Gamma Process FastMNMF for Separating an Unknown Number of Sound Sources

Yoshiaki Bando*[†], Kouhei Sekiguchi[†‡], and Kazuyoshi Yoshii[‡†]
*National Institute of Advanced Industrial Science and Technology, Japan [†]AIP, RIKEN, Japan
[‡]Graduate School of Informatics, Kyoto University, Japan

***Abstract*—This paper presents a blind source separation (BSS) method to separate an unknown number of sound sources. A state-of-the-art BSS method called fast multichannel nonnegative matrix factorization (FastMNMF) represents the power spectral density of sources with an NMF model and their spatial covariance matrices (SCMs) with a jointly-diagonalizable (JD) full-rank model. Thanks to the JD SCMs, this method can separate more realistic reverberant and noisy mixtures compared to the conventional rank-1 spatial models. In this paper, we extend FastMNMF to work with an unknown number of sound sources based on a Bayesian non-parametric framework. Because FastMNMF can be considered as nonnegative tensor factorization (NTF) on a diagonalized spectrogram, we utilize a gamma process to this NTF by introducing a latent source activation variable to encourage the shrinkage of the redundant source classes. The proposed inference is formulated as a variational expectation-maximization algorithm for jointly estimating the NTF parameters and diagonalizer. We experimentally confirmed that the proposed method robustly separated an unknown number of sources while the conventional FastMNMF required careful parameter selection depending on the actual number of sources.**

***Index Terms*—Blind source separation, FastMNMF, gamma process, Bayesian signal processing**

I. Introduction

Blind source separation (BSS) is a technique to separate sound source signals from a multichannel mixture recording with few prior information about sources and microphones [1]–[3]. The recent BSS methods have been investigated by formulating probabilistic generative models of the observed mixture signal, and they can be categorized into two types: mixture and factor models [4]–[9]. The mixture model assumes that each time-frequency (TF) bin of the observed multichannel mixture follows a multivariate Gaussian distribution with a spatial covariance matrix (SCM) [4]–[6]. Assuming that only one source signal is dominant in each TF bin, the SCM is switched according to the dominant source. In the factor model, on the other hand, the TF bin is formulated with the sum of SCMs for all the source signals [7]–[9]. Since an audio mixture signal is represented by a sum of source signals, the factor model is a more natural representation of audio signals rather than the mixture model.

The advancement of the efficient SCM representation has been offering significant performance benefits to the factor-based BSS [2], [8]–[12]. Since the SCMs without any constraints have a too high degree of freedom, their estimation is unstable and has frequency permutation ambiguity. To solve this problem, multichannel nonnegative matrix factorization (MNMF) [9] has been proposed by assuming an NMF model on the power spectral density (PSD) of each source. The NMF model represents the PSD as a product of spectral basis vectors and their activation vectors. Independent low-rank matrix analysis (ILRMA) [2], which performs fast and stable inference of MNMF, has been proposed by assuming that the rank of SCMs is only one (rank-1) and the number of sources is equal to that of microphones (determined condition). To mitigate these assumptions, FastMNMF [11], [12] has recently been proposed by utilizing a joint diagonalization (JD) technique. In the FastMNMF model, each of the SCMs is represented by a weighted sum of a common set of rank-1 SCMs. As the parameters of SCMs are efficiently reduced, FastMNMF achieved both high performance and fast inference.

