

Chapter 11

Recent Advances in Multichannel Source Separation and Denoising Based on Source Sparseness

Nobutaka Ito, Shoko Araki and Tomohiro Nakatani

Abstract This chapter deals with multichannel source separation and denoising based on sparseness of source signals in the time-frequency domain. In this approach, time-frequency masks are typically estimated based on clustering of source location features, such as time and level differences between microphones. In this chapter, we describe the approach and its recent advances. Especially, we introduce a recently proposed clustering method, *observation vector clustering*, which has attracted attention for its effectiveness. We introduce algorithms for observation vector clustering based on a complex Watson mixture model (cWMM), a complex Bingham mixture model (cBMM), and a complex Gaussian mixture model (cGMM). We show through experiments the effectiveness of observation vector clustering in source separation and denoising.

11.1 Introduction

When a desired sound is recorded by distant microphones, it is mixed with other sounds, which often degrade speech quality and intelligibility as well as automatic speech recognition (ASR) performance. To resolve this problem, techniques such as source separation, denoising, and dereverberation have been studied extensively. This chapter focuses on source separation and denoising; see [1] for dereverberation.

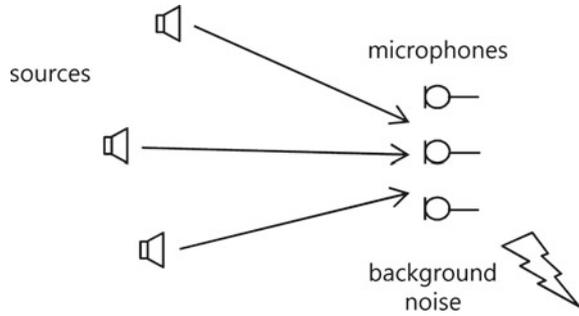
Figure 11.1 illustrates source separation and denoising we deal with in this paper. Suppose we record $N (\geq 1)$ source signals in the presence of background noise by using $M (\geq 2)$ microphones. Our goal is to estimate each source signal from the observed signals. Note that there is not only a multichannel approach [2–5] using mul-

N. Ito (✉) · S. Araki · T. Nakatani
NTT Communication Science Laboratories, NTT Corporation,
2-4 Hikaridai, Seikacho, Kyoto, Sorakugun 619-0237, Japan
e-mail: ito.nobutaka@lab.ntt.co.jp

S. Araki
e-mail: araki.shoko@lab.ntt.co.jp

T. Nakatani
e-mail: nakatani.tomohiro@lab.ntt.co.jp

Fig. 11.1 Source separation and denoising we deal with in this paper



multiple microphones but also a single-channel approach using a single microphone [6–9]. A main advantage of the multichannel approach is that it can perform source separation and denoising with little or even no distortion in the desired source signal.

Especially, multichannel source separation and denoising based on source sparseness [10–20] have turned out to be highly effective and robust in the real world [16, 17, 19, 20]. Various signals including speech are known to have sparseness in the time-frequency domain: a small percentage of the time-frequency components of a signal capture a large percentage of its overall energy [10]. The source sparseness is often exploited by assuming that the observed signals are *dominated* by a single source signal or by background noise at each time-frequency point. We call this a *sparseness assumption*. The dominating source signal or background noise at each time-frequency point can be represented by *masks*. Once we have obtained these masks, we can estimate the source signals either by applying the masks directly to the observed signals (*masking*) [10–14, 16, 17, 19, 21] or by applying beamformers designed based on the masks [15, 18, 20, 22].

The key to the effectiveness of this approach is accurate estimation of the masks, which is usually performed based on either spatial information [10–20] or spectral information [21, 22]. We focus on the former, which employs source location features extracted from the observed signals, such as time and level differences between microphones. The sparseness assumption implies that the source location features form clusters, each of which corresponds to a source signal or the background noise. These clusters can be found by clustering the source location features to obtain the masks. This is typically done by fitting a mixture model to the features, where the appropriate design of the features and the mixture model is significant to mask estimation accuracy.

In this chapter, we introduce a recently proposed clustering method, *observation vector clustering*, which has attracted attention for its effectiveness [11, 13, 15–20]. This method has been employed in many evaluation campaigns successfully [17, 20]. We introduce algorithms for observation vector clustering based on a complex Watson mixture model (cWMM), a complex Bingham mixture model (cBMM), and a complex Gaussian mixture model (cGMM).

The rest of this chapter is organized as follows. Section 11.2 overviews source separation and denoising based on the observation vector clustering. Section 11.3

introduces algorithms for observation vector clustering based on the cWMM, the cBMM, and the cGMM. Section 11.4 describes experiments, and Sect. 11.5 concludes this chapter.

11.2 Source Separation and Denoising Based on Observation Vector Clustering

This section overviews source separation and denoising based on observation vector clustering. Figure 11.2 shows the overall processing flow of this method. In *mask estimation*, masks are estimated from the observed signals. In *source signal estimation*, source signals are estimated by masking or beamforming based on the estimated masks.

11.2.1 Mask Estimation

Figure 11.3 shows the processing flow of *mask estimation* in Fig. 11.2. In *feature extraction*, a source location feature vector is extracted from the observed signals. In *frequency-wise clustering*, clustering of the extracted feature vector is performed in each frequency bin. As a result, posterior probabilities are obtained, which indicate how much the individual clusters contribute to each time-frequency point. In *permutation alignment*, the masks are obtained from the posterior probabilities; the details will be explained later.

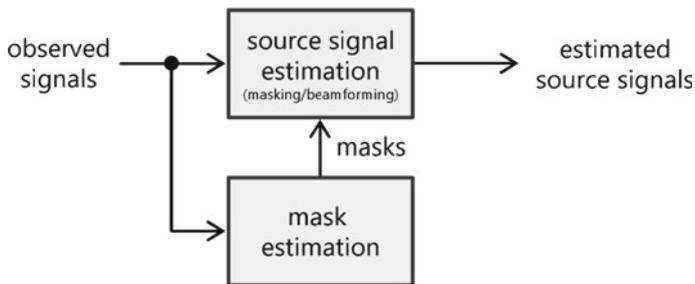


Fig. 11.2 Overall processing flow of source separation and denoising based on observation vector clustering

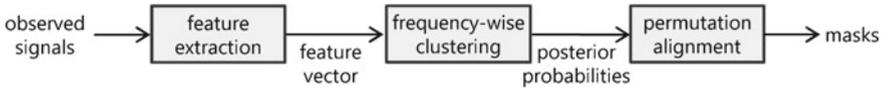


Fig. 11.3 Processing flow of *mask estimation* in Fig. 11.2

Feature Extraction

In *feature extraction* in Fig. 11.3, a source location feature vector is extracted at each time-frequency point. Conventionally, time and level differences between microphones were often employed as source location features. In contrast, in the observation vector clustering, we operate directly on an observation vector composed of multichannel complex spectra.

Let $y_{tf}^{(m)} \in \mathbb{C}$ denote the observed signal at the m th microphone in the short-time Fourier transform (STFT) domain. Here, $m \in \{1, \dots, M\}$ denotes the microphone index; $t \in \{1, \dots, T\}$ the frame index; $f \in \{1, \dots, F\}$ the frequency bin index; M the number of microphones in the array; T the number of frames; F the number of frequency bins up to the Nyquist frequency. We define the *observation vector* by $\mathbf{y}_{tf} \triangleq [y_{tf}^{(1)} \ y_{tf}^{(2)} \ \dots \ y_{tf}^{(M)}]^\top \in \mathbb{C}^M$, where the superscript \top denotes transposition.

