PERMUTATION-FREE CGMM: COMPLEX GAUSSIAN MIXTURE MODEL WITH INVERSE WISHART MIXTURE MODEL BASED SPATIAL PRIOR FOR PERMUTATION-FREE SOURCE SEPARATION AND SOURCE COUNTING

Juan Azcarreta, Nobutaka Ito, Shoko Araki, Tomohiro Nakatani

NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan
{ito.nobutaka, araki.shoko, nakatani.tomohiro}@lab.ntt.co.jp

ABSTRACT

Here we propose a *permutation-free cGMM* (PF-cGMM), a new probabilistic model of observed mixtures, which can resolve permutation ambiguity between frequency bins, and is applicable even when the number of sources is unknown. A recently proposed complex Gaussian mixture model (cGMM) is highly effective for frequency bin-wise clustering when the number of sources is known. However, it cannot resolve the permutation ambiguity, and is inapplicable when the number of sources is unknown. The proposed PF-cGMM is an extension of the cGMM, which resolves these issues. The resolution of the permutation ambiguity can be realized by a spatial prior called a *complex inverse Wishart mixture model* (cIWMM). The absence of the permutation ambiguity facilitates source counting, which is performed by hierarchical clustering in this paper. Experiments showed that the PF-cGMM was able to (1) resolve the permutation ambiguity and (2) realize source separation even when the number of sources was unknown with little performance degradation compared to when it was known.

Index Terms— Microphone array signal processing, source separation, time-frequency masks, permutation ambiguity, source counting.

1. INTRODUCTION

A source signal (*e.g.,* speech) is often *sparse* in the sense that only a small fraction of its time-frequency components capture a large fraction of its overall energy [1]. If observe signals are composed of such sparse source signals, they can be well approximated by only one dominant source signal at each time-frequency point. The dominant source signal is indicated by *masks*, *i.e.,* the presence probabilities of the source signals. Once obtained, these masks can be utilized for various array signal processing, such as source separation [1], denoising [2], and source localization [3]. For example, source separation is realized by retaining the time-frequency components dominated by the source signal of interest while suppressing the others. Recently, this mask-based approach has turned out to be highly effective, and was employed in the best-performing system [2, 4] in noise-robust automatic speech recognition (ASR) challenges CHiME-3 [5] and CHiME-4. In this approach, the accuracy of mask estimation is critical for the overall performance.

There are two main approaches to mask estimation, which have different advantages and disadvantages [6]: a deep neural network (DNN)-based approach [7–9] and a spatial clustering-based approach [1, 2, 10–13]. In the former, a DNN is trained in advance on training data so that it can estimate the masks from the observed spectral features. In the latter, on the other hand, the masks are estimated by unsupervised clustering of the observed spatial features. Here, we focus on the latter, which has the advantage over the former that it can more easily deal with mixtures of more than one source signals, and it is insensitive to mismatch between the training and the test conditions.

Among the mask estimation methods based on the latter approach, those based on time and level differences between microphones [1] are especially well known. Recently, *cGMM-based mask estimation* [14] has turned out to be highly effective, which was employed in the above best-performing system [2, 4] in CHiME-3 and CHiME-4. While the conventional cGMM is highly effective for frequency bin-wise clustering, it cannot resolve permutation ambiguity. That is, the cGMM cannot group together the bin-wise clusters corresponding to the same source in different frequency bins. This is a major obstacle in applying the cGMM to source separation for an unknown number of sources or in an online manner.

In contrast, the proposed PF-cGMM can resolve the permutation ambiguity based on a spatial prior called a *complex inverse Wishart mixture model* (cIWMM). The cIWMM does not require the prior knowledge of directions of arrival (DOA) of the sources, because it has source DOAs as hidden variables. Furthermore, the permutation-free nature of the PF-cGMM facilitates source counting, which is performed by hierarchical clustering in this paper. Consequently, the PF-cGMM realizes source separation even for an unknown number of sources.

The rest of the paper is organized as follows. Section 2 describes the signal model and the conventional cGMM. Section 3 describes the proposed PF-cGMM. Section 4 compares the source separation performance of the PF-cGMM and conventional methods including the cGMM. Finally, Section 5 concludes the paper.

