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

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

Intelligent Sensing and Communications Group, School of Engineering, Newcastle University, NE1 7RU, UK.
Email: {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.

## Introduction

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.

- **Drawback 1:**

Limits the BSS application in real scenarios.

- **Drawback 2:**

In the underdetermined cases, some techniques are not applicable.

In order to overcome this limitation, some approaches have been proposed to determine the number of sources for BSS [1].

- **Solution 1:**

The multimodal (audio-video) information is exploited and the video modality is used to determine the number of speakers and assist the source separation process.

- **Solution 2:**

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, there still exist two main limitations in the above mentioned approaches:

- **Limitation 1:**

In some cases, only the speech mixtures are available, which limits the application of the multimodal methods.

- **Limitation 2:**

In full Bayesian inference methods, the maximum possible number of sources and full set of parameters need to be initialized.

**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.

The CGS method is used with the DP to obtain the samples from the joint distribution. Then, the CRP is exploited to cluster samples into different components and obtain the number of sources [2].

- **Advantage 1:**

Only the latent parameters related to the observed data and hyperparamters of the prior distribution are exploited

- **Advantage 2:**

The computational cost of this approach is also relatively low.

Hence, the proposed method can be used to estimate the number of sources from the speech mixtures which are generated in the reverberant environment.

## Proposed Method

**Model and Dirichlet Process Description:**

In the frequency domain, the Fourier transform of the left and right channels are $L(\omega,t)$ and $R(\omega,t)$, respectively [3].

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 $N(\omega,t)$ represents the Fourier transform of the noise and $H(\omega)$ is the ratio of Fourier transforms of the impulse responses [3].

The proposed approach is a model-based clustering method, which assumes the parameters of data points are generated by a mixture model.

In the Dirichlet process mixture model (DPMM), considering each data point $x_i$ is generated from a distribution defined by parameter $\theta_i$ and $\boldsymbol{\theta} = \{\theta_1, ..., \theta_N\}$.

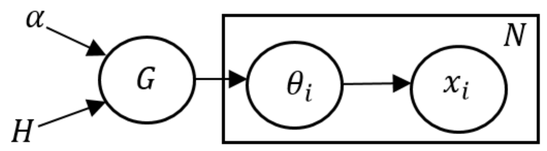
By using DP, the model can be described as:

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

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

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

where $\alpha$ is the concentration parameter to control the dispersion of the new distribution $G$, $H$ is the basement distribution and $\theta_i$ is generated from the new discrete distribution $G$. $F$ is a distribution, which generates $x_i$.




- For every observed data point $x_i$, there is a distribution and a parameter set $\theta_i$ which is generated from the $i$th distribution.
- $\boldsymbol{\theta}$ must have some identical factors, which indicates these parameters come from the same distribution.
- The latent variables are associated with these parameters and represent the clusters of the observed data.

**Collapsed Gibbs Sampling And Data Clustering:**

For the observed data point $x_N$, its cluster assignment $z_N$ is considered as the latent variable. 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 $\omega$ is the parameter set of this distribution. It can be obtained as:

$$p(x_N|x_{1:N-1}) = \int_\omega p(x_N|\omega)p(\omega|x_{1:N-1})d\omega \tag{5}$$

The predictive distribution of $p(\theta_N|\theta_{1:N-1})$ can be expressed similarly as:

$$p(\theta_N|\theta_{1:N-1}) = \int_G p(\theta_N|G)p(G|\theta_{1:N-1})dG \tag{6}$$

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

- 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:

$$p(z_N = m|z_{1:N-1}) = \frac{p(z_N = m, z_{1:N-1})}{p(z_{1:N-1})} \tag{7}$$

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 is the expression of the probability that the new data point $x_N$ belongs to the component $m$.
- The probability of new data belonging to the existing component is:

$$p(z_N = m|z_{1:N-1}, \alpha) = \frac{\sum_{m=1}^{C} n_{m,N-1}}{(N+\alpha-1)} = \frac{N-1}{(N+\alpha-1)} \tag{8}$$

