COMPLEX ANGULAR CENTRAL GAUSSIAN MIXTURE MODEL FOR DIRECTIONAL STATISTICS IN MASK-BASED MICROPHONE ARRAY SIGNAL PROCESSING

Nobutaka Ito, Shoko Araki, and Tomohiro Nakatani

NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seika-cho, "Keihanna Science City" Kyoto 619-0237 Japan

ABSTRACT

Microphone array signal processing based on time-frequency masks has been applied successfully to various tasks including source separation, denoising, source localization, and source counting. Aiming to improve the performance of these techniques, here we propose a mask estimation method based on a *complex Angular Central Gaussian Mixture Model (cACGMM)* for multichannel observed signals. Compared to a conventional *complex Watson Mixture Model (cWMM)*, the proposed cACGMM can model not only rotationally symmetrical but also elliptical distributions. Therefore, the cACGMM can better approximate the distribution of observed data, which is generally not rotationally symmetrical. In source separation simulations with real recorded impulse responses, the cACGMM resulted in an average 1.2 dB improvement of the Signal-to-Distortion Ratio (SDR) over the cWMM.

Index Terms— Time-frequency masks, microphone array signal processing, complex angular central Gaussian distributions.

I. INTRODUCTION

Array signal processing based on masks has been applied to various tasks. These tasks include source separation [1]–[13], denoising [10], [13]–[16], source localization [5], [17], and source counting [4], [9], [12], [18]. Real-world evaluations have demonstrated effectiveness and robustness of this approach [7], [19].

The approach is based on the *sparseness* assumption that the observed signals contain at most one source signal at each time-frequency point [1]. This assumption reduces the above tasks to the estimation of masks, which indicate which signal component is present at each time-frequency point. The masks can be estimated by clustering source location features such as time difference and level difference between microphones.

In this paper, we employ *directional statistics* as the source location features. Real-world experiments have shown that the directional statistics enable accurate mask estimation even in reverberant environments [7]. To define the directional statistics, let us denote by $\boldsymbol{y}_{tf} \in \mathbb{C}^M$ the vector composed of the observed signals at all microphones in the short-time Fourier transform domain, where t denotes the frame index, f the frequency bin index, and M the number of microphones. The directional statistics are defined as the unit vector [7]

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

where $\|\cdot\|$ denotes the Euclidean norm. The directional statistics (1) lie on the unit hypersphere centered at the origin, and form clusters corresponding to signal components.

Conventionally, the directional statistics (1) have been modeled by a *complex Watson Mixture Model (cWMM)* [6], [7]. The cWMM is composed of *complex Watson distributions*, which are rotationally symmetrical. Therefore, the cWMM has a limited ability to approximate the distribution of the directional statistics, which is generally not rotationally symmetrical, depending on various conditions such as the array geometry and acoustic transfer characteristics.

To overcome this limitation, here we propose a *complex Angular Central Gaussian Mixture Model (cACGMM)* for the directional statistics (1). The cACGMM is composed of *complex Angular Central Gaussian (cACG) distributions* [20], which can represent not only rotationally symmetrical but also elliptical distributions. Therefore, the cACGMM can better approximate the distribution of the directional statistics. We present an Expectation-Maximization (EM) algorithm for estimating the masks based on the cACGMM.

The rest of this paper is organized as follows. Section II describes mask-based microphone array signal processing, and reviews the conventional cWMM for estimating the masks. Section III describes the proposed cACGMM and the EM algorithm. Section 4 describes source separation simulations, and Section 5 concludes the paper.

II. BACKGROUND

II-A. Mask-based Microphone Array Signal Processing

Here we formulate blind source separation based on time-frequency masks. See the references for noise reduction [10], [14], [16], multisource localization [5], [17], and source enumeration [4], [12], [18].