2. BACKGROUND

2.1. Signal Model

Suppose we observe $N(\geq 2)$ concurrent speech signals using $M(\geq 2)$ microphones. In the short-time Fourier transform (STFT) domain the observed signals, $\boldsymbol{y}_{tf} \triangleq \begin{bmatrix} y_{tf}^{(1)} & \cdots & y_{tf}^{(M)} \end{bmatrix}^{\mathsf{T}}$, can be modeled as follows

$$\boldsymbol{y}_{tf} = \sum_{n=1}^{N} s_{tf}^{(n)} \boldsymbol{h}_{f}^{(n)} + \boldsymbol{v}_{tf}. \tag{1}$$

Here, $y_{tf}^{(m)}$ denotes the observed signal at the mth microphone, $t \in \{1, \ldots, T\}$ the frame index, $f \in \{1, \ldots, F\}$ the frequency bin index, T the number of frames, F the number of frequency bins up to the Nyquist frequency, $s_{tf}^{(n)}$ the STFT of the nth source signal,

$\boldsymbol{h}_f^{(n)} \triangleq \begin{bmatrix} h_f^{(1,n)} & \cdots & h_{tf}^{(M,n)} \end{bmatrix}^{\mathsf{T}}$, the time-invariant transfer function from the nth source to the microphones, and the superscript $^{\mathsf{T}}$ transposition. $\boldsymbol{v}_{tf} \triangleq \begin{bmatrix} v_{tf}^{(1)} & \cdots & v_{tf}^{(M)} \end{bmatrix}^{\mathsf{T}}$ denotes the background noise.

2.2. Complex Gaussian Mixture Model (cGMM)

We model the likelihood of the observation vector $\boldsymbol{y}_{tf}$ with a complex Gaussian mixture model (cGMM) [14]. The cGMM is defined by

$$p(\boldsymbol{y}_{tf}|\Theta) = \sum_{n=0}^{N} \alpha_f^{(n)} \mathcal{N}\Big(\boldsymbol{y}_{tf}; \boldsymbol{0}, \phi_{tf}^{(n)} \boldsymbol{\Sigma}_f^{(n)}\Big), \tag{2}$$

where $n = 0$ corresponds to the background noise $\boldsymbol{v}_{tf}$, and $n \in \{1, \ldots, N\}$ to a speaker [15]; $\mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the complex Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$:

$$\mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \triangleq \frac{1}{\det(\pi \boldsymbol{\Sigma})} \exp\big[-(\boldsymbol{y} - \boldsymbol{\mu})^{\mathsf{H}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu})\big], \tag{3}$$

where the superscript $^{\mathsf{H}}$ denotes Hermitian transposition. $\boldsymbol{\Sigma}_f^{(n)}$ denotes the spatial covariance matrix [16], and models the source location of the nth speaker (or the spatial characteristics of the background noise for $n = 0$). $\phi_{tf}^{(n)}$ models the spectrogram of the nth source signal.

Since the clustering is performed at each frequency bin independently, clusters with the same label n at different frequency bins may not necessarily correspond to the same source signal. This ambiguity is referred as the permutation ambiguity and makes it difficult to extend the cGMM to online processing. In the following section, we tackle this issue by introducing prior information over the spatial covariance matrix $\boldsymbol{\Sigma}_f^{(n)}$.

3. PROPOSED METHOD

3.1. Complex Inverse Wishart Mixture Model (cIWMM) for Spatial Prior

Let us introduce the complex inverse Wishart distribution

$$\mathcal{IW}(\boldsymbol{\Sigma}; \boldsymbol{\Psi}, \nu) \triangleq \frac{(\det \boldsymbol{\Psi})^{\nu}}{\pi^{M(M-1)/2} \prod_{m=1}^{M} \Gamma(\nu - m + 1)} \frac{\exp[-\mathrm{tr}(\boldsymbol{\Psi} \boldsymbol{\Sigma}^{-1})]}{(\det \boldsymbol{\Sigma})^{\nu + M}} \tag{4}$$

defined on the set of all Hermitian positive-definite matrices of size $M \times M$. $\boldsymbol{\Psi}_f^{(k)}$ denotes the scale matrix, and models a prior estimate of $\boldsymbol{\Sigma}_f^{(n)}$ when the speaker is at the kth potential source location. $\nu_f^{(k)}$ denotes the degrees of freedom, and models the degrees of deviation of $\boldsymbol{\Sigma}_f^{(n)}$ from $\boldsymbol{\Psi}_f^{(k)}$ when the speaker is at the kth potential source location. Both $\boldsymbol{\Psi}_f^{(k)}$ and $\nu_f^{(k)}$ are pre-trained on training data or computed theoretically under the planewave assumption, and constitute the probabilistic spatial dictionary.

