MODELING AUDIO DIRECTIONAL STATISTICS USING A COMPLEX BINGHAM MIXTURE MODEL FOR BLIND SOURCE EXTRACTION FROM DIFFUSE NOISE

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

Mask estimation is a central task in blind signal processing including source separation, denoising, and multi-source localization. In this paper, we define a complex Bingham mixture model (cBMM), and propose it as a model of directional statistics for mask estimation. The complex Bingham distribution can represent not only rotationally symmetric but also rotationally asymmetric distributions. Therefore, it can precisely model stochastic variation of the directional statistics due to reverberation, noise, source movement, *etc.*, which is not necessarily rotationally symmetric. In an experimental evaluation, the proposed cBMM outperformed a conventional complex Watson mixture model (cWMM) in terms of blind source extraction from diffuse noise, reducing the word error rate by 0.91% absolute on CHiME-3 challenge data.

Index Terms— Complex Bingham distribution, denoising, diffuse noise, clustering.

1. INTRODUCTION

Time-frequency clustering of observed signals is fundamental to blind acoustic signal processing including source separation [1–7], denoising [8, 9], and multi-source localization [10]. The clustering can be performed based on modeling the probability distribution of source location features with a mixture model. As such features, most conventional methods employ estimated time differences of arrival (TDOAs) between microphones. However, reverberation degrades the accuracy of TDOA estimation, which in turn degrades the performance of such methods.

To address the above issue, more recent methods [4–9] utilize directional statistics [11] as source location features. In this paper, the term "directional statistics" refers to a normalized M-dimensional observation vector in the short-time Fourier transform (STFT) domain, where M denotes the number of microphones. It has been shown experimentally that the directional statistics enable effective blind signal processing even in reverberant environments [5].

Conventionally, the directional statistics were modeled by a complex Watson mixture model (cWMM) [4] or its variant [5]. However, the complex Watson distribution cannot precisely model the stochastic variation of the directional statistics caused by reverberation, noise, source movement, *etc.* This is because the complex Watson distribution is rotationally symmetric about an axis, while such variation is rotationally asymmetric in general. Therefore, there is room to improve the clustering accuracy further by utilizing a model that can also represent rotationally asymmetric distributions.

In this paper, we define a complex Bingham mixture model (cBMM), and propose it as a model of directional statistics for mask estimation. The complex Bingham distribution can represent not only rotationally symmetric but also rotationally asymmetric distributions.

The rest of the paper is structured as follows. Section 2 formulates the blind signal enhancement problem including source separation and denoising. Section 3 reviews the conventional cWMM of directional statistics for blind signal enhancement, and Section 4 describes the proposed cBMM. Section 5 describes an experimental evaluation in terms of blind source extraction from diffuse noise. Section 6 concludes the paper.

2. PROBLEM FORMULATION

2.1. Blind Signal Enhancement

Suppose observed signals at $M(\geq 2)$ microphones are mixtures of $N(\geq 2)$ acoustic signals each of which is either a target signal or noise. In this paper, we consider *blind signal enhancement*, *i.e.*, estimation of each target signal using the observed signals only. We assume that N is known.

The above general problem includes the following special problems: (a) *source separation*, in which all acoustic signals are target signals; (b) *denoising*, in which one of the $N = 2$ acoustic signals is noise and the other is a target signal; (c) *joint source separation and denoising*, in which one of the $N \geq 3$ acoustic signals is noise and the others are target signals.

2.2. Our Focus: Mask Estimation

Under the assumption that each target signal is sparse in the time-frequency domain, all the above problems boil down to

mask estimation. This is true not only for the noiseless case (a) as per [5] but also for the noisy cases (b) and (c) as explained in the following.

Consider the noisy cases (b) and (c). We assume that, at each time-frequency point, the observed signals contain the noise and take one of the following N states

- $S^{(n)}\,(1 \leq n \leq N-1)$: the state in which the observed signals contain the nth target signal plus the noise,
- $S^{(N)}$: the state in which the observed signals contain the noise only.