We employ the observation vector \mathbf{y}_{tf} as the feature vector \mathbf{z}_{tf} :

$$\mathbf{z}_{tf} = \mathbf{y}_{tf}. \tag{11.1}$$

In this case, \mathbf{z}_{tf} lies in the complex linear space \mathbb{C}^M . Alternatively, we can also employ a normalized observation vector $\frac{\mathbf{y}_{tf}}{\|\mathbf{y}_{tf}\|}$ as the feature vector \mathbf{z}_{tf} :

$$\mathbf{z}_{tf} = \frac{\mathbf{y}_{tf}}{\|\mathbf{y}_{tf}\|}, \tag{11.2}$$

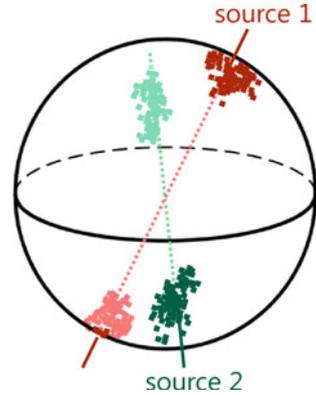
where $\|\cdot\|$ denotes the Euclidean norm. In this case, \mathbf{z}_{tf} lies on the unit hypersphere S^{M-1} in \mathbb{C}^M centered at the origin, because $\|\mathbf{z}_{tf}\| = 1$ (see Fig. 11.4).

In the following, we describe our modeling of the observation vector \mathbf{y}_{tf} . We consider both noiseless and noisy cases.

First, we consider the noiseless case, where $N (\geq 2)$ source signals are recorded by the microphones without noise. The number of sources, N , is assumed to be given throughout this chapter. In this noiseless case, \mathbf{y}_{tf} is modeled by $\mathbf{y}_{tf} = \sum_{n=1}^N s_{tf}^{(n)} \mathbf{h}_{tf}^{(n)}$. Here, $s_{tf}^{(n)}$ denotes the n th source signal in the STFT domain, and $\mathbf{h}_{tf}^{(n)}$ denotes the steering vector for the n th source. The steering vector $\mathbf{h}_{tf}^{(n)}$ represents the acoustic transfer characteristics from the n th source to the microphones. Under the sparseness assumption, the above model can be approximated by $\mathbf{y}_{tf} = s_{tf}^{(v)} \mathbf{h}_{tf}^{(v)}$, where $v = d_{tf}$ denotes the index of the source signal that dominates \mathbf{y}_{tf} at the time-frequency point (t, f) . Here, both $s_{tf}^{(n)}$ and $\mathbf{h}_{tf}^{(n)}$ are unknown.

Next, we consider the noisy case, where $N (\geq 1)$ source signal(s) are recorded by the microphones in the presence of background noise. In this case, \mathbf{y}_{tf} is modeled

Fig. 11.4 Example of the source location feature vector for two sources. Here, \mathbb{C}^M has been simplified to \mathbb{R}^3 for illustration



by $\mathbf{y}_{tf} = \sum_{n=1}^N s_{tf}^{(n)} \mathbf{h}_{tf}^{(n)} + \mathbf{v}_{tf}$, where \mathbf{v}_{tf} denotes the contribution of the background noise to \mathbf{y}_{tf} . Under the sparseness assumption, this model can be approximated by

$$\mathbf{y}_{tf} = \begin{cases} s_{tf}^{(v)} \mathbf{h}_{tf}^{(v)} + \mathbf{v}_{tf}, & \text{if } d_{tf} = v \in \{1, \dots, N\}, \\ \mathbf{v}_{tf}, & \text{if } d_{tf} = 0. \end{cases} \tag{11.3}$$

Here, d_{tf} denotes the index of the source signal or the background noise that dominates \mathbf{y}_{tf} at the time-frequency point (t, f) , where the case $d_{tf} = 0$ corresponds to the background noise and the cases $d_{tf} \in \{1, \dots, N\}$ to the source signals. Note that the background noise \mathbf{v}_{tf} is assumed to be contained in \mathbf{y}_{tf} at all time-frequency points, because it is usually not sparse. $s_{tf}^{(n)}$, $\mathbf{h}_{tf}^{(n)}$, and \mathbf{v}_{tf} are all unknown.

In both cases, our goal is to estimate $s_{tf}^{(n)}$ given \mathbf{y}_{tf} .

Frequency-Wise Clustering

In *frequency-wise clustering* in Fig. 11.3, clustering of the feature vector \mathbf{z}_{tf} is performed in each frequency bin. As a result, the posterior probability $\tilde{\gamma}_{tf}^{(k)}$ is obtained for each cluster k , which indicates how much the k th cluster contributes to the time-frequency point (t, f) .

The clustering can be performed by fitting a mixture model

$$p(\mathbf{z}_{tf} | \Theta_f) = \sum_k \alpha_f^{(k)} p(\mathbf{z}_{tf} | \tilde{d}_{tf} = k, \Theta_f) \tag{11.4}$$

to \mathbf{z}_{tf} . Here, \tilde{d}_{tf} denotes the index of the cluster that \mathbf{z}_{tf} belongs to; $\alpha_f^{(k)} \triangleq P(\tilde{d}_{tf} = k | \Theta_f)$ the prior probability of $\tilde{d}_{tf} = k$; $p(\mathbf{z}_{tf} | \tilde{d}_{tf} = k, \Theta_f)$ the conditional probability density function of \mathbf{z}_{tf} under $\tilde{d}_{tf} = k$; \sum_k the sum over all possible values of k (i.e., $\sum_{k=1}^K$ for the noiseless case; $\sum_{k=0}^K$ for the noisy case); Θ_f the set of all model parameters in (11.4). $\alpha_f^{(k)}$ satisfies $\sum_k \alpha_f^{(k)} = 1$ and $\alpha_f^{(k)} \geq 0$.

Θ_f is estimated by the maximization of the log-likelihood function

$$L(\Theta_f) = \sum_{t=1}^T \ln p(\mathbf{z}_{tf} | \Theta_f), \quad (11.5)$$

which can be done by the expectation-maximization (EM) algorithm. Once Θ_f has been estimated, we obtain the posterior probability $\tilde{\gamma}_{tf}^{(k)}$ based on Bayes' theorem [23] as follows:

$$\tilde{\gamma}_{tf}^{(k)} \triangleq P(\tilde{d}_{tf} = k | \mathbf{z}_{tf}, \Theta_f) \quad (11.6)$$

$$= \frac{\alpha_f^{(k)} p(\mathbf{z}_{tf} | \tilde{d}_{tf} = k, \Theta_f)}{\sum_l \alpha_f^{(l)} p(\mathbf{z}_{tf} | \tilde{d}_{tf} = l, \Theta_f)}. \quad (11.7)$$

Here, $\tilde{\gamma}_{tf}^{(k)}$ satisfies $\sum_k \tilde{\gamma}_{tf}^{(k)} = 1$ and $\tilde{\gamma}_{tf}^{(k)} \geq 0$.

Permutation Alignment

In *permutation alignment* in Fig. 11.3, the masks are obtained by using the posterior probabilities $\tilde{\gamma}_{tf}^{(k)}$.

The index k of the clusters and the index n of the source signals and the background noise do not necessarily coincide, but there is permutation ambiguity between them. This implies that $\tilde{\gamma}_{tf}^{(k)}$ for the same k may correspond to different source signals at different frequencies. Therefore, we need to permute the cluster indexes k so that each k corresponds to the same source signal or background noise in all frequency bins, which is called permutation alignment. As a result of the permutation alignment, we obtain the masks $\gamma_{tf}^{(n)}$.

Many methods have been proposed for permutation alignment [16, 24–26]. Especially, Sawada et al. [16] has proposed an effective method based on correlation of posterior probabilities $\tilde{\gamma}_{tf}^{(k)}$ between frequencies.

11.2.2 Source Signal Estimation

In *source signal estimation* in Fig. 11.2, source signals are estimated by masking or beamforming based on the estimated masks.

Masking

When masking is employed, the source signals are estimated by multiplying an observed signal by the estimated masks $\gamma_{tf}^{(n)}$ as follows:

$$\hat{s}_{tf}^{(n)} = \gamma_{tf}^{(n)} y_{tf}^{(\mu)}. \quad (11.8)$$

Here, μ denotes the index of the reference microphone.

Beamforming

Here we consider the noisy case. Among many types of beamformers, we focus on the MVDR beamformer. The MVDR beamformer is especially suitable for the front end of ASR, because it can perform source separation and denoising without distorting the desired source signal.