The complex inverse Wishart distribution has been employed as a previous probability for the covariance matrix when the source DOAs or the room characteristics where known [17, 18]. In this work, we model the unknown DOAs as a PSD based on a complex inverse Wishart mixture model (cIWMM) over the spatial covariance matrix $\boldsymbol{\Sigma}_f^{(n)}$, for a speaker n as

$$p\Big(\big\{\boldsymbol{\Sigma}_f^{(n)}\big\}_f\Big) = \sum_{k=1}^{K} \beta^{(n,k)} \prod_{f=1}^{F} \mathcal{IW}\Big(\boldsymbol{\Sigma}_f^{(n)}; \boldsymbol{\Psi}_f^{(k)}, \nu_f^{(k)}\Big), \tag{5}$$
$$\forall n \in \{1, \ldots, N\}.$$

Here, k denotes the index of a potential speaker location, and K the number of potential speaker locations. $\beta^{(n,k)}$ denotes a mixture weight and models the prior probability of the nth speaker being at the kth potential speaker location.

As for the prior distribution over the spatial covariance matrix of the background noise, $\boldsymbol{\Sigma}_f^{(0)}$, we employ the following complex inverse Wishart distribution:

$$p\Big(\big\{\boldsymbol{\Sigma}_f^{(0)}\big\}_f\Big) = \prod_{f=1}^{F} \mathcal{IW}\Big(\boldsymbol{\Sigma}_f^{(0)}; \boldsymbol{\Psi}_f^{(0)}, \nu_f^{(0)}\Big). \tag{6}$$

Under the assumption of isotropic noise, we set $\boldsymbol{\Psi}_f^{(0)} \propto \boldsymbol{I}$, where $\boldsymbol{I}$ denotes the identity matrix. Alternatively, we can also set $\boldsymbol{\Psi}_f^{(0)} \propto \boldsymbol{\Gamma}_f$, where $\boldsymbol{\Gamma}_f$ denotes the spatial covariance matrix for the spherically or the cylindrically isotropic noise [19, 20].

The parameter set Θ now is explicitly defined as

$$\Theta \triangleq \big\{\alpha_f^{(n)}, \boldsymbol{\Sigma}_f^{(n)}, \beta^{(n,k)}\big\}, \tag{7}$$

where the range of n is $0 \leq n \leq N$ for $\alpha^{(n)}$ and $\boldsymbol{\Sigma}_f^{(n)}$ and $1 \leq n \leq N$ for $\beta^{(n,k)}$. We denote the likelihood of the observed data $\mathcal{Y}$ and the prior distribution over Θ by

$$p(\mathcal{Y}|\Theta) = \prod_{t=1}^{T} \prod_{f=1}^{F} p(\boldsymbol{y}_{tf}|\Theta), \tag{8}$$

$$p(\Theta) = \prod_{n=0}^{N} p\big(\boldsymbol{\Sigma}_1^{(n)}, \ldots, \boldsymbol{\Sigma}_F^{(n)}\big). \tag{9}$$