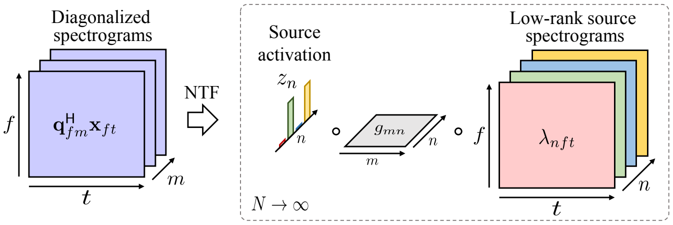

Fig. 1. Overview of the proposed gamma process FastMNMF

Most of the existing BSS methods require the number of sound sources in an observed mixture as a hyperparameter. The performance of the methods often changes inconveniently depending on this parameter. A popular solution is to jointly count and separate sound sources, which has mainly been studied for the mixture models by clustering the TF bins of the observed mixture signal [6], [13]–[16]. Preparing a sufficiently large number of source classes, a sparse prior distribution is assumed on the activation of each source to encourage the shrinkage of redundant sources. Bayesian non-parametric priors such as the Dirichlet process [13], [16] have also been utilized for representing the arbitrary number of sources. This approach, however, has performance limitations due to the disjointness assumption of the mixture model. Few studies aim at the factor models, and only one study combined a rank-1 spatial model and a Bayesian non-parametric framework based on the Beta-process [17].

In this paper, we propose Bayesian non-parametric FastMNMF for separating an unknown number of sources with the strong JD full-rank spatial model (Fig. 1). FastMNMF can also be regarded as a joint optimization of diagonalizer (linear

2021 29th European Signal Processing Conference (EUSIPCO) | 978-9-0827-9706-0/21/$31.00 ©202110.23919/EUSIPCO54536.2021.9616224

transform) estimation and nonnegative tensor factorization (NTF) [18] for the power spectrograms of the transformed domain. Since NTF is a multichannel extension of NMF, we can perform source number estimation with a gamma process [19]. This framework introduces nonnegative sparse variables for encouraging the shrinkage of redundant source spectrograms in a Bayesian manner. The proposed inference is formulated as a variational expectation-maximization (VEM) algorithm that iteratively and alternately updates the NTF parameters and the diagonalizer.

The main contribution of this study is to combine the state-of-the-art BSS method with Bayesian source number estimation. The existing BSS methods with source counting were based on the mixture models [6], [13], [15] or the factor model with the rank-1 spatial model [17]. The proposed method is based on FastMNMF, which is the factor model with the JD full-rank spatial model. Thanks to this spatial model, our method can work in more reasonable echoic and noisy conditions than the methods based on the rank-1 spatial model. We experimentally show that the proposed method can robustly separate sound sources even when the number of sound sources is not given in advance.

II. Background

Let an M-channel mixture signal $\mathbf{x}_{ft} \in \mathbb{C}^M$ include N source signals $s_{nft} \in \mathbb{C}$, the relationship between $\mathbf{x}_{ft}$ and s_{nft} can be formulated as follows:

$$\mathbf{x}_{ft} = \sum_{n=1}^{N} \mathbf{a}_{nf} s_{nft}, \tag{1}$$

where $\mathbf{a}_{nf} \in \mathbb{C}^M$ is a steering vector for source n, and $t = 1, \ldots, T$ and $f = 1, \ldots, F$ represent time and frequency indices, respectively. Since there is scale ambiguity between $\mathbf{a}_{nf}$ and s_{nft}, most of BSS methods estimate the source image $\mathbf{x}_{nft} = \mathbf{a}_{nf} s_{nft}$ from an input observed mixture $\mathbf{x}_{ft}$. A typical BSS approach is to estimate the source image based on a statistical model that combines a source model and a spatial model, representing the PSD of each source signal and the sound propagation process, respectively.

A. Source models

A popular formulation of a source model assumes that the source signal s_{nft} follows a complex Gaussian distribution [8]–[10]:

$$s_{nft} \sim \mathcal{N}_{\mathbb{C}}\left(0, \lambda_{nft}\right), \tag{2}$$

were $\lambda_{nft} \in \mathbb{R}_+$ represents the PSD of source n. The NMF model further assumes the PSD λ_{nft} to be low-rank with spectral basis vectors $\mathbf{w}_{nk} = [w_{nk1}, \ldots, w_{nkF}] \in \mathbb{R}_+^F$ and their activation vectors $\mathbf{h}_{nk} = [h_{nk1}, \ldots, h_{nkT}] \in \mathbb{R}_+^T$:

$$\lambda_{nft} = \sum_{k=1}^{K} w_{nkf} h_{nkt}, \tag{3}$$

where K represents the number of the spectral bases.