The output of the MVDR beamformer is given by

$$\hat{s}_{\mu}^{(n)} = \frac{\mathbf{h}_{\mu}^{(n)H} (\Phi_f^y)^{-1} \mathbf{y}_{\mu}}{\mathbf{h}_{\mu}^{(n)H} (\Phi_f^y)^{-1} \mathbf{h}_{\mu}^{(n)}}. \tag{11.9}$$

Φ_f^y denotes the covariance matrix of \mathbf{y}_{μ} , which can be estimated by

$$\hat{\Phi}_f^y = \frac{1}{T} \sum_{t=1}^T \mathbf{y}_{\mu} \mathbf{y}_{\mu}^H. \tag{11.10}$$

In the MVDR beamformer, accurate estimation of the steering vector $\mathbf{h}_{\mu}^{(n)}$ is crucial.

Conventionally, $\mathbf{h}_{\mu}^{(n)}$ was estimated based on the assumptions of planewave propagation and a known array geometry. These assumptions are often violated in the real world, and lead to degraded performances of the MVDR beamformer and therefore ASR. Here we present mask-based steering vector estimation, which does not rely on these assumptions, and therefore is more robust in the real world.

First, a covariance matrix $\Psi_f^{(n)}$ corresponding to the n th source signal plus the background noise is estimated by

$$\Psi_f^{(n)} = \frac{\sum_{t=1}^T \gamma_{\mu}^{(n)} \mathbf{y}_{\mu} \mathbf{y}_{\mu}^H}{\sum_{t=1}^T \gamma_{\mu}^{(n)}}, \tag{11.11}$$

and a noise covariance matrix $\Psi_f^{(0)}$ is estimated by

$$\Psi_f^{(0)} = \frac{\sum_{t=1}^T \gamma_{\mu}^{(0)} \mathbf{y}_{\mu} \mathbf{y}_{\mu}^H}{\sum_{t=1}^T \gamma_{\mu}^{(0)}}. \tag{11.12}$$

The noise contribution to $\Psi_f^{(n)}$ is reduced by subtracting $\Psi_f^{(0)}$ from $\Psi_f^{(n)}$. The steering vector $\mathbf{h}_{\mu}^{(n)}$ is estimated as a principal eigenvector of the resultant matrix $\Psi_f^{(n)} - \Psi_f^{(0)}$.

11.3 Mask Estimation Based on Modeling Directional Statistics

Several mixture models for the feature vector \mathbf{z}_{tf} have been proposed to estimate the masks accurately. These mixture models include the cWMM, the cBMM, and the cGMM, which are specific examples of the general mixture model (11.4).

11.3.1 Mask Estimation Based on Complex Watson Mixture Model (cWMM)

Sawada et al. [13, 16] and Tran Vu et al. [15] have proposed to estimate masks based on modeling the feature vector (11.2) by a *complex Watson mixture model (cWMM)*. The cWMM is composed of *complex Watson distributions* of Mardia et al. [27], and the complex Watson distribution is an extension of a real Watson distribution of Watson [28].

The probability density function (PDF) of the cWMM is given by

$$p(\mathbf{z}_{tf}; \Theta_{\mathbf{W},f}) = \sum_k \alpha_f^{(k)} p_{\mathbf{W}}(\mathbf{z}_{tf}; \mathbf{a}_f^{(k)}, \kappa_f^{(k)}), \quad (11.13)$$

where $p_{\mathbf{W}}$ denotes a complex Watson distribution

$$p_{\mathbf{W}}(\mathbf{z}; \mathbf{a}, \kappa) \triangleq \frac{(M-1)!}{2\pi^M \mathcal{H}(1, M; \kappa)} \exp(\kappa |\mathbf{a}^H \mathbf{z}|^2). \quad (11.14)$$

Both the complex Watson distribution and the cWMM are defined on the unit hypersphere in \mathbb{C}^M :

$$S^{M-1} \triangleq \left\{ \mathbf{z} \in \mathbb{C}^M \mid \|\mathbf{z}\| = 1 \right\}, \quad (11.15)$$

which is illustrated in Fig. 11.4. Each complex Watson distribution in (11.13) models the distribution of \mathbf{z}_{tf} for a cluster. k denotes the cluster index.

$$\Theta_{\mathbf{W},f} \triangleq \left\{ \alpha_f^{(k)}, \mathbf{a}_f^{(k)}, \kappa_f^{(k)} \mid \forall k \right\} \quad (11.16)$$

denotes the set of all model parameters of the cWMM (11.13), where $\alpha_f^{(k)}$ satisfies

$$\alpha_f^{(k)} \geq 0, \quad (11.17)$$

$$\sum_k \alpha_f^{(k)} = 1, \quad (11.18)$$

$\mathbf{a}_f^{(k)}$ denotes a parameter representing the mean orientation of \mathbf{z}_{if} for the k th cluster satisfying

$$\|\mathbf{a}_f^{(k)}\| = 1, \quad (11.19)$$

and $\kappa \in \mathbb{R}$ denotes a parameter representing the concentration of the distribution of \mathbf{z}_{if} for the k th cluster. H denotes conjugate transposition; \mathcal{K} the confluent hypergeometric function of the first kind, also known as the Kummer function, which is defined by the following power series:

$$\mathcal{K}(\xi, \eta; \kappa) \triangleq 1 + \frac{\xi}{\eta} \frac{\kappa}{1!} + \frac{\xi(\xi+1)}{\eta(\eta+1)} \frac{\kappa^2}{2!} + \dots \quad (11.20)$$

To analyze the behavior of (11.14) as a function of \mathbf{z} , note that (11.14) depends on \mathbf{z} through the term $|\mathbf{a}^H \mathbf{z}|$ only and increases[decreases] monotonically as $|\mathbf{a}^H \mathbf{z}|$ increases when $\kappa > 0$ [$\kappa < 0$]. Note also that

$$0 \leq |\mathbf{a}^H \mathbf{z}| \leq 1, \quad (11.21)$$

which follows from the Cauchy-Schwartz inequality and $\|\mathbf{z}\| = \|\mathbf{a}\| = 1$. Therefore, for $\kappa > 0$ [$\kappa < 0$], (11.14) has the global minima[maxima] at

$$\{\mathbf{z} \in S^{M-1} \mid |\mathbf{a}^H \mathbf{z}| = 0\}, \quad (11.22)$$

increases[decreases] monotonically as $|\mathbf{a}^H \mathbf{z}|$ increases, and has the global maxima[minima] at

$$\{\mathbf{z} \in S^{M-1} \mid |\mathbf{a}^H \mathbf{z}| = 1\}. \quad (11.23)$$

Note that (11.22) equals

$$\{\mathbf{z} \in S^{M-1} \mid \mathbf{a}^H \mathbf{z} = 0\}, \quad (11.24)$$

and (11.23) equals

$$\{\exp(j\theta)\mathbf{a} \mid \theta \in [0, 2\pi)\}. \quad (11.25)$$

It is straightforward to see that, for $\kappa = 0$, (11.14) is constant (i.e., uniform distribution on S^{M-1}). Based on the above property, we impose a constraint

$$\kappa_f^{(k)} > 0, \quad (11.26)$$

which is appropriate for our application.

Once the model parameters $\Theta_{W,f}$ have been estimated, the posterior probability $\tilde{\gamma}_{if}^{(k)}$ can be obtained based on Bayes' theorem [23] by

$$\tilde{\gamma}_{if}^{(k)} \leftarrow \frac{\alpha_f^{(k)} p_{\text{W}}(\mathbf{z}_{if}; \mathbf{a}_f^{(k)}, \kappa_f^{(k)})}{\sum_l \alpha_f^{(l)} p_{\text{W}}(\mathbf{z}_{if}; \mathbf{a}_f^{(l)}, \kappa_f^{(l)})}. \quad (11.27)$$