- The probability of new data belonging to a new component is expressed as:

$$p(z_N = C_{new}|z_{1:N-1}, \alpha) = \frac{\alpha}{(N+\alpha-1)} \tag{9}$$

- The likelihood expression of the new data is expressed as :

$$\mathcal{L}_{new} = \mathcal{L}_{old} \times \frac{p(x_{N,m}|\boldsymbol{\phi})}{p(x_{1:N-1,m}|\boldsymbol{\phi})} \tag{10}$$

where $\boldsymbol{\phi} = \{\beta_0, \mu_0, \nu_0, \mathbf{S}_0\}$

- Value of the probability of the data point belonging to the existing component $m$ can be expressed as:

$$\frac{n_{m,N-1}}{(N+\alpha-1)} \times \mathcal{L}_{old} \times p(x_N|x_{1:N-1,m}, \boldsymbol{\phi}) \tag{11}$$

- Value of the probability of the data point belonging a new component can be expressed as:

$$\frac{\alpha}{(N+\alpha-1)} \times \mathcal{L}_{old} \times p(x_N|x_{1:N-1,m}, \boldsymbol{\phi}) \tag{12}$$

where $n_{m,N-1}$ is the number of variables in $z_{1:N-1}$, which belongs to component $m$.

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.

## Experimental Results

- **Mixtures:** Using the TIMIT database and real BRIRs.
- **Azimuth between sources:** 15 ° to 75 ° with the step size of 15 °.
- **Number of mixtures:** 20 pairs for each of the determined and underdetermined cases.
- **Room settings for the real BRIRs:**

| 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 |

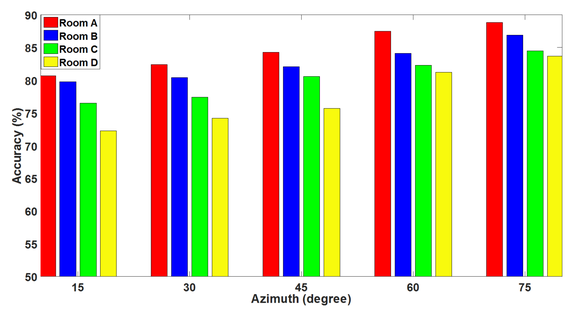



Figure 1: Average accuracy of estimation from the mixtures which are generated with TIMIT database and the BRIRs in different rooms and azimuths for two sources scenarios.

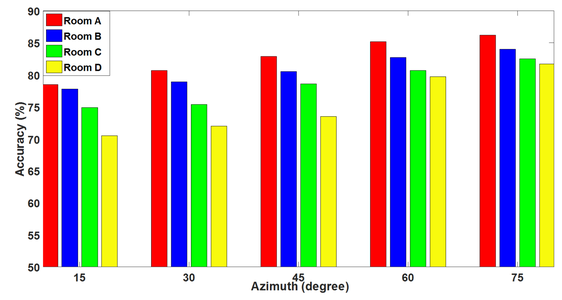



Figure 2: Average accuracy of estimation from the mixtures which are generated with TIMIT database and the BRIRs in different rooms and azimuths for three sources scenarios.

| 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 % |

Table 1: The average accuracies of the proposed method with mixtures generated by five sources scenarios and different azimuths for the real BRIRs.

**Discussions:**

- When the azimuths become larger, the average accuracies of clustering are improved.
- The average accuracy of the proposed method is inversely proportional to the number of sources in mixtures, when compared at the same room environments and azimuths.
- The average accuracies of the proposed method with five source scenarios in all cases are above 70 %.
- The proposed method has robust estimation performance when the room environments have high reverberation time and sources are also physically close to each other.

## Conclusions

The proposed method can 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.

## Selected references

[1] 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.

[2] 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.

[3] 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.