Suppose we observe mixtures of $N(\geq 2)$ source signals with $M(\geq 2)$ microphones with N known. The observed

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

signals $\boldsymbol{y}_{tf}$ can be modeled by the sum of the N signal components:

$$\boldsymbol{y}_{tf} = \sum_{n=1}^{N} \boldsymbol{s}_{tf}^{(n)}, \tag{2}$$

where $\boldsymbol{s}_{tf}^{(n)} \in \mathbb{C}^M$ denotes the nth signal component. The sparseness reduces (2) to

$$\boldsymbol{y}_{tf} = \boldsymbol{s}_{tf}^{(\nu)} \quad \text{with} \quad \nu = d_{tf}, \tag{3}$$

where d_{tf} denotes the index of the source signal present at (t, f).

Source separation aims to estimate $\boldsymbol{s}_{tf}^{(n)}$ from $\boldsymbol{y}_{tf}$. Given the time-frequency mask $\gamma_{tf}^{(n)}$, we can estimate $\boldsymbol{s}_{tf}^{(n)}$ by, *e.g.*,

$$\hat{\boldsymbol{s}}_{tf}^{(n)} = \gamma_{tf}^{(n)} \boldsymbol{y}_{tf}. \tag{4}$$

$\gamma_{tf}^{(n)}$ is defined as the posterior probability of $d_{tf} = n$ given $\boldsymbol{y}_{tf}$. $\gamma_{tf}^{(n)}$ satisfies $0 \leq \gamma_{tf}^{(n)} \leq 1$ and $\sum_{n=1}^{N} \gamma_{tf}^{(n)} = 1$.

II-B. Conventional Complex Watson Mixture Model

Sawada *et al.* [7] and Tran Vu *et al.* [6] proposed to model the directional statistics (1) by a *complex Watson Mixture Model (cWMM)*

$$p(\boldsymbol{z}_{tf}; \Theta_f) = \sum_{k=1}^{K} \alpha_f^{(k)} \mathcal{W}\Big(\boldsymbol{z}_{tf}; \boldsymbol{a}_f^{(k)}, \kappa_f^{(k)}\Big) \tag{5}$$

defined on the unit hypersphere

$$\mathcal{S} \triangleq \Big\{ \boldsymbol{z} \in \mathbb{C}^M \Big| \|\boldsymbol{z}\| = 1 \Big\}. \tag{6}$$

k denotes the mixture component index; K the number of mixture components (*e.g.*, $K = N$ for source separation); $\alpha_f^{(k)}$ a mixture weight satisfying $0 \leq \alpha_f^{(k)} \leq 1$ and $\sum_{k=1}^{K} \alpha_f^{(k)} = 1$; $\Theta_f \triangleq \Big\{ \alpha_f^{(k)}, \boldsymbol{a}_f^{(k)}, \kappa_f^{(k)} \Big| {}^{\forall} k \Big\}$ the set of the model parameters.

$\mathcal{W}(\boldsymbol{z}; \boldsymbol{a}, \kappa)$ denotes the complex Watson distribution [21] with mean orientation $\boldsymbol{a} \in \mathcal{S}$ and concentration $\kappa \geq 0$:

$$\mathcal{W}(\boldsymbol{z}; \boldsymbol{a}, \kappa) \triangleq \frac{(M-1)!}{2\pi^M \mathcal{K}(1, M; \kappa)} \mathrm{e}^{\kappa |\boldsymbol{a}^{\mathsf{H}} \boldsymbol{z}|^2} \tag{7}$$

defined on $\mathcal{S}$. It models the distribution of $\boldsymbol{z}_{tf}$ for each signal component in the mixture model (5). $\boldsymbol{a}$ indicates the mode location of (7). Indeed, for $\kappa > 0$, (7) attains a maximum when $\boldsymbol{z} \parallel \boldsymbol{a}$, *i.e.*, when $\boldsymbol{z} = e^{\mathrm{j}\theta} \boldsymbol{a}$ for some $\theta \in \mathbb{R}$. κ controls the concentration of the distribution (7) at the mode. For $\kappa = 0$, (7) reduces to the uniform distribution on $\mathcal{S}$ given by $\mathcal{U}(\boldsymbol{z}) \triangleq \frac{(M-1)!}{2\pi^M}$. $\mathcal{K}(a, b; x)$ denotes the Kummer function; $^{\mathsf{H}}$ conjugate transposition.