3.2. Objective Function: Posterior Probability

The parameters Θ are estimated so that the posterior probability $p(\Theta|\mathcal{Y})$ or equivalently the log-posterior probability $\ln p(\Theta|\mathcal{Y})$ is maximized. Introducing a weight γ to balance between the likelihood and the prior, we have the following objective function:

$$\begin{aligned} L(\Theta) &\triangleq \ln p(\mathcal{Y}|\Theta) + \gamma \ln p(\Theta) \\ &= \sum_{t=1}^{T} \sum_{f=1}^{F} \ln \Big[\sum_{n=0}^{N} \alpha_f^{(n)} \mathcal{N}\Big(\boldsymbol{y}_{tf}; \boldsymbol{0}, \phi_{tf}^{(n)} \boldsymbol{\Sigma}_f^{(n)}\Big)\Big] \\ &\quad + \gamma \sum_{n=1}^{N} \ln \Big[\sum_{k=1}^{K} \beta^{(n,k)} \prod_{f=1}^{F} \mathcal{IW}\Big(\boldsymbol{\Sigma}_f^{(n)}; \boldsymbol{\Psi}_f^{(k)}, \nu_f^{(k)}\Big)\Big] \\ &\quad + \gamma \sum_{f=1}^{F} \ln \mathcal{IW}\Big(\boldsymbol{\Sigma}_f^{(0)}; \boldsymbol{\Psi}_f^{(0)}, \nu_f^{(0)}\Big). \end{aligned} \tag{10}$$

3.3. Model Parameter Estimation Based on Majorization-Minimization

Since $L(\Theta)$ has no closed-form solution, we derive update equations for the parameters Θ using a majorization-minimization (MM)

technique. Let Φ be a set of auxiliary variables defined by $\Phi \triangleq \{\lambda_{tf}^{(n)}, \mu^{(n,k)}\}$ satisfying

$$\sum_{n=0}^{N} \lambda_{tf}^{(n)} = 1, \qquad \sum_{k=1}^{K} \mu^{(n,k)} = 1, \tag{11}$$

where the range of n is $1 \leq n \leq N$ for $\mu^{(n,k)}$. Applying Jensen's inequality, we can obtain an auxiliary function $Q(\Theta, \Phi)$ as a lower bound of $L(\Theta)$ and tangent to the current estimate of Θ [21]. In a MM approach, we estimate Θ by maximizing the auxiliary function $Q(\Theta, \Phi)$ rather than the actual function $L(\Theta)$. Therefore, we can estimate Θ by iterating the following two steps alternately, which increases $L(\Theta)$ monotonically.

3.3.1. Majorization Step

First, we update Φ by

$$\lambda_{tf}^{(n)} = \frac{\alpha_f^{(n)} \mathcal{N}\left(\boldsymbol{y}_{tf}; \boldsymbol{0}, \phi_{tf}^{(n)} \boldsymbol{\Sigma}_f^{(n)}\right)}{\sum_{\nu=0}^{N} \alpha_f^{(\nu)} \mathcal{N}\left(\boldsymbol{y}_{tf}; \boldsymbol{0}, \phi_{tf}^{(\nu)} \boldsymbol{\Sigma}_f^{(\nu)}\right)}, \tag{12}$$

$$\mu^{(n,k)} = \frac{\beta^{(n,k)} \prod_{f=1}^{F} \mathcal{IW}\left(\boldsymbol{\Sigma}_f^{(n)}; \boldsymbol{\Psi}_f^{(k)}, \nu_f^{(k)}\right)}{\sum_{l=1}^{K} \beta^{(n,l)} \prod_{f=1}^{F} \mathcal{IW}\left(\boldsymbol{\Sigma}_f^{(n)}; \boldsymbol{\Psi}_f^{(l)}, \nu_f^{(l)}\right)}, \tag{13}$$

so that $Q(\Theta, \Phi)$ is maximized.

3.3.2. Minimization Step

Then, we increase $Q(\Theta, \Phi)$ by updating Θ. Partial differentiation of $Q(\Theta, \Phi)$ with respect to Θ leads to the following update rules:

$$\alpha_f^{(n)} \leftarrow \frac{1}{T} \sum_{t=1}^{T} \lambda_{tf}^{(n)}, \tag{14}$$

$$\beta^{(n,k)} \leftarrow \mu^{(n,k)}, \tag{15}$$

$$\boldsymbol{\Sigma}_f^{(n)} \leftarrow \frac{\sum_{t=1}^{T} \lambda_{tf}^{(n)} \frac{1}{\phi_{tf}^{(n)}} \boldsymbol{y}_{tf} \boldsymbol{y}_{tf}^{\mathsf{H}} + \gamma \sum_{k=1}^{K} \mu^{(n,k)} \boldsymbol{\Psi}_f^{(k)}}{\sum_{t=1}^{T} \lambda_{tf}^{(n)} + \gamma \sum_{k=1}^{K} \mu^{(n,k)} \left(\nu_f^{(k)} + M\right)}, \tag{16}$$