The number of bases K is an important parameter of NMF because K controls the complexity of the source spectrograms. Using a Bayesian non-parametric framework [19], we can automatically determine this parameter. Here, the key idea is to prepare a sufficiently large number of bases and encourage shrinkage of redundant bases by introducing a latent activation variable $z_{nk} \in \mathbb{R}_+$ as follows:

$$\lambda_{nft} = \sum_{k=1}^{K} z_{nk} w_{nkf} h_{nkt}. \tag{4}$$

This shrinkage is conducted by putting a sparse gamma prior on z_{nk} as follows:

$$z_{nk} \sim \mathrm{Gamma}\left(a^z/K, ca^z\right), \tag{5}$$

where $a^z \in \mathbb{R}_+$ is a hyperparameter to control how many basis vectors are likely to be active, and c is a scale parameter. In this model, w_{nkf} and h_{nkt} are also assumed to follow gamma distributions as conjugate priors of the Gaussian likelihood:

$$w_{nkf} \sim \mathrm{Gamma}\left(a^w, a^w\right), \quad h_{nkt} \sim \mathrm{Gamma}\left(a^h, a^h\right), \tag{6}$$

where $a^w \in \mathbb{R}_+$ and $a^h \in \mathbb{R}_+$ are hyperparameters to control the sparseness of w_{nkf} and h_{nkt}, respectively. As the number of bases K increases towards infinity, z_{nk} approximates a sample from a gamma process where the number of z_{nk} having large values follows a Poisson distribution [19].

B. Spatial models

From Eqs. (1) and (2), we obtain the following spatial model with a zero-mean multivariate Gaussian distribution:

$$\mathbf{x}_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \sum_{n=1}^{N} \lambda_{nft} \mathbf{H}_{nf}\right), \tag{7}$$

where $\mathbf{H}_{nf} = \mathbf{a}_{nf} \mathbf{a}_{nf}^{\mathsf{H}} \in \mathbb{S}_+$ is a rank-1 SCM for source n at frequency f. ILRMA assumes this spatial model and the NMF source model under the determined condition ($N = M$) to avoid the likelihood being a degenerate distribution. The inference of ILRMA can be efficiently conducted by the iterative projection (IP) rules [7] for $\mathbf{a}_{nf}$ and multiplicative update rules for the NMF parametersz. A Bayesian ILRMA utilizing the gamma process, introduced in Sec. II-A, is also proposed for automatically determining the number of basis vectors [20], [21].

The full-rank spatial model [8], which is used in MNMF [9], assumes the SCM $\mathbf{H}_{nf}$ to be a full-rank matrix. This model works with any number of sound sources N regardless of the number of microphones M. In addition, the full-rank SCM can handle fluctuation of the steering vectors (e.g., small echoic conditions) and diffuse noise. This model, however, takes heavy computational costs and often fails at bad local optima due to its high degree of freedom.

To efficiently reduce the number of parameters for an SCM while making its expression capability high enough, a JD full-rank spatial model has been proposed [10]–[12]. The model assumes that the SCMs for all the sources to be jointly diagonalizable as follows:

$$\mathbf{x}_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \sum_{n=1}^{N} \lambda_{nft} \mathbf{Q}_f^{-1} \mathrm{diag}(\mathbf{g}_{nf}) \mathbf{Q}_f^{-\mathsf{H}}\right), \tag{8}$$