To estimate the model parameters $\Theta_{\text{W},f}$, the cWMM (11.13) is fitted to the feature vector \mathbf{z}_{if} , e.g., based on the maximization of the log-likelihood function

$$\sum_{t=1}^T \ln p(\mathbf{z}_{if}; \Theta_{\text{W},f}). \quad (11.28)$$

This is realized by, e.g., an expectation-maximization (EM) algorithm [23], which consists in alternate iteration of an E-step and an M-step. The E-step consists in updating the posterior probability $\gamma_{if}^{(k)}$ by (11.27) using current estimates of the model parameters $\Theta_{\text{W},f}$. The M-step consists in updating the model parameters $\Theta_{\text{W},f}$ using the posterior probability $\gamma_{if}^{(k)}$, which is realized by applying the following update rules:

$$\alpha_f^{(k)} \leftarrow \frac{1}{T} \sum_{t=1}^T \tilde{\gamma}_{if}^{(k)}, \quad (11.29)$$

$$\mathbf{R}_f^{(k)} \leftarrow \frac{\sum_{t=1}^T \tilde{\gamma}_{if}^{(k)} \mathbf{z}_{if} \mathbf{z}_{if}^{\text{H}}}{\sum_{t=1}^T \tilde{\gamma}_{if}^{(k)}}, \quad (11.30)$$

$$(\lambda_f^{(k)}, \mathbf{a}_f^{(k)}) \leftarrow \text{the largest eigenvalue and a corresponding eigenvector of } \mathbf{R}_f^{(k)}, \quad (11.31)$$

$$\mathbf{a}_f^{(k)} \leftarrow \frac{\mathbf{a}_f^{(k)}}{\|\mathbf{a}_f^{(k)}\|}, \quad (11.32)$$

$$\kappa_f^{(k)} \leftarrow \frac{M \lambda_f^{(k)} - 1}{2 \lambda_f^{(k)} (1 - \lambda_f^{(k)})} \left[1 + \sqrt{1 + \frac{4(M+1) \lambda_f^{(k)} (1 - \lambda_f^{(k)})}{M-1}} \right]. \quad (11.33)$$

See Appendix 1 for derivation of this EM algorithm.

A major limitation of the cWMM lies in that the complex Watson distribution (11.14) can represent a distribution that is rotationally symmetric about the axis \mathbf{a} (see Fig. 11.5). Indeed, as we have already noted, (11.14) is a function of $|\mathbf{a}^{\text{H}} \mathbf{z}|$, which can be regarded as the cosine of the angle between \mathbf{a} and \mathbf{z} . However, the distribution of the feature vector \mathbf{z}_{if} for each cluster is not necessarily rotationally symmetric, depending on various conditions such as the array geometry and acoustic transfer characteristics. The cWMM therefore has a limited ability to approximate the distribution of \mathbf{z}_{if} , which results in degraded mask estimation accuracy and therefore degraded performance of source separation and denoising. This motivates us to consider more flexible distributions, which are described in Sects. 11.3.2 and 11.3.3.

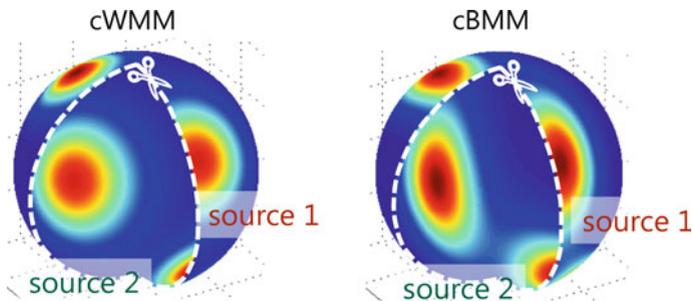


Fig. 11.5 Illustration of the cWMM and the cBMM for two sources

11.3.2 Mask Estimation Based on Complex Bingham Mixture Model (cBMM)

To overcome the above limitation of the cWMM, Ito et al. have proposed to estimate masks based on modeling the feature vector (11.2) by a *complex Bingham mixture model (cBMM)* [29]. The cBMM is composed of *complex Bingham distributions* of Kent [30], and the complex Bingham distribution is an extension of the real Bingham distribution of Bingham [31]. The complex Bingham distribution can represent not only rotationally symmetric but also elliptical distributions on the unit hypersphere (see Fig. 11.5), and can therefore better approximate the distribution of the feature vector \mathbf{z}_{tf} than the complex Watson distribution. As a result, the cBMM can improve mask estimation accuracy and therefore source separation and denoising performance compared to the cWMM.

The PDF of the cBMM is given by

$$p(\mathbf{z}_{tf}; \Theta_{B,f}) = \sum_k \alpha_f^{(k)} p_B(\mathbf{z}_{tf}; \mathbf{B}_f^{(k)}), \tag{11.34}$$

where p_B denotes a complex Bingham distribution

$$p_B(\mathbf{z}; \mathbf{B}) \triangleq c(\mathbf{B})^{-1} \exp(\mathbf{z}^H \mathbf{B} \mathbf{z}). \tag{11.35}$$

Here, $c(\mathbf{B})$ denotes the following function defined for a Hermitian matrix \mathbf{B} .

$$c(\mathbf{B}) \triangleq \left[2\pi^M \sum_{m=1}^M \frac{\exp(\beta_m)}{\prod_{l \neq m} (\beta_m - \beta_l)} \right], \tag{11.36}$$

where $\beta_m, m = 1, \dots, M$, denote the eigenvalues of \mathbf{B} . Both the complex Bingham distribution and the cBMM are defined on the unit hypersphere S^{M-1} . Each complex Bingham distribution in (11.34) models the distribution of \mathbf{z}_{tf} for a cluster.

$$\Theta_{B,f} \triangleq \left\{ \alpha_f^{(k)}, \mathbf{B}_f^{(k)} \mid \forall k \right\} \quad (11.37)$$

denotes the set of all model parameters of the cBMM, where $\mathbf{B}_f^{(k)}$ is a Hermitian parameter matrix, which represents not only the location and the concentration, but also the direction and the shape, of the complex Bingham distribution. Note that the expression for the normalization factor in (11.35) is valid only when the eigenvalues of \mathbf{B} are all distinct, which is always satisfied in practice.

Once the model parameters $\Theta_{B,f}$ have been estimated, the posterior probability $\tilde{\gamma}_{tf}^{(k)}$ can be obtained by

$$\tilde{\gamma}_{tf}^{(k)} \leftarrow \frac{\alpha_f^{(k)} p_B(\mathbf{z}_{tf}; \mathbf{B}_f^{(k)})}{\sum_l \alpha_f^{(l)} p_B(\mathbf{z}_{tf}; \mathbf{B}_f^{(l)})}. \quad (11.38)$$

As in the cWMM case, $\Theta_{B,f}$ can be estimated by the maximum likelihood method based on the EM algorithm. The E-step consists in updating $\tilde{\gamma}_{tf}^{(k)}$ by (11.38) using the current $\Theta_{B,f}$ value. The M-step consists in updating $\Theta_{B,f}$ using $\tilde{\gamma}_{tf}^{(k)}$, which is realized by applying the following update rules:

$$\alpha_f^{(k)} \leftarrow \frac{1}{T} \sum_{t=1}^T \tilde{\gamma}_{tf}^{(k)}, \quad (11.39)$$

$$\mathbf{R}_f^{(k)} \leftarrow \frac{\sum_{t=1}^T \tilde{\gamma}_{tf}^{(k)} \mathbf{z}_{tf} \mathbf{z}_{tf}^H}{\sum_{t=1}^T \tilde{\gamma}_{tf}^{(k)}}, \quad (11.40)$$

$$(\lambda_{fm}^{(k)}, \mathbf{a}_{fm}^{(k)}) \leftarrow \begin{array}{l} \text{the } m\text{th largest eigenvalue and a corresponding eigenvector} \\ \text{of } \mathbf{R}_f^{(k)}, \end{array} \quad (11.41)$$

$$\mathbf{a}_{fm}^{(k)} \leftarrow \frac{\mathbf{a}_{fm}^{(k)}}{\|\mathbf{a}_{fm}^{(k)}\|}, \quad (11.42)$$

$$\mathbf{B}_f^{(k)} \leftarrow \sum_{m=1}^M \left(-\frac{1}{\lambda_{fm}^{(k)}} + \frac{1}{\lambda_{f1}^{(k)}} \right) \mathbf{a}_{fm}^{(k)} \mathbf{a}_{fm}^{(k)H}. \quad (11.43)$$

See Appendix 2 for derivation of the above algorithm.