$$\forall n \in \{1, \ldots, N\},$$

$$\boldsymbol{\Sigma}_f^{(0)} \leftarrow \frac{\sum_{t=1}^{T} \lambda_{tf}^{(0)} \frac{1}{\phi_{tf}^{(0)}} \boldsymbol{y}_{tf} \boldsymbol{y}_{tf}^{\mathsf{H}} + \gamma \boldsymbol{\Psi}_f^{(0)}}{\sum_{t=1}^{T} \lambda_{tf}^{(0)} + \gamma \left(\nu_f^{(0)} + M\right)}, \tag{17}$$

$$\phi_{tf}^{(n)} \leftarrow \frac{1}{M} \boldsymbol{y}_{tf}^{\mathsf{H}} \left(\boldsymbol{\Sigma}_f^{(n)}\right)^{-1} \boldsymbol{y}_{tf}. \tag{18}$$

3.3.3. Permutation Step

To avoid convergence to permuted solutions, after every MM iteration we permute the spatial covariance matrix at each frequency, so that $Q(\Theta, \Phi)$ is maximized as:

$$\Pi_f \leftarrow \arg\max_{\Pi} \sum_{n=1}^{N} \sum_{k=1}^{K} \beta^{(n,k)} \ln \mathcal{IW}\left(\boldsymbol{\Sigma}_f^{(\Pi(n))}; \boldsymbol{\Psi}_f^{(k)}, \nu_f^{(k)}\right), \tag{19}$$

where $\Pi : \{1, ..., C\} \rightarrow \{1, ..., C\}$ is a permutation on the set $\{1, ..., C\}$. The nondecrease of the likelihood function through this modified MM iteration is guaranteed.

3.4. Source Counting Based on Hierarchical Clustering

If the number of clusters L, is greater than N, the estimated DOAs defined by $\beta^{(n,k)}$ are grouped together forming N clusters along N distinct directions . Hence, by applying hierarchical clustering [22] over the estimated DOAs it is possible to calculate N.

Therefore, the final masks $\mathcal{M}_{tf}^{(n)}(n = 1, ..., N)$ can be obtained by merging the posterior probabilities, $\lambda_{tf}^{(l)}(l = 1, ..., L)$, which belong to the same group $J^{(n)}$, as follows

$$\mathcal{M}_{tf}^{(n)} \leftarrow \sum_{l \in J^{(n)}} \lambda_{tf}^{(l)}, \tag{20}$$

and $\mathcal{M}_{tf}^{(0)}$ is obtained by $\mathcal{M}_{tf}^{(0)} \leftarrow \lambda_{tf}^{(0)}$.

3.5. Discussion

We also proposed a permutation-free clustering method based on time-varying mixture weights [23]. This conventional method relies on the temporal activation pattern of each source signal to resolve permutation ambiguity. This method requires a certain duration of data to utilize the temporal information effectively, which may be an obstacle in online processing. In contrast, the proposed method relies on spatial information instead of temporal by modeling the spatial prior by a dictionary-based cIWMM. Therefore, the proposed method is more suited for online procesing, which is a remarkable advantage over the conventional method [23].

4. EXPERIMENTS

4.1. Experimental Conditions

We conducted simulations to demonstrate the effectiveness of the proposed PF-cGMM. In the first simulation (Sec. 4.2), we compared the PF-cGMM with three conventional models, namely, cGMM [14], cWMM [12, 13] and cBMM [24] in terms of source separation with known N. In the second simulation (Sec. 4.3), the PF-cGMM was tested for unknown N.

We generated observed signals by convolving 8 s-long clean speech signals with measured room impulse responses. These room impulse responses were measured in an experimental room using three microphones ($M = 3$) located as in Figure 1. We averaged the signal-to-distortion ratio (SDR) [25] for 16 trials with different combinations of speech signals and different distances between sources and the array centroid.