Suppose we are given masks $\mathcal{M}_{tf}^{(n)}$ for $n = 1, \ldots, N$, where $\mathcal{M}_{tf}^{(n)}$ takes 1 at the time-frequency points (t, f) corresponding to the state $S^{(n)}$ and takes 0 elsewhere. Here, t denotes the frame index and f the frequency bin index. We can compute the covariance matrix of the nth target signal plus the noise using (t, f) with $\mathcal{M}_{tf}^{(n)} = 1\,(1 \leq n \leq N-1)$ and that of noise using (t, f) with $\mathcal{M}_{tf}^{(N)} = 1$. Subtraction of these covariance matrices yields the covariance matrix of the nth target signal, with which one can design a beamformer such as the multichannel Wiener filter [8] for signal enhancement. Therefore, (b) and (c) have boiled down to mask estimation.

In the remainder, we focus on mask estimation.

3. CONVENTIONAL METHOD

3.1. Mask Estimation by Clustering Directional Statistics

Mask estimation is usually performed by extracting at each time-frequency point, and clustering into N clusters, a feature vector which represent the direction of sound arrival. In such an approach, the signal enhancement performance depends on the clustering accuracy, which in turn depends on the design of the feature vector and of its model distribution.

Let us denote by $\boldsymbol{y}_{tf} \in \mathbb{C}^M$ the vector composed of the observed signals at the M microphones at (t, f). A conventional method [5] uses a feature vector defined by

$$\boldsymbol{z}_{tf} \triangleq \frac{\boldsymbol{y}_{tf}}{\|\boldsymbol{y}_{tf}\|}, \tag{1}$$

where $\|\cdot\|$ denotes the 2 norm. (1) lies on the unit hypersphere of $\mathbb{C}^M$. Such data in a spherical sample space cannot be treated properly by ordinary statistics for linear sample spaces, and call for *directional statistics* [11]. In this paper, we also refer to $\boldsymbol{z}_{tf}$ as directional statistics.

In the remainder, we mainly focus on how to model the directional statistics (1).

3.2. Complex Watson Mixture Model (cWMM) of Directional Statistics

The conventional method [5] models the probability distribution of the directional statistics $\boldsymbol{z}_{tf}$ at each f with a complex Watson mixture model (cWMM)

$$p(\boldsymbol{z}_{tf}; \Theta_f) = \sum_{n=1}^{N} \alpha_f^{(n)} \mathcal{W}\Big(\boldsymbol{z}_{tf}; \boldsymbol{a}_f^{(n)}, \kappa_f^{(n)}\Big). \tag{2}$$

Here, a complex Watson distribution [11]

$$\mathcal{W}\Big(\boldsymbol{z}_{tf}; \boldsymbol{a}_f^{(n)}, \kappa_f^{(n)}\Big) \propto \exp\Big[\kappa_f^{(n)} \big|\boldsymbol{a}_f^{(n)\mathsf{H}} \boldsymbol{z}_{tf}\big|^2\Big], \tag{3}$$

which is used in directional statistics and defined on the unit hypersphere, models the distribution of $\boldsymbol{z}_{tf}$ for each acoustic signal. Here, $^{\mathsf{H}}$ denotes Hermitian transposition. The unit vector $\boldsymbol{a}_f^{(n)}$ represents the location of the distribution of $\boldsymbol{z}_{tf}$ for each acoustic signal, $\kappa_f^{(n)}$ the concentration of the distribution, and the mixture weight $\alpha_f^{(n)}$ the height of the distribution. $\alpha_f^{(n)}$ satisfies $\sum_{n=1}^{N} \alpha_f^{(n)} = 1$ and $0 \leq \alpha_f^{(n)} \leq 1$. $\Theta_f \triangleq \Big\{\alpha_f^{(n)}, \boldsymbol{a}_f^{(n)}, \kappa_f^{(n)} \Big| {}^{\forall} n\Big\}$ denotes the set of the model parameters.

We fit (2) to the observed distribution of $\boldsymbol{z}_{tf}$ by estimating Θ_f by the maximum likelihood method or the maximum *a posteriori* method. Using estimated parameters $\hat{\Theta}_f \triangleq \Big\{\hat{\alpha}_f^{(n)}, \hat{\boldsymbol{a}}_f^{(n)}, \hat{\kappa}_f^{(n)} \Big| {}^{\forall} n\Big\}$, we obtain a mask estimate $\hat{\mathcal{M}}_{tf}^{(n)}$ as the posterior probability of (t, f) corresponding to each acoustic signal as follows:

$$\hat{\mathcal{M}}_{tf}^{(n)} = \frac{\hat{\alpha}_f^{(n)} \mathcal{W}\Big(\boldsymbol{z}_{tf}; \hat{\boldsymbol{a}}_f^{(n)}, \hat{\kappa}_f^{(n)}\Big)}{\sum_{k=1}^{N} \hat{\alpha}_f^{(k)} \mathcal{W}\Big(\boldsymbol{z}_{tf}; \hat{\boldsymbol{a}}_f^{(k)}, \hat{\kappa}_f^{(k)}\Big)}. \tag{4}$$