Θ_f can be estimated by, *e.g.*, the maximum likelihood method. $\gamma_{tf}^{(k)}$ can be estimated using Θ_f by

$$\gamma_{tf}^{(k)} = \frac{\alpha_f^{(k)} \mathcal{W}\Big(\boldsymbol{z}_{tf}; \boldsymbol{a}_f^{(k)}, \kappa_f^{(k)}\Big)}{\sum_{\kappa=1}^{K} \alpha_f^{(\kappa)} \mathcal{W}\Big(\boldsymbol{z}_{tf}; \boldsymbol{a}_f^{(\kappa)}, \kappa_f^{(\kappa)}\Big)}. \tag{8}$$

Since Θ_f and $\gamma_{tf}^{(k)}$ are computed in each frequency bin independently, it generally varies from frequency bin to frequency bin which cluster k corresponds to which signal component n. This is called a permutation problem. To realize source separation, it is necessary to relabel the clusters so that each cluster k corresponds to the same signal component n in all frequency bins. In [7], this is done based on correlation of $\gamma_{tf}^{(k)}$ between frequency bins.

The complex Watson distribution (5) is rotationally symmetrical about the axis $\boldsymbol{a}$. However, as already pointed out, the distribution of $\boldsymbol{z}_{tf}$ for each signal component is not necessarily rotationally symmetrical.

III. PROPOSED METHOD

III-A. Complex Angular Central Gaussian Mixture Model

We propose to model the directional statistics by a *complex Angular Central Gaussian Mixture Model (cACGMM)*

$$p(\boldsymbol{z}_{tf}; \Theta_f) = \sum_{k=1}^{K} \alpha_f^{(k)} \mathcal{A}\Big(\boldsymbol{z}_{tf}; \boldsymbol{B}_f^{(k)}\Big) \tag{9}$$

defined on the hypersphere $\mathcal{S}$. Here we define Θ_f by $\Theta_f \triangleq \Big\{ \alpha_f^{(k)}, \boldsymbol{B}_f^{(k)} \Big| {}^{\forall} k \Big\}$.

$\mathcal{A}(\boldsymbol{z}; \boldsymbol{B})$ denotes a *complex Angular Central Gaussian (cACG) distribution* [20]

$$\mathcal{A}(\boldsymbol{z}; \boldsymbol{B}) \triangleq \frac{(M-1)!}{2\pi^M \det(\boldsymbol{B})} \frac{1}{(\boldsymbol{z}^{\mathsf{H}} \boldsymbol{B}^{-1} \boldsymbol{z})^M} \tag{10}$$

defined on $\mathcal{S}$. It models the distribution of $\boldsymbol{z}_{tf}$ for each signal component in the mixture model (9). The positive-definite Hermitian matrix $\boldsymbol{B}$ controls not only the mode location and the concentration, but also the rotation and the shape, of the distribution (10). (10) has a global maximum when $\boldsymbol{z}$ is a principal eigenvector of $\boldsymbol{B}$.

III-B. Expectation-Maximization Algorithm for Mask Estimation

Θ_f is estimated by maximizing the log-likelihood function

$$L(\Theta_f) = \sum_{t=1}^{T} \ln \sum_{k=1}^{K} \alpha_f^{(k)} \mathcal{A}\Big(\boldsymbol{z}_{tf}; \boldsymbol{B}_f^{(k)}\Big). \tag{11}$$

T denotes the number of frames. One can derive an EM algorithm [22], which iterates the following update rules