The observed signals were sampled at 8 kHz, the frame length was 1024 points (128 ms), the frame shift 256 points (32 ms), and the number of iterations 30. The hyperparameters of the cIWMM were set at $\nu_f^{(k)} = M + 1.1$ and $\boldsymbol{\Psi}_f^{(k)} = \left(\nu_f^{(k)} - M\right) \boldsymbol{h}_f^{(k)} \boldsymbol{h}_f^{(k)\mathsf{H}}$. The vector $\boldsymbol{h}_f^{(k)}$ denotes the steering vector for the direction k based on a planewave assumption. We set the strength of the prior, γ, at 5. The product in (13) was replaced by a sum for numerical stability.

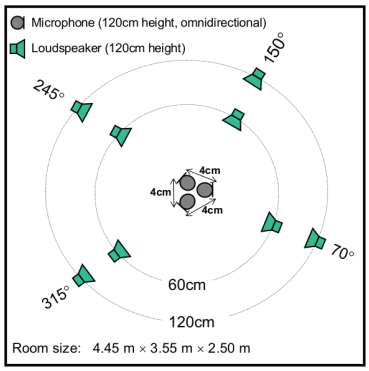

Fig. 1: Experimental setup

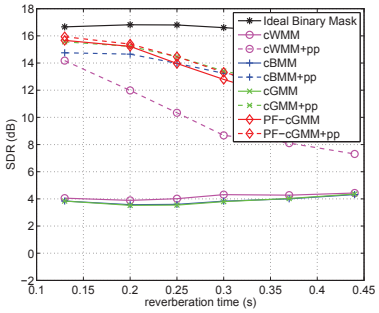

Fig. 2: Overdetermined case ($N = 2$, $M = 3$)

4.2. Source Separation Simulations

Figures 2–4 show SDR as a function of the reverberation time RT_{60}. The solid lines are results with postprocessing [13] for solving the permutation ambiguity, while the dashed lines are results without the postprocessing. The cWMM, the cBMM, the cGMM, and the PF-cGMM combined with the postprocessing are denoted by pcWMM, pcBMM, pcGMM, and pPF-cGMM, respectively. Regarding the SDR with the postprocessing, the pPF-cGMM and the pcGMM worked best, the pcBMM slightly worse, and the pcWMM significantly worse. While the SDR for the proposed PF-cGMM degraded little without the postprocessing, those for the conventional models degraded significantly. Consequently, the SDR without the postprocessing was much higher for the proposed PF-cGMM than for the conventional models. For high reverberation times, the PF-cGMM gave a smaller SDR than the pcGMM, which is probably because of the planewave assumption in the cIWMM. This can be compensated by designing the cIWMM based on the data measured in the test environment, or preprocessing the observed signals for dereverberation.

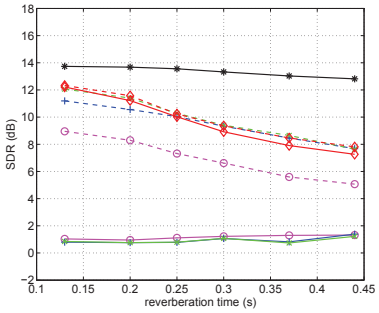

Fig. 3: Determined case ($N = 3$, $M = 3$)

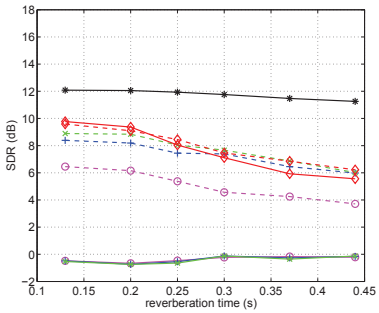

Fig. 4: Underdetermined case ($N = 4$, $M = 3$)

Table 1: Source counting accuracy of the proposed method. The ratio (%) of the trials with correct source counting to the total number of trials (16) is shown.

	reverberation time (ms)					
N	130	200	350	300	370	440
2	94%	75%	81%	75%	81%	88%
3	100%	94%	94%	94%	81%	94%
4	88%	94%	68%	75%	88%	81%

4.3. Source Counting Results

Table 1 shows the source counting accuracy of the proposed method, when N was unknown. The threshold for grouping the DOAs was set to 25 degrees. We set the number of clusters L at 6, which exceeded the number of sources $N = 2, 3, 4$. We see that the proposed method counted sources correctly with a high probability except for the most reverberant case.