3.3. Drawback: Limited Expressiveness

(3) has the tight constraint of being rotationally symmetric about an axis $\boldsymbol{a}_f^{(n)}$. However, the distribution of $\boldsymbol{z}_{tf}$ for each acoustic signal is not necessarily rotationally symmetric, depending on the array geometry, the acoustic transfer characteristics, *etc.* Therefore, (2) may not be able to model the distribution accurately, which means that there is room for further improvement in mask estimation accuracy.

4. PROPOSED METHOD

4.1. Complex Bingham Mixture Model (cBMM) of Directional Statistics

To improve the mask estimation accuracy, we define a complex Bingham mixture model (cBMM) by

$$p(\boldsymbol{z}_{tf}; \Theta_f) = \sum_{n=1}^{N} \alpha_f^{(n)} \mathcal{B}\Big(\boldsymbol{z}_{tf}; \boldsymbol{B}_f^{(n)}\Big), \tag{5}$$

and propose to use it as the model distribution of the directional statistics. The element distribution

$$\mathcal{B}\Big(\boldsymbol{z}_{tf};\boldsymbol{B}_f^{(n)}\Big) \propto \exp\Big(\boldsymbol{z}_{tf}^{\mathsf{H}}\boldsymbol{B}_f^{(n)}\boldsymbol{z}_{tf}\Big) \tag{6}$$

is a complex Bingham distribution [11] used in directional statistics. The Hermitian matrix $\boldsymbol{B}_f^{(n)}$ represents not only the location and the concentration, but also the direction and the shape, of the distribution of $\boldsymbol{z}_{tf}$ corresponding to each acoustic signal. The set of parameters, Θ_f, is defined by $\Theta_f \triangleq \Big\{\alpha_f^{(n)}, \boldsymbol{B}_f^{(n)}\Big|^\forall n\Big\}$ this time.

We can easily confirm that the complex Watson distribution (3) is a special complex Bingham distribution (6) with a constraint

$$\boldsymbol{B}_f^{(n)} = \kappa_f^{(n)}\boldsymbol{a}_f^{(n)}\boldsymbol{a}_f^{(n)\mathsf{H}}. \tag{7}$$

In contrast, there is no constraint on $\boldsymbol{B}_f^{(n)}$ for the complex Bingham distribution except the hermiticity. Therefore, the complex Bingham distribution can represent various elliptically-shaped distributions on the hypersphere, and model the distribution of $\boldsymbol{z}_{tf}$ for each acoustic signal more precisely.

Since $\|\boldsymbol{z}_{tf}\| = 1$, $\mathcal{B}\Big(\boldsymbol{z}_{tf};\boldsymbol{B}_f^{(n)}\Big) = \mathcal{B}\Big(\boldsymbol{z}_{tf};\boldsymbol{B}_f^{(n)} + \xi\boldsymbol{I}\Big)$. Here, ξ is an arbitrary real number, and $\boldsymbol{I}$ the $M \times M$ identity matrix. Hereafter, we remove this indeterminacy by determining ξ so that the maximum eigenvalue of $\boldsymbol{B}_f^{(n)}$ becomes zero.

Suppose the eigenvalues of $\boldsymbol{B}_f^{(n)}$ are all distinct, which is always satisfied in practice. Then, the normalization constant for (6) is given by

$$c\Big(\boldsymbol{B}_f^{(n)}\Big) \triangleq 2\pi^M \sum_{i=1}^{M} \frac{\exp\Big(\lambda_f^{(n,i)}\Big)}{\prod_{j\neq i}\Big(\lambda_f^{(n,i)} - \lambda_f^{(n,j)}\Big)}, \tag{8}$$

so that

$$\mathcal{B}\Big(\boldsymbol{z}_{tf};\boldsymbol{B}_f^{(n)}\Big) = c\Big(\boldsymbol{B}_f^{(n)}\Big)^{-1}\exp\Big(\boldsymbol{z}_{tf}^{\mathsf{H}}\boldsymbol{B}_f^{(n)}\boldsymbol{z}_{tf}\Big). \tag{9}$$

Here, $\lambda_f^{(n,i)}\ (i = 1, \ldots, M)$ denote the eigenvalues of $\boldsymbol{B}_f^{(n)}$ with $\lambda_f^{(n,1)} < \cdots < \lambda_f^{(n,M)} = 0$.