where $\mathbf{Q}_f = [\mathbf{q}_{f1}, \ldots, \mathbf{q}_{fM}]^{\mathsf{H}} \in \mathbb{C}^{M \times M}$ is a diagonalizer, and $\mathbf{g}_{nf} \in \mathbb{R}_+^M$ is a nonnegative vector. The diagonalizer $\mathbf{Q}_f$ is estimated by the fast and stable IP algorithm, and the nonnegative vector $\mathbf{g}_{nf}$ is estimated with the multiplicative update rule. FastMNMF assumes this JD full-rank spatial model with the NMF source model [11], [12]. It is also reported that FastMNMF can improve the separation performance and initialization sensitivity by making $\mathbf{g}_{fn}$ into $\mathbf{g}_n \in \mathbb{R}_+^M$ for sharing the parameter over all the frequencies [12]. Importantly, FastMNMF can be regarded as the joint optimization of the diagonalization for the observed signal and the NTF on the transformed domain $\mathbf{Q}_f \mathbf{x}_{ft} \in \mathbb{C}^M$. Since the transformed spectrogram follows a multivariate Gaussian distribution with a diagonal covariance matrix, we obtain the likelihood function for g_{nm} and λ_{nft} as follows:

$$\mathbf{q}_{fm}^{\mathsf{H}} \mathbf{x}_{ft} \sim \mathcal{N}_{\mathbb{C}} \left(0, \sum_{n=1}^{N} g_{nm} \lambda_{nft} \right). \tag{9}$$

The maximization of this zero-mean complex Gaussian likelihood corresponds to the minimization of the Itakura-Saito divergence between $|\mathbf{q}_{fm}^{\mathsf{H}} \mathbf{x}_{ft}|^2$ and $\sum_{n=1}^{N} g_{nm} \lambda_{nft}$.

III. Gamma Process FastMNMF

The proposed method combines FastMNMF that has the powerful JD spatial model with a gamma process that can handle an unknown number of sources in a Bayesian manner. We first formulate the proposed generative model for a multichannel mixture recording and describe its inference based on a VEM algorithm.

A. Model formulation

To handle the unknown number of sources with a gamma process, we introduce a latent gain variable $z_n \in \mathbb{R}_+$ to Eq. (9) with a sufficiently large number N as follows:

$$\mathbf{q}_{fm}^{\mathsf{H}} \mathbf{x}_{ft} \sim \mathcal{N}_{\mathbb{C}} \left(\mathbf{0}, \sum_{n=1}^{N} z_n g_{nm} \lambda_{nft} \right). \tag{10}$$

The latent variable z_n is assumed to follow a sparse gamma distribution to construct a gamma process by taking $N \to \infty$:

$$z_n \sim \mathrm{Gamma}\left(a^z/N, c a^z\right), \tag{11}$$

where $a^z \in \mathbb{R}_+$ is a hyperparameter that controls how many sources are likely to be active, and $c \in \mathbb{R}_+$ is a scale parameter that is set to $\frac{1}{FTM} \sum_{f,t,m} |x_{ftm}|^2$. We also put a gamma distribution on g_{nm} as a conjugate prior:

$$g_{nm} \sim \mathrm{Gamma}\left(a^g, a^g\right) \tag{12}$$

where $a^g \in \mathbb{R}_+$ is the hyperparameter to control the sparseness of g_{nm}. As described in Sec. II-A, the number of sufficiently active sources approximately follows a Poisson distribution.

The PSD λ_{nft} is formulated with the NMF source model Eq. (3) with conjugate gamma priors as follows:

$$w_{nkf} \sim \mathrm{Gamma}\left(a^w, a^w\right), \tag{13}$$

$$h_{nkt} \sim \mathrm{Gamma}\left(a^h, a^h K\right), \tag{14}$$

where $a^w \in \mathbb{R}_+$ and $a^h \in \mathbb{R}_+$ are the hyperparameters to control the sparseness of w_{nkf} and h_{nkt}, respectively.

B. Variational expectation-maximization inference

The goal of our VEM inference is to estimate the posterior distribution of the model parameters $\mathbf{\Theta} = \{\mathbf{Z}, \mathbf{