5. CONCLUSIONS

We introduced the PF-cGMM, a cGMM with a spatial prior based on the cIWMM. Experiments showed that the PF-cGMM can perform source separation without causing the permutation ambiguity and estimate the number of sources accurately.

The future work includes an online implementation and application to the real-data processing, such as meetings.

6. REFERENCES

[1] Ö. Yılmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. SP*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[2] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. Fabian, M. Espi, T. Higuchi, S. Araki, and T. Nakatani, "The NTT CHiME-3 system: Advances

in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. ASRU*, Dec. 2015, pp. 436–443.

[3] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA estimation for multiple sparse sources with arbitrarily arranged multiple sensors," *Journal of Signal Processing Systems*, vol. 63, no. 3, pp. 265–275, June 2011.

[4] J. Du, Y.-H. Tu, L. Sun, F. Ma, H.-K. Wang, J. Pan, C. Liu, J.-D. Chen, and C.-H. Lee, "The USTC-iFlytek system for CHiME-4 challenge," in *Proc. CHiME2016*, 2016.

[5] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. ASRU*, Dec. 2015, pp. 504–511.

[6] T. Nakatani, N. Ito, T. Higuchi, S. Araki, and K. Kinoshita, "Integrating DNN-based and spatial clustering-based mask estimation for robust MVDR beamforming," in *Proc. ICASSP*, Mar. 2017.

[7] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. ASLP*, vol. 21, no. 7, pp. 1381–1390, July 2013.

[8] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. ICASSP*, Mar. 2016.

[9] H. Erdogan, J. Hershey, S. Watanabe, M. Mandel, and J. Le Roux, "Improved MVDR beamforming using single-channel mask prediction networks," in *Proc. Interspeech*, 2016.

[10] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 87, no. 8, pp. 1833–1847, Aug. 2007.

[11] M. Mandel, R. Weiss, and D. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. ASLP*, vol. 18, no. 2, pp. 382–394, Feb. 2010.

[12] D. H. Tran Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *Proc. ICASSP*, Mar. 2010, pp. 241–244.

[13] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. ASLP*, vol. 19, no. 3, pp. 516–527, Mar. 2011.

[14] N. Ito, S. Araki, T. Yoshioka, and T. Nakatani, "Relaxed disjointness based clustering for joint blind source separation and dereverberation," in *Proc. IWAENC*, Sept. 2014, pp. 268–272.

[15] N. Ito, S. Araki, M. Delcroix, and T. Nakatani, "Probabilistic spatial dictionary based online adaptive beamforming for meeting recognition in noisy and reverberant environments," in *Proc. ICASSP*, Mar. 2017.

[16] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. ASLP*, vol. 18, no. 7, pp. 1830–1840, Sep. 2010.

[17] T. Higuchi, N. Ito, S. Araki, T. Yoshioka, M. Delcroix, and T. Nakatani, "Online MVDR beamformer based on complex Gaussian mixture model with spatial prior for noise robust ASR," *IEEE Trans. ASLP*, vol. 25, no. 4, pp. 780–793, Apr. 2017.

[18] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Spatial location priors for Gaussian model based reverberant audio source separation," *EURASIP J. Adv. Signal Process.*, doi: 10.1186/1687-6180-2013-149, 2013.

[19] R. Cook, R. Waterhouse, R. Berendt, S. Edelman, and M. Thompson Jr., "Measurement of correlation coefficients in reverberant sound fields," *J. Acoust. Soc. Am.*, vol. 27, no. 6, pp. 1072–1077, Nov. 1955.

[20] G. Elko, "Spatial coherence functions for differential microphones in isotropic noise fields," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. Springer, 2001, pp. 61–85.

[21] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[22] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. New York: Springer, 2001.

[23] N. Ito, S. Araki, and T. Nakatani, "Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors," in *Proc. ICASSP*, May 2013, pp. 3238–3242.

[24] ——, "Modeling audio directional statistics using a complex Bingham mixture model for blind source extraction from diffuse noise," in *Proc. ICASSP*, Mar. 2016, pp. 465–468.

[25] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.