4.2. Maximum-Likelihood Parameter Estimation Based on the Expectation-Maximization Algorithm

We estimate Θ_f by maximizing the log-likelihood function of the observed data given by

$$L(\Theta_f) = \sum_{t=1}^{T} \ln \sum_{n=1}^{N} \alpha_f^{(n)} \mathcal{B}\Big(\boldsymbol{z}_{tf};\boldsymbol{B}_f^{(n)}\Big). \tag{10}$$

T denotes the number of frames. We can derive an algorithm for optimizing (10) based on the expectation-maximization algorithm [12]. The auxiliary Q function is given by

$$Q\Big(\Theta_f;\Theta_f'\Big) = \sum_{t=1}^{T}\sum_{n=1}^{N} \mathcal{M}_{tf}^{(n)\prime} \ln\Big[\alpha_f^{(n)}\mathcal{B}\Big(\boldsymbol{z}_{tf};\boldsymbol{B}_f^{(n)}\Big)\Big], \tag{11}$$

where

$$\mathcal{M}_{tf}^{(n)\prime} \triangleq \frac{\alpha_f^{(n)\prime}\mathcal{B}\Big(\boldsymbol{z}_{tf};\boldsymbol{B}_f^{(n)\prime}\Big)}{\sum_{k=1}^{N}\alpha_f^{(k)\prime}\mathcal{B}\Big(\boldsymbol{z}_{tf};\boldsymbol{B}_f^{(k)\prime}\Big)}, \tag{12}$$

and $\Theta_f' \triangleq \Big\{\alpha_f^{(n)\prime}, \boldsymbol{B}_f^{(n)\prime}\Big|^\forall n\Big\}$ denotes the current estimate of Θ_f.

Substituting (9) in (11), we have

$$\begin{aligned} Q\Big(\Theta_f;\Theta_f'\Big) = \sum_{n=1}^{N}\Bigg(\sum_{t=1}^{T}\mathcal{M}_{tf}^{(n)\prime}\Bigg)\Bigg\{ &\ln\alpha_f^{(n)} \\ &- \ln c\Big(\boldsymbol{B}_f^{(n)}\Big) + \mathrm{tr}\Big[\boldsymbol{B}_f^{(n)}\boldsymbol{R}_f^{(n)}\Big]\Bigg\}, \end{aligned} \tag{13}$$

where $\boldsymbol{R}_f^{(n)} \triangleq \Big(\sum_{t=1}^{T}\mathcal{M}_{tf}^{(n)\prime}\boldsymbol{z}_{tf}\boldsymbol{z}_{tf}^{\mathsf{H}}\Big)\Big/\Big(\sum_{t=1}^{T}\mathcal{M}_{tf}^{(n)\prime}\Big)$. Applying the Lagrangian multiplier method with $\sum_{n=1}^{N}\alpha_f^{(n)} = 1$, we obtain the following update rule for $\alpha_f^{(n)}$

$$\alpha_f^{(n)} = \frac{1}{T}\sum_{t=1}^{T}\mathcal{M}_{tf}^{(n)\prime}. \tag{14}$$