Note that the parameter matrix $\mathbf{B}_f^{(k)}$ has the following indeterminacy

$$p_{\mathbf{B}}(\mathbf{z}_{tf}; \mathbf{B}_f^{(k)}) = p_{\mathbf{B}}(\mathbf{z}_{tf}; \mathbf{B}_f^{(k)} + \xi \mathbf{I}), \forall \xi \in \mathbb{R}, \quad (11.44)$$

which follows from $\|\mathbf{z}_{tf}\| = 1$. Here, \mathbf{I} denotes the $M \times M$ identity matrix. To remove this indeterminacy, in the above algorithm, ξ has been determined so that the largest eigenvalue of $\mathbf{B}_f^{(k)}$ equals zero.

11.3.3 Mask Estimation Based on Complex Gaussian Mixture Model (cGMM)

As an alternative method, Ito et al. have proposed to estimate masks based on modeling the feature vector (11.1) by a *complex (time-varying) Gaussian mixture model (cGMM)* [32], inspired by Duong et al. [33]. Note that the cGMM models the observation vector itself in (11.1), instead of its normalized version in (11.2). The cGMM is composed of complex Gaussian distributions, where the covariance matrices are parametrized by time-invariant spatial covariance matrices and time-variant power parameters.

The PDF of the cGMM is given by

$$p(\mathbf{z}_{tf}; \Theta_{G,f}) = \sum_k \alpha_f^{(k)} p_{\mathbf{G}}(\mathbf{z}_{tf}; 0, \phi_{tf}^{(k)} \mathbf{B}_f^{(k)}), \quad (11.45)$$

where $p_{\mathbf{G}}$ denotes a complex Gaussian distribution

$$p_{\mathbf{G}}(\mathbf{z}; \mathbf{g}, \Sigma) \triangleq \frac{1}{\pi^M \det \Sigma} \exp[-(\mathbf{z} - \mathbf{g})^H \Sigma^{-1} (\mathbf{z} - \mathbf{g})], \quad (11.46)$$

with \mathbf{g} being the mean and Σ the covariance matrix. Both the complex Gaussian distribution and the cGMM are defined in \mathbb{C}^M . Each complex Gaussian distribution in (11.45) models the distribution of \mathbf{z}_{tf} for a cluster.

$$\Theta_{G,f} \triangleq \left\{ \alpha_f^{(k)}, \mathbf{B}_f^{(k)} \middle| \forall k \right\} \cup \left\{ \phi_{tf}^{(k)} \middle| \forall k, \forall t \right\} \quad (11.47)$$

denotes the set of all model parameters of the cGMM, where $\mathbf{B}_f^{(k)}$ is a scaled covariance matrix modeling the direction of the observation vector in (11.1) (i.e., the normalized observation vector (11.2)), and $\phi_{tf}^{(k)}$ is a power parameter modeling the magnitude of the observation vector.

Once the model parameters $\Theta_{G,f}$ have been estimated, the posterior probability $\tilde{\gamma}_{tf}^{(k)}$ can be obtained by

$$\tilde{\gamma}_{tf}^{(k)} \leftarrow \frac{\alpha_f^{(k)} p_G(\mathbf{y}_{tf}; \mathbf{0}, \phi_{tf}^{(k)} \mathbf{B}_f^{(k)})}{\sum_l \alpha_f^{(l)} p_G(\mathbf{y}_{tf}; \mathbf{0}, \phi_{tf}^{(l)} \mathbf{B}_f^{(l)})}. \quad (11.48)$$

As in the cWMM and the cBMM cases, $\Theta_{G,f}$ can be estimated by the maximum likelihood method based on the EM algorithm. The E-step consists in updating $\tilde{\gamma}_{tf}^{(k)}$ by (11.48) using the current $\Theta_{G,f}$ value. The M-step consists in updating $\Theta_{G,f}$ using $\tilde{\gamma}_{tf}^{(k)}$, which is realized by applying the following update rules:

$$\alpha_f^{(k)} \leftarrow \frac{1}{T} \sum_{t=1}^T \tilde{\gamma}_{tf}^{(k)}, \quad (11.49)$$

$$\mathbf{B}_f^{(k)} \leftarrow \frac{\sum_{t=1}^T \tilde{\gamma}_{tf}^{(k)} \mathbf{y}_{tf} \mathbf{y}_{tf}^H / \phi_{tf}^{(k)}}{\sum_{t=1}^T \tilde{\gamma}_{tf}^{(k)}}, \quad (11.50)$$

$$\phi_{tf}^{(k)} \leftarrow \frac{1}{M} \mathbf{y}_{tf}^H \left(\mathbf{B}_f^{(k)} \right)^{-1} \mathbf{y}_{tf}. \quad (11.51)$$

See Appendix 3 for derivation of the above algorithm.

11.4 Experimental Evaluation

We conducted source separation and denoising experiments to verify the effectiveness of observation vector clustering introduced in this chapter.

11.4.1 Source Separation

We first describe the source separation experiment. We assumed that the number of sources was known. We generated observed signals by convolving 8s-long English speech signals with room impulse responses measured in an experimental room (see Fig. 11.6). 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 [16]. Source signal estimates were obtained based on masking as in (11.8).

Figure 11.7 shows the signal-to-distortion ratio (SDR) [34] as a function of the reverberation time RT_{60} , and Fig. 11.8 shows an example of source separation results. The SDRs were averaged over 16 trials with eight combinations of speech signals and two distances between a loudspeaker and the array center. The azimuths of sources were 70° and 150° for $N = 2$, and 70° , 150° , and 245° for $N = 3$.

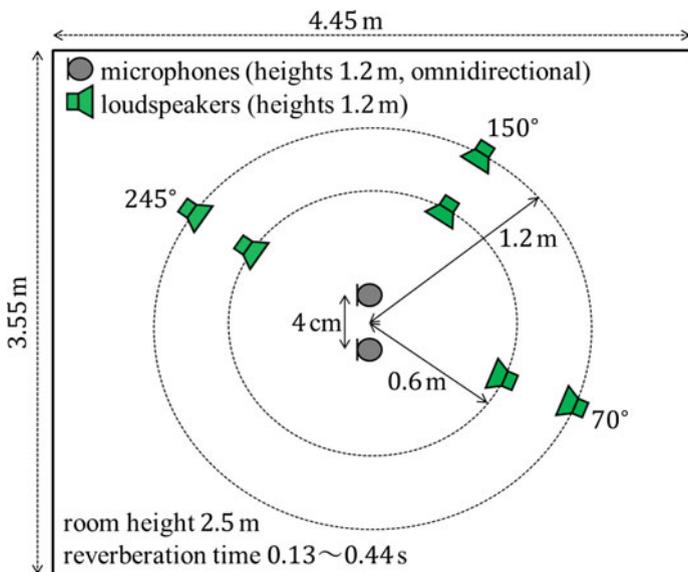


Fig. 11.6 Configurations in room impulse response measurement

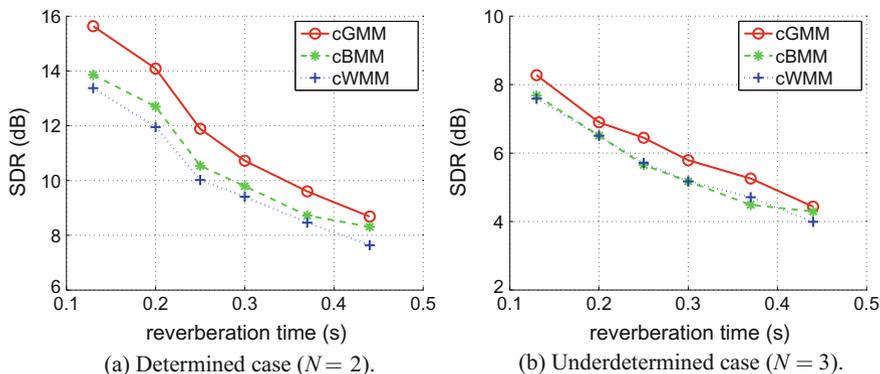


Fig. 11.7 Signal-to-distortion ratio (SDR) as a function of the reverberation time RT_{60}

11.4.2 Denoising

Now we move on to the denoising experiment. The performance was measured by the word error rate (WER) of ASR on the CHiME-3 task [35]. The CHiME-3 task consists in recognition of 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 [35].