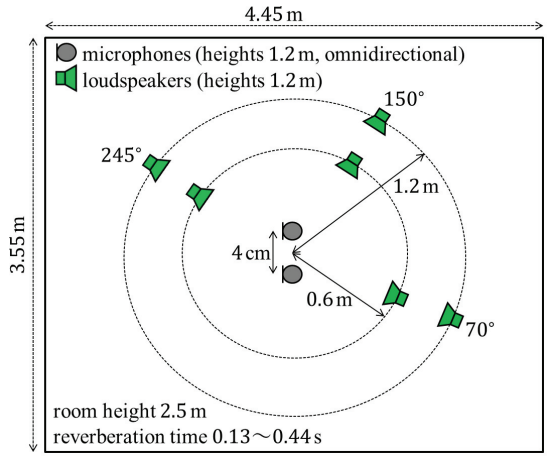


Fig. 1. Configurations in room impulse response measurement.

alternately:

$$\gamma_{tf}^{(k)} = \frac{\alpha_f^{(k)} \dfrac{1}{\det\left(\boldsymbol{B}_f^{(k)}\right)} \dfrac{1}{\left[\boldsymbol{z}_{tf}^{\mathsf{H}} \left(\boldsymbol{B}_f^{(k)}\right)^{-1} \boldsymbol{z}_{tf}\right]^M}}{\displaystyle\sum_{\kappa=1}^{K} \alpha_f^{(\kappa)} \frac{1}{\det\left(\boldsymbol{B}_f^{(\kappa)}\right)} \frac{1}{\left[\boldsymbol{z}_{tf}^{\mathsf{H}} \left(\boldsymbol{B}_f^{(\kappa)}\right)^{-1} \boldsymbol{z}_{tf}\right]^M}}, \quad (12)$$

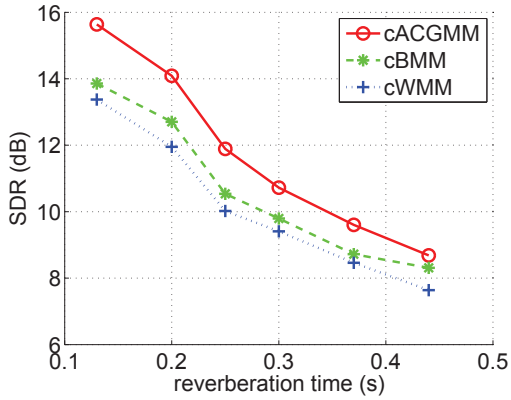
$$\alpha_f^{(k)} = \frac{1}{T} \sum_{t=1}^{T} \gamma_{tf}^{(k)}, \quad (13)$$

$$\boldsymbol{B}_f^{(k)} = M \frac{\displaystyle\sum_{t=1}^{T} \gamma_{tf}^{(k)} \frac{\boldsymbol{z}_{tf} \boldsymbol{z}_{tf}^{\mathsf{H}}}{\boldsymbol{z}_{tf}^{\mathsf{H}} \left(\boldsymbol{B}_f^{(k)}\right)^{-1} \boldsymbol{z}_{tf}}}{\displaystyle\sum_{t=1}^{T} \gamma_{tf}^{(k)}}. \quad (14)$$

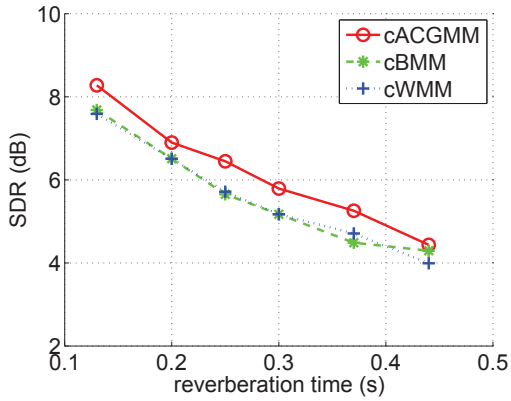
In Appendix, we describe a close relation between the proposed cACGMM and a recently proposed *Time-Varying complex Gaussian Mixture Model (TV-cGMM)* [11]. Real-world experiments demonstrated the effectiveness of the TV-cGMM for source separation [13] and noise reduction [15], [19]. Especially, the TV-cGMM was employed in our CHiME-3 challenge system [19], which won the first place [23].