In the following, we derive the update rule for $\boldsymbol{B}_f^{(n)}$. Let us assume that the eigenvalues of $\boldsymbol{R}_f^{(n)}$ are all positive and distinct as follows: $0 < l_f^{(n,1)} < \cdots < l_f^{(n,M)}$. Suppose the eigenvalue decompositions of $\boldsymbol{B}_f^{(n)}$ and $\boldsymbol{R}_f^{(n)}$ are given by $\boldsymbol{B}_f^{(n)} = \boldsymbol{U}_f^{(n)}\boldsymbol{\Lambda}_f^{(n)}\boldsymbol{U}_f^{(n)\mathsf{H}}$ and $\boldsymbol{R}_f^{(n)} = \boldsymbol{V}_f^{(n)}\boldsymbol{L}_f^{(n)}\boldsymbol{V}_f^{(n)\mathsf{H}}$. Then, as in [11],

$$\boldsymbol{U}_f^{(n)} = \boldsymbol{V}_f^{(n)}. \tag{15}$$

Setting the partial differentiation of (13) with respect to $\lambda_f^{(n,i)}$ to zero yields

$$\frac{\partial \ln c\Big(\boldsymbol{B}_f^{(n)}\Big)}{\partial\lambda_f^{(n,i)}} = l_f^{(n,i)}, i = 1, \ldots, M-1. \tag{16}$$

Under high concentration, (16) can be approximately solved as

$$\lambda_f^{(n,i)} \sim -\frac{1}{l_f^{(n,i)}}, i = 1, \ldots, M-1. \tag{17}$$

Note that $\lambda_f^{(n,M)} = 0$.

In the E step we compute (12). In the M step, we update Θ_f by (14), (15), and (17).

5. EXPERIMENTAL EVALUATION IN TERMS OF BLIND DIFFUSE NOISE REDUCTION

We evaluated the proposed cBMM applied to blind denoising (*i.e.*, the above case (b)) in terms of automatic speech recognition performance on the CHiME-3 corpus [13]. The CHiME-3 is a task of recognizing WSJ-5K prompts read from, and recorded by, a tablet device equipped with $M = 6$ microphones in four noisy public areas: on the bus (BUS), cafe (CAF), pedestrian area (PED), and street junction (STR). For further details about the data, we refer the readers to [13].

For both the proposed cBMM and the conventional cWMM, denoising was performed by using the multichannel Wiener filter [8] designed using the estimated masks. Under the assumption that the noise is diffuse, we modeled the noise cluster with a uniform distribution on the hypersphere (*i.e.*, a special cWMM with $\kappa_f^{(n)} = 0$, or a special cBMM with $\boldsymbol{B}_f^{(n)} = 0$). Based on the common amplitude modulation property of speech, we used time-dependent instead of frequency-dependent mixture weights [7]. The frame length and the frame shift were 64 ms and 16 ms, respectively, and the window was hann.

ASR was performed by a DNN-HMM system trained on 18 hours of multicondition data, where a fully-connected DNN with 10 hidden layers was used.

The word error rate (WER) for the real data of the development set, averaged over all environments, was as follows:

- no denoising: 14.29 %,
- denoising with the conventional cWMM: 9.28 %,
- denoising with the proposed cBMM: 8.37 %.

This result shows the effectiveness of the proposed method.

6. CONCLUSIONS

In this paper, we proposed to model directional statistics of multichannel audio signals using the cBMM. The method has been applied to blind source extraction from diffuse noise, and the superiority of the proposed method to the conventional method was demonstrated in the ASR experiment on the CHiME-3 dataset.

The future work includes the evaluation of the proposed method in the case with multiple, and possibly an unknown number of, sources.

7. 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] 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.

[3] Y. Izumi, N. Ono, and S. Sagayama, "Sparseness-based 2ch BSS using the EM algorithm in reverberant environment," in *Proc. WASPAA*, Oct. 2007, pp. 147–150.

[4] 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.

[5] 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.

[6] J. Taghia, N. Mohammadiha, and A. Leijon, "A variational Bayes approach to the underdetermined blind source separation with automatic determination of the number of sources," in *Proc. ICASSP*, Mar. 2012, pp. 253–256.

[7] 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.

[8] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel MMSE-based framework for speech source separation and noise reduction," *IEEE Trans. ASLP*, vol. 21, no. 9, pp. 1913–1928, Sept. 2013.

[9] T. Nakatani, S. Araki, T. Yoshioka, M. Delcroix, and M. Fujimoto, "Dominance based integration of spatial and spectral features for speech enhancement," *IEEE Trans. ASLP*, vol. 21, no. 12, pp. 2516–2531, Dec. 2013.

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

[11] K.V. Mardia and P.E. Jupp, *Directional Statistics*, John Wiley & Sons, West Sussex, 2000.

[12] A.P. Dempster, N.M. Laird, and D.B. 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.

[13] 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.