.

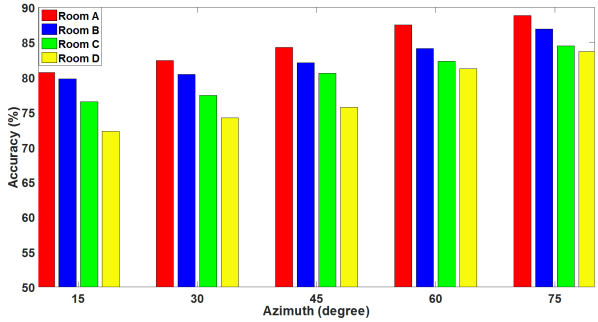

Fig. 2: Average accuracy of estimation from the mixtures which are generated with TIMIT database and the BRIRs [22] in different rooms and azimuths for two sources scenarios.

From Figures 2 & 3, we can observe that when the azimuths become larger, the average accuracies of clustering are improved. For example, it can be seen from Figure 2 that in room A, the performance of estimation accuracy is improved from around 80 % to around 87 % with the increase of azimuth from 15° to 75°.

In addition, the experimental results in Figures 2 & 3 confirm that the proposed method always performs better at lower

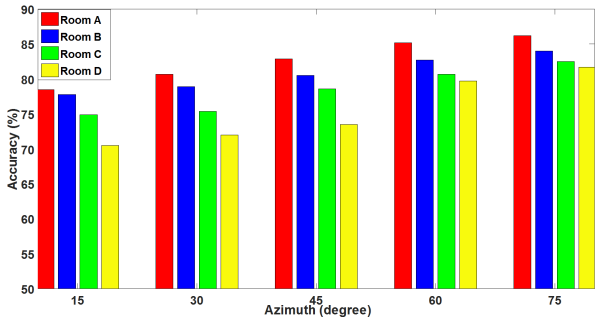

Fig. 3: Average accuracy of estimation from the mixtures which are generated with TIMIT database and the BRIRs [22] in different rooms and azimuths for three sources scenarios.

$RT60$s at all azimuths. And the average accuracy of the proposed method is inversely proportional to the number of sources in mixtures, e.g. the accuracies in Figure 2 are better than Figure 3, when compared at the same room environments and azimuths.

Table 2 shows that the average accuracies of the proposed method with five source scenarios in all cases are above 70 %.

Table 2: The average accuracies of the proposed method with mixtures generated by five sources scenarios and different azimuths for the real BRIRs [22].

Room	A	B	C	D
15°&30°	75.1 %	74.7 %	71.3 %	70.0 %
30°&45°	77.3 %	74.9 %	72.6 %	70.1 %
45°&60°	80.1 %	78.3 %	75.1 %	71.2 %
60°&75°	82.4 %	79.9 %	75.7 %	72.7 %

Because the algorithm with CGS requires less amount of parameters, the computational complexity and cost are reduced. Using the proposed method on a 3.5 GHz Intel Core i5, the average running time of the proposed method is 62.3 s while the running time with the Gibbs sampling exploited in [7] is 97.5 s, the improvement of the running time is around 36.1 %.

The above experimental results confirm that the proposed method has robust estimation performance when the room environments have high reverberation time and sources are also physically close to each other.

5. CONCLUSIONS

In this paper, we proposed a method to automatically determine the number of sources from the binaural speech mixtures. By using the DP and the CGS, the prior knowledge of possible maximum number of sources was no longer assumed and only the hyperparameters of the prior distribution were exploited. The proposed method also reduces the computational complexity while estimating the number of speakers from the speech mixtures, which are generated in a reverberant room environments.

6. REFERENCES

[1] C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, pp. 975–979, 1953.

[2] A. Hyvarinen and E. Oja, *Independent Component Analysis*, Wiley, 2001.

[3] S. M. Naqvi, M. Yu, and J. A. Chambers, "A Multimodal Approach to Blind Source Separation of Moving Sources," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, pp. 895–910, 2010.

[4] M. S. Salman, S. M. Naqvi, A. Rehman, W. Wang, and J. A. Chambers, "Video-Aided Model-Based Source Separation in Real Reverberant Rooms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1900–1912, 2013.

[5] 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," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

[6] J. Taghia and A. Leijon, "Separation of unknown number of sources," *IEEE Signal Processing Letters*, vol. 21, no. 5, pp. 625–629, 2014.

[7] O. Walter, L. Drude, and R. Haeb-Umbach, "Source counting in speech mixtures by nonparametric Bayesian estimation of an infinite Gaussian mixture model," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[8] T. Otsuka, K. Ishiguro, H. Sawada, and H. G. Okuno, "Bayesian Nonparametrics for Microphone Array Processing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 493–504, 2014.

[9] C. E. Antoniak, "Mixtures of Dirichlet process with applications to Bayesian nonparametric problems," *Annals of Statistics*, vol. 2, pp. 1152–1174, 1974.

[10] H. Ishwaran and L. F. James, "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, vol. 456, no. 96, pp. 161–168, 2001.

[11] R. Das, *Collapsed Gibbs Sampler for Dirichlet Process Gaussian Mixture Models (DPGMM)*, Technical report, Carnegie Mellon University , United States , 2014.

[12] C. Wang, Y. Chen, and K. J. R. Liu, "Sequential Chinese Restaurant Game," *IEEE Transactions on Signal Processing*, vol. 61, no. 3, pp. 571–584, 2013.

[13] B. Rivet, W. Wang, S. M. Naqvi, and J. A. Chambers, "Audiovisual speech source separation: An overview of key methodologies," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 125–134, 2014.

[14] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[15] T. Barker and T. Virtanen, "Blind Separation of Audio Mixtures Through Nonnegative Tensor Factorization of Modulation Spectrograms," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2377–2389, 2016.

[16] 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.

[17] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer, 2004.

[18] Y. Xu, *Non parametric Bayes and application to relational model*, University of Technology Sydney , Australia , 2015.

[19] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2009.

[20] J. Yin and J. Wang, "A model-based approach for text clustering with outlier detection," in *IEEE International Conference on Data Engineering (ICDE)*, 2016.

[21] Z. Y. Zohny, S. M. Naqvi, and J. A. Chambers, "Variational EM for clustering interaural phase cues in messl for blind source separation of speech," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[22] C. Hummersone, *Binaural Room Impulse Response Measurements*, Surrey University , United Kingdom , 2011.

[23] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, *DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM*, 1993.