IV. SOURCE SEPARATION SIMULATIONS

We conducted simulations to compare the proposed cACGMM with two conventional models, namely, the cWMM and a *complex Bingham Mixture Model (cBMM)* [16] in terms of source separation with known N. We generated observed signals by convolving 8s-long English speech signals with room impulse responses measured in an experimental room (see Fig. 1). The sampling frequency of the observed signals was 8 kHz; the frame length 1024 points (128 ms); the frame shift 256 points (32 ms); the number of EM iterations 100. The permutation problem was resolved by Sawada's method [7]. The source signals were estimated using the estimated masks by (4).



(a) Determined case ($N = 2$).



(b) Underdetermined case ($N = 3$).

Fig. 2. Signal-to-Distortion Ratio (SDR) as a function of the reverberation time RT_{60}. The figure shows SDRs averaged over 16 trials with eight combinations of speech signals and two distances between the loudspeaker and the array center. The azimuths of sources were 70° and 150° for $N = 2$, and 70°, 150°, and 245° for $N = 3$.

Fig 2 shows the Signal-to-Distortion Ratio (SDR) [24] as a function of the reverberation time RT_{60}. The cACGMM outperformed the other models consistently. It improved the SDR by an average of 1.2 dB compared to the cWMM, and 0.9 dB compared to the cBMM. The superiority of the cACGMM to the cWMM is attributed to its ability to model elliptical distributions.

V. CONCLUSIONS

We proposed the cACGMM for modeling the directional statistics. In simulations, the cACGMM outperformed the cBMM and the cWMM in terms of source separation.

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] S. Araki, T. Nakatani, H. Sawada, and S. Makino, "Blind sparse source separation for unknown number of sources using Gaussian mixture model fitting with Dirichlet prior," in *Proc. ICASSP*, Apr. 2009, pp. 33–36.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] N. Ito, S. Araki, and T. Nakatani, "Permutation-free clustering of relative transfer function features for blind source separation," in *Proc. EUSIPCO*, Sept. 2015, pp. 409–413.

[13] S. Araki, M. Okada, and T. Nakatani, "Spatial correlation model based observation vector clustering and MVDR beamforming for meeting recognition," in *Proc. ICASSP*, Mar. 2016, pp. 385–389.

[14] 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.

[15] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *Proc. ICASSP*, Mar. 2016, pp. 5210–5214.

[16] N. Ito, S. Araki, and T. Nakatani, "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.

[17] 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.

[18] L. Drude, A. Chinaev, D.H. Tran Vu, and R. Haeb-Umbach, "Source counting in speech mixtures using a variational EM approach for complex Watson mixture models," in *Proc. ICASSP*, May 2014, pp. 6834–6838.

[19] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W.J. 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.

[20] J.T. Kent, "Data analysis for shapes and images," *Journal of Statistical Planning and Inference*, vol. 57, no. 2, pp. 181–193, Feb. 1997.

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

APPENDIX

The cACGMM is closely related to a *Time-Varying complex Gaussian Mixture Model (TV-cGMM)* [11]. The TV-cGMM is defined in $\mathbb{C}^M$ by

$$p(\boldsymbol{y}_{tf};\Theta_f)=\sum_{k=1}^{K}\alpha_f^{(k)}\mathcal{N}\Big(\boldsymbol{y}_{tf};0,\phi_{tf}^{(k)}\boldsymbol{B}_f^{(k)}\Big). \tag{15}$$

$\boldsymbol{y}_{tf}$ denotes the observed signals; $\mathcal{N}(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma})$ the complex Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$:

$$\mathcal{N}(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma})\triangleq\frac{1}{\det(\pi\boldsymbol{\Sigma})}\mathrm{e}^{-(\boldsymbol{y}-\boldsymbol{\mu})^{\mathsf{H}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}. \tag{16}$$

$\phi_{tf}^{(k)}$ models the source spectrogram; $\boldsymbol{B}_f^{(k)}$ denotes a spatial covariance matrix [25]; Θ_f denotes the parameter set.