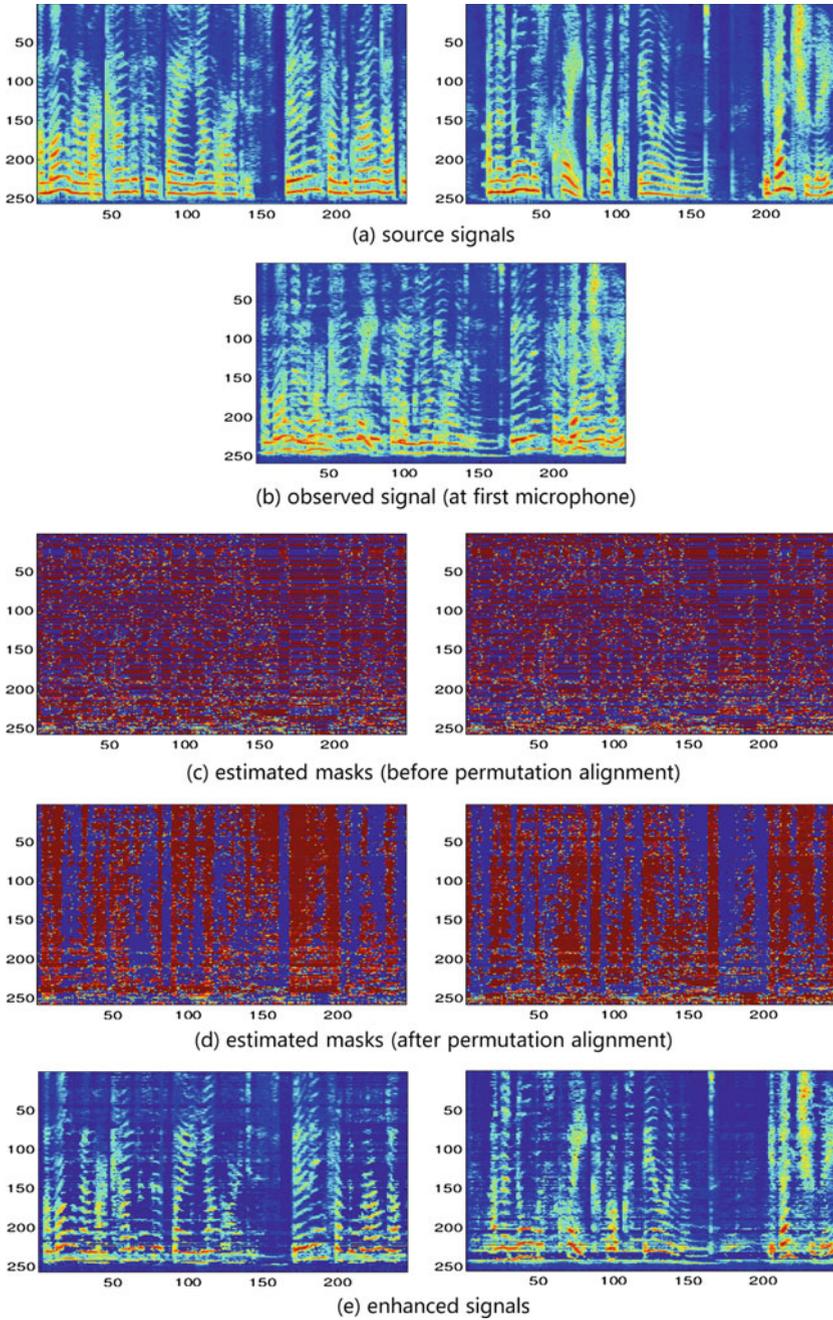


Fig. 11.8 Example of source separation results for $N = 2$ and $RT_{60} = 130$ ms. The horizontal axis represents the time, and the vertical the frequency. To focus on low frequencies, which contain most speech energy, only the frequency range of 0 to 2 kHz is shown. The temporal range shown corresponds to 10 s

Denoising was performed by using MVDR beamformers designed using the estimated masks as in Sect. 11.2.2. Assuming that the background noise arrive from all directions equally (i.e., noise is diffuse), we set $\kappa_f^{(0)} = 0$ for the cWMM, $\mathbf{B}_f^{(0)} = \mathbf{0}$ for the cBMM, and $\mathbf{B}_f^{(0)} = \mathbf{I}$ for the cGMM. Permutation alignment was performed by the method proposed in [36], which is based on a common amplitude modulation property of speech. The frame length and the frame shift were 64 ms and 16 ms, respectively, and the window was hann.

ASR was performed by using a DNN-HMM-based acoustic model with a fully connected DNN (10 hidden layers) and an RNN-based language model. The acoustic model was trained on 18 hours of multicondition data.

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 cWMM: 10.2 %,
- denoising with the cBMM: 8.3 %,
- denoising with the cGMM: 9.3 %.

We see that the WER has been reduced significantly by mask-based MVDR beamforming.

11.5 Conclusions

In this chapter, we described multichannel source separation and denoising based on source sparseness. Particularly, we introduced recently proposed framework of observation vector clustering, which have been shown to be effective and robust in the real world. We also introduced specific algorithms for observation vector clustering, based on the cWMM, the cBMM, and the cGMM.

Appendix 1 Derivation of cWMM-Based Mask Estimation Algorithm

Here we derive the cWMM-based mask estimation algorithm in Sect. 11.3.1. The derivation of the E-step is straightforward and omitted. The update rules for the M-step is obtained by maximizing the following Q-function with respect to $\Theta_{W,f}$:

$$Q(\Theta_{W,f}) \triangleq \sum_{t=1}^T \sum_k \tilde{\gamma}_{tf}^{(k)} \ln \left[\alpha_f^{(k)} p_W(\mathbf{z}_{tf}; \mathbf{a}_f^{(k)}, \kappa_f^{(k)}) \right] \quad (11.52)$$

$$= \sum_k \left(\sum_{t=1}^T \tilde{\gamma}_{tf}^{(k)} \right) \ln \alpha_f^{(k)} - \sum_k \left(\sum_{t=1}^T \tilde{\gamma}_{tf}^{(k)} \right) \ln \mathcal{L}(1, M; \kappa_f^{(k)}) \quad (11.53)$$

$$\begin{aligned}
& + \sum_k \kappa_f^{(k)} \mathbf{a}_f^{(k)\text{H}} \left(\sum_{t=1}^T \tilde{\gamma}_{tf}^{(k)} \mathbf{z}_{tf} \mathbf{z}_{tf}^{\text{H}} \right) \mathbf{a}_f^{(k)} + C \\
& = \sum_k \left(\sum_{t=1}^T \tilde{\gamma}_{tf}^{(k)} \right) \left[\ln \alpha_f^{(k)} - \ln \mathcal{H} \left(1, M; \kappa_f^{(k)} \right) + \kappa_f^{(k)} \mathbf{a}_f^{(k)\text{H}} \mathbf{R}_f^{(k)} \mathbf{a}_f^{(k)} \right] + C.
\end{aligned} \tag{11.54}$$

Here, $\mathbf{R}_f^{(k)}$ is defined by

$$\mathbf{R}_f^{(k)} \triangleq \frac{\sum_{t=1}^T \tilde{\gamma}_{tf}^{(k)} \mathbf{z}_{tf} \mathbf{z}_{tf}^{\text{H}}}{\sum_{t=1}^T \tilde{\gamma}_{tf}^{(k)}}, \tag{11.55}$$

and C denotes a constant independent of $\Theta_{W,f}$.

The update rule for $\alpha_f^{(k)}$ is obvious: note the constraint (11.18) and apply the Lagrangian multiplier method.

The update rule for $\mathbf{a}_f^{(k)}$ is obtained by maximizing $Q(\Theta_{W,f})$ subject to (11.19). Noting (11.26), we see that this is equivalent to maximizing $\mathbf{a}_f^{(k)\text{H}} \mathbf{R}_f^{(k)} \mathbf{a}_f^{(k)}$ subject to (11.19). From the linear algebra, $\mathbf{a}_f^{(k)}$ is therefore a unit-norm principal eigenvector of $\mathbf{R}_f^{(k)}$.