The EM algorithm for the TV-cGMM is given by [11], [13], [15]

$$\gamma_{tf}^{(k)}=\frac{\alpha_f^{(k)}\mathcal{N}\Big(\boldsymbol{y}_{tf};0,\phi_{tf}^{(k)}\boldsymbol{B}_f^{(k)}\Big)}{\sum_{\kappa=1}^{K}\alpha_f^{(\kappa)}\mathcal{N}\Big(\boldsymbol{y}_{tf};0,\phi_{tf}^{(\kappa)}\boldsymbol{B}_f^{(\kappa)}\Big)}, \tag{17}$$

$$\alpha_f^{(k)}=\frac{1}{T}\sum_{t=1}^{T}\gamma_{tf}^{(k)}, \tag{18}$$

$$\boldsymbol{B}_f^{(k)}=\frac{\displaystyle\sum_{t=1}^{T}\gamma_{tf}^{(k)}\frac{1}{\phi_{tf}^{(k)}}\boldsymbol{y}_{tf}\boldsymbol{y}_{tf}^{\mathsf{H}}}{\displaystyle\sum_{t=1}^{T}\gamma_{tf}^{(k)}}, \tag{19}$$

$$\phi_{tf}^{(k)}=\frac{1}{M}\boldsymbol{y}_{tf}^{\mathsf{H}}\Big(\boldsymbol{B}_f^{(k)}\Big)^{-1}\boldsymbol{y}_{tf}. \tag{20}$$

We can show that the EM algorithm (17)–(20) for the TV-cGMM is equivalent to the EM algorithm (12)–(14) for the cACGMM. To show this, we plug in (20) to (17) and (19), which yields

$$\gamma_{tf}^{(k)}=\frac{\alpha_f^{(k)}\dfrac{1}{\det\Big(\boldsymbol{B}_f^{(k)}\Big)}\dfrac{1}{\Big[\boldsymbol{y}_{tf}^{\mathsf{H}}\Big(\boldsymbol{B}_f^{(k)}\Big)^{-1}\boldsymbol{y}_{tf}\Big]^{M}}}{\displaystyle\sum_{\kappa=1}^{K}\alpha_f^{(\kappa)}\frac{1}{\det\Big(\boldsymbol{B}_f^{(\kappa)}\Big)}\frac{1}{\Big[\boldsymbol{y}_{tf}^{\mathsf{H}}\Big(\boldsymbol{B}_f^{(\kappa)}\Big)^{-1}\boldsymbol{y}_{tf}\Big]^{M}}}, \tag{21}$$

$$\boldsymbol{B}_f^{(k)}=M\frac{\displaystyle\sum_{t=1}^{T}\gamma_{tf}^{(k)}\frac{\boldsymbol{y}_{tf}\boldsymbol{y}_{tf}^{\mathsf{H}}}{\boldsymbol{y}_{tf}^{\mathsf{H}}\Big(\boldsymbol{B}_f^{(k)}\Big)^{-1}\boldsymbol{y}_{tf}}}{\displaystyle\sum_{t=1}^{T}\gamma_{tf}^{(k)}}. \tag{22}$$

Since $\boldsymbol{y}_{tf}\propto\boldsymbol{z}_{tf}$, (21) coincides with (12), and (22) with (14). Therefore, the EM algorithms for the cACGMM and the TV-cGMM are equivalent.

It can be also proved that, if the observed signals $\boldsymbol{y}_{tf}$ follow the TV-cGMM (15), then the directional statistics $\boldsymbol{z}_{tf}\triangleq\frac{\boldsymbol{y}_{tf}}{\|\boldsymbol{y}_{tf}\|}$ follow the cACGMM (9). The proof is omitted here, and will be presented in our future publications.

These facts imply that the cACGMM can achieve the same mask estimation accuracy as the TV-cGMM using less information (*i.e.*, direction information $\boldsymbol{z}_{tf}$ only).