The update rule for $\kappa_f^{(k)}$ is obtained by maximizing

$$-\left(\sum_{t=1}^T \tilde{\gamma}_{tf}^{(k)} \right) \ln \mathcal{H} \left(1, M; \kappa_f^{(k)} \right) + \left(\sum_{t=1}^T \tilde{\gamma}_{tf}^{(k)} \right) \kappa_f^{(k)} \mathbf{a}_f^{(k)\text{H}} \mathbf{R}_f^{(k)} \mathbf{a}_f^{(k)}. \tag{11.56}$$

Since $\mathbf{a}_f^{(k)}$ is a unit-norm principal eigenvector of $\mathbf{R}_f^{(k)}$, we have

$$\mathbf{a}_f^{(k)\text{H}} \mathbf{R}_f^{(k)} \mathbf{a}_f^{(k)} = \lambda_f^{(k)}, \tag{11.57}$$

where $\lambda_f^{(k)}$ is the principal eigenvalue of $\mathbf{R}_f^{(k)}$. Therefore, we have the following nonlinear equation for $\kappa_f^{(k)}$:

$$\frac{\partial}{\partial \kappa_f^{(k)}} \mathcal{H} \left(1, M; \kappa_f^{(k)} \right) = \lambda_f^{(k)} \mathcal{H} \left(1, M; \kappa_f^{(k)} \right). \tag{11.58}$$

Using (3.8) in [37], (11.58) is approximately solved as follows:

$$\kappa_f^{(k)} = \frac{M \lambda_f^{(k)} - 1}{2 \lambda_f^{(k)} (1 - \lambda_f^{(k)})} \left[1 + \sqrt{1 + \frac{4(M+1) \lambda_f^{(k)} (1 - \lambda_f^{(k)})}{M-1}} \right]. \tag{11.59}$$

Appendix 2 Derivation of cBMM-Based Mask Estimation Algorithm

Here we derive the cBMM-based mask estimation algorithm in Sect. 11.3.2. The update rule for the E-step is obvious. The update rules for the M-step is obtained by maximizing the following Q-function with respect to $\Theta_{B,f}$:

$$Q(\Theta_{B,f}) \triangleq \sum_{t=1}^T \sum_k \tilde{\gamma}_{tf}^{(k)} \ln \left[\alpha_f^{(k)} p_B(\mathbf{z}_{tf}; \mathbf{B}_f^{(k)}) \right] \quad (11.60)$$

$$= \sum_k \left(\sum_{t=1}^T \tilde{\gamma}_{tf}^{(k)} \right) \ln \alpha_f^{(k)} - \sum_k \left(\sum_{t=1}^T \tilde{\gamma}_{tf}^{(k)} \right) \ln c(\mathbf{B}_f^{(k)}) \quad (11.61)$$

$$+ \sum_k \text{tr} \left[\mathbf{B}_f^{(k)} \left(\sum_{t=1}^T \tilde{\gamma}_{tf}^{(k)} \mathbf{z}_{tf} \mathbf{z}_{tf}^H \right) \right] \\ = \sum_k \left(\sum_{t=1}^T \tilde{\gamma}_{tf}^{(k)} \right) \left[\ln \alpha_f^{(k)} - \ln c(\mathbf{B}_f^{(k)}) + \text{tr}(\mathbf{B}_f^{(k)} \mathbf{R}_f^{(k)}) \right]. \quad (11.62)$$

Here, $c(\mathbf{B})$ is defined by (11.36), and $\mathbf{R}_f^{(k)}$ by (11.55).

The update rule for $\alpha_f^{(k)}$ is obvious.

To derive the update rule for $\mathbf{B}_f^{(k)}$, let us denote the m th largest eigenvalue of $\mathbf{R}_f^{(k)}$ by $\lambda_{fm}^{(k)}$ and a corresponding unit-norm eigenvector by $\mathbf{v}_{fm}^{(k)}$. We assume that $\lambda_{fm}^{(k)}$, $m = 1, \dots, M$, are all distinct and positive, which is always true in practice. $\mathbf{R}_f^{(k)}$ is represented as

$$\mathbf{R}_f^{(k)} = \sum_{m=1}^M \lambda_{fm}^{(k)} \mathbf{v}_{fm}^{(k)} \mathbf{v}_{fm}^{(k)H}. \quad (11.63)$$

From a result in [38], $\mathbf{v}_{fm}^{(k)}$, $m = 1, \dots, M$, are also the eigenvectors of $\mathbf{B}_f^{(k)}$. Hence, $\mathbf{B}_f^{(k)}$ is represented in the form

$$\mathbf{B}_f^{(k)} = \sum_{m=1}^M \beta_{fm}^{(k)} \mathbf{v}_{fm}^{(k)} \mathbf{v}_{fm}^{(k)H}. \quad (11.64)$$

Substituting (11.63) and (11.64) into (11.62) and disregarding terms independent of $\beta_{fm}^{(k)}$, $m = 1, \dots, M$, we have

$$\left(\sum_{t=1}^T \tilde{\gamma}_{tf}^{(k)} \right) \left[-\ln c(\mathbf{B}_f^{(k)}) + \sum_{m=1}^M \lambda_{fm}^{(k)} \beta_{fm}^{(k)} \right]. \quad (11.65)$$

Therefore, we have

$$\frac{\partial \ln c(\mathbf{B}_f^{(k)})}{\partial \beta_{fm}^{(k)}} = \lambda_{fm}^{(k)}. \quad (11.66)$$

Using an approximation in [38], this nonlinear equation can be approximately solved as follows:

$$\beta_{fm}^{(k)} \sim -\frac{1}{\lambda_{fm}^{(k)}}. \quad (11.67)$$

Substituting (11.67) into (11.64) and adding a matrix of the form $\xi \mathbf{I}$ so that the largest eigenvalue of $\mathbf{B}_f^{(k)}$ is zero, we obtain the following update rule for $\mathbf{B}_f^{(k)}$:

$$\mathbf{B}_f^{(k)} \leftarrow \sum_{m=1}^M \left(-\frac{1}{\lambda_{fm}^{(k)}} + \frac{1}{\lambda_{f1}^{(k)}} \right) \mathbf{v}_{fm}^{(k)} \mathbf{v}_{fm}^{(k)H}. \quad (11.68)$$

Appendix 3 Derivation of cGMM-Based Mask Estimation Algorithm

Here we derive the cGMM-based mask estimation algorithm in Sect. 11.3.3. The derivation of the E-step is straightforward and omitted. The update rules for the M-step is obtained by maximizing the following Q-function with respect to $\Theta_{G,f}$:

$$Q(\Theta_{G,f}) \triangleq \sum_{t=1}^T \sum_k \tilde{\gamma}_{tf}^{(k)} \ln \left[\alpha_f^{(k)} p_G(\mathbf{y}_{tf}; 0, \phi_{tf}^{(k)} \mathbf{B}_f^{(k)}) \right] \quad (11.69)$$

$$\begin{aligned} &= \sum_k \left(\sum_{t=1}^T \tilde{\gamma}_{tf}^{(k)} \right) \ln \alpha_f^{(k)} - M \sum_{t=1}^T \sum_k \tilde{\gamma}_{tf}^{(k)} \ln \phi_{tf}^{(k)} \\ &\quad - \sum_k \left(\sum_{t=1}^T \tilde{\gamma}_{tf}^{(k)} \right) \ln \det \mathbf{B}_f^{(k)} - \sum_{t=1}^T \sum_k \frac{\tilde{\gamma}_{tf}^{(k)}}{\phi_{tf}^{(k)}} \mathbf{y}_{tf}^H \left(\mathbf{B}_f^{(k)} \right)^{-1} \mathbf{y}_{tf} + C \end{aligned} \quad (11.70)$$

$$\begin{aligned} &= \sum_k \left(\sum_{t=1}^T \tilde{\gamma}_{tf}^{(k)} \right) \ln \alpha_f^{(k)} - M \sum_{t=1}^T \sum_k \tilde{\gamma}_{tf}^{(k)} \ln \phi_{tf}^{(k)} \\ &\quad - \sum_k \left(\sum_{t=1}^T \tilde{\gamma}_{tf}^{(k)} \right) \ln \det \mathbf{B}_f^{(k)} - \sum_k \text{tr} \left[\left(\mathbf{B}_f^{(k)} \right)^{-1} \left(\sum_{t=1}^T \frac{\tilde{\gamma}_{tf}^{(k)}}{\phi_{tf}^{(k)}} \mathbf{y}_{tf} \mathbf{y}_{tf}^H \right) \right] + C. \end{aligned} \quad (11.71)$$

Here, C denotes a constant independent of $\Theta_{G,f}$.

The update rule for $\alpha_f^{(k)}$ is obvious.

From (11.70), the update rule for $\phi_{tf}^{(k)}$ is given by

$$\phi_{tf}^{(k)} = \frac{1}{M} \mathbf{y}_{tf}^H (\mathbf{B}_f^{(k)})^{-1} \mathbf{y}_{tf}. \quad (11.72)$$

As for $\mathbf{B}_f^{(k)}$, it should satisfy

$$-\left(\sum_{t=1}^T \tilde{\gamma}_{tf}^{(k)}\right) (\mathbf{B}_f^{(k)})^{-1} + (\mathbf{B}_f^{(k)})^{-1} \left(\sum_{t=1}^T \frac{\tilde{\gamma}_{tf}^{(k)}}{\phi_{tf}^{(k)}} \mathbf{y}_{tf} \mathbf{y}_{tf}^H\right) (\mathbf{B}_f^{(k)})^{-1} = 0. \quad (11.73)$$

Therefore, the update rule for $\mathbf{B}_f^{(k)}$ is

$$\mathbf{B}_f^{(k)} = \frac{\sum_{t=1}^T \tilde{\gamma}_{tf}^{(k)} \mathbf{y}_{tf} \mathbf{y}_{tf}^H / \phi_{tf}^{(k)}}{\sum_{t=1}^T \tilde{\gamma}_{tf}^{(k)}}. \quad (11.74)$$

References

1. P.A. Naylor, N.D. Gaubitch, *Speech Dereverberation*. (Springer, 2009)
2. M. Brandstein, D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. (Springer, 2001)
3. R. Zelinski, A microphone array with adaptive post-filtering for noise reduction in reverberant rooms, in *Proceeding of ICASSP* (1988), pp. 2578–2581
4. S. Gannot, D. Burshtein, E. Weinstein, Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Trans. SP* **49**(8), 1614–1626 (2001)
5. S. Doclo, M. Moonen, GSVD-based optimal filtering for single and multimicrophone speech enhancement. *IEEE Trans. SP* **50**(9), 2230–2244 (2002)
6. S.F. Boll, Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. ASSP* **ASSP-27**(2), 113–120 (1979)
7. Y. Ephraim, D. Malah, Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. ASSP* **32**(6), 1109–1121 (1984)
8. R. Miyazaki, H. Saruwatari, T. Inoue, Y. Takahashi, K. Shikano, K. Kondo, Musical-noise-free speech enhancement based on optimized iterative spectral subtraction. *IEEE Trans. ASLP* **20**(7), 2080–2094 (2012)
9. P. Smaragdīs, Probabilistic decompositions of spectra for sound separation, in *Blind Speech Separation*, ed. by S. Makino, T.-W. Lee, H. Sawada (Springer, 2007), pp. 365–386
10. Ö. Yılmaz, S. Rickard, Blind separation of speech mixtures via time-frequency masking. *IEEE Trans. SP* **52**(7), 1830–1847 (2004)
11. S. Araki, H. Sawada, R. Mukai, S. Makino, Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors. *Signal Process.* **87**(8), 1833–1847 (2007)
12. Y. Izumi, N. Ono, S. Sagayama, Sparseness-based 2ch BSS using the EM algorithm in reverberant environment, in *Proceeding of WASPAA* (2007), pp. 147–150
13. H. Sawada, S. Araki, S. Makino, A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures, in *Proceeding of WASPAA* (2007), pp. 139–142
14. M.I. Mandel, R.J. Weiss, D.P.W. Ellis, Model-based expectation-maximization source separation and localization. *IEEE Trans. ASLP* **18**(2), 382–394 (2010)
15. D.H. Tran Vu, R. Haeb-Umbach, Blind speech separation employing directional statistics in an expectation maximization framework, in *Proceeding of ICASSP* (2010), pp. 241–244
16. H. Sawada, S. Araki, S. Makino, Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. *IEEE Trans. ASLP* **19**(3), 516–527 (2011)

17. M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, A. Ogawa, T. Hori, S. Watanabe, M. Fujimoto, T. Yoshioka, T. Oba, Y. Kubo, M. Souden, S.-J. Hahm, A. Nakamura, Speech recognition in the presence of highly non-stationary noise based on spatial, spectral and temporal speech/noise modeling combined with dynamic variance adaptation, in *Proceeding of CHiME 2011 Workshop on Machine Listening in Multisource Environments* (2011), pp. 12–17
18. M. Souden, S. Araki, K. Kinoshita, T. Nakatani, H. Sawada, A multichannel MMSE-based framework for speech source separation and noise reduction. *IEEE Trans. ASLP* **21**(9), 1913–1928 (2013)
19. T. Nakatani, S. Araki, T. Yoshioka, M. Delcroix, M. Fujimoto, Dominance based integration of spatial and spectral features for speech enhancement. *IEEE Trans. ASLP* **21**(12), 2516–2531 (2013)
20. T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W.J. Fabian, M. Espi, T. Higuchi, S. Araki, T. Nakatani, The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices, in *Proceeding of ASRU* (2015), pp. 436–443
21. Y. Wang, D. Wang, Towards scaling up classification-based speech separation. *IEEE Trans. ASLP* **21**(7), 1381–1390 (2013)
22. J. Heymann, L. Drude, R. Haeb-Umbach, Neural network based spectral mask estimation for acoustic beamforming, in *Proceeding of ICASSP* (2016), pp. 196–200
23. C. Bishop, *Pattern Recognition and Machine Learning*. (Springer, 2006)
24. N. Murata, S. Ikeda, A. Ziehe, An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing* **41**(1–4), 1–24 (2001)
25. H. Sawada, R. Mukai, S. Araki, S. Makino, A robust and precise method for solving the permutation problem of frequency-domain blind source separation. *IEEE Trans. SAP* **12**(5), 530–538 (2004)
26. H. Sawada, S. Araki, S. Makino, Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS, in *Proceeding of IEEE International Symposium on Circuits and Systems (ISCAS)* (2007), pp. 3247–3250
27. K.V. Mardia, I.L. Dryden, The complex Watson distribution and shape analysis. *J. Roy. Stat. Soc.: Ser. B (Stat. Methodol.)* **61**(4), 913–926 (1999)
28. G. Watson, Equatorial distributions on a sphere. *Biometrika* **52**, 193–201 (1965)
29. N. Ito, S. Araki, T. Nakatani, Modeling audio directional statistics using a complex Bingham mixture model for blind source extraction from diffuse noise, in *Proceeding of ICASSP* (2016), pp. 465–468
30. J.T. Kent, The complex Bingham distribution and shape analysis. *J. Roy. Stat. Soc.: Ser. B (Methodol.)* **56**(2), 285–299 (1994)
31. C. Bingham, An antipodally symmetric distribution on the sphere. *Ann. Stat.* **2**, 1201–1205 (1974)
32. N. Ito, S. Araki, T. Yoshioka, T. Nakatani, Relaxed disjointness based clustering for joint blind source separation and dereverberation, in *Proceeding of IWAENC* (2014), pp. 268–272
33. N.Q.K. Duong, E. Vincent, R. Gribonval, Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Trans. ASLP* **18**(7), 1830–1840 (2010)
34. E. Vincent, R. Gribonval, C. Févotte, Performance measurement in blind audio source separation. *IEEE Trans. ASLP* **14**(4), 1462–1469 (2006)
35. J. Barker, R. Marxer, E. Vincent, S. Watanabe, The third ‘CHiME’ speech separation and recognition challenge: Dataset, task and baselines, in *Proceeding of ASRU* (2015), pp. 504–511
36. N. Ito, S. Araki, T. Nakatani, Permutation-free clustering of relative transfer function features for blind source separation, in *Proceeding of EUSIPCO* (2015), pp. 409–413
37. S. Sra, D. Karp, The multivariate Watson distribution: maximum-likelihood estimation and other aspects. *J. Multivar. Anal.* **114**, 256–269 (2013)
38. K.V. Mardia, P.E. Jupp, *Directional Statistics*. (Wiley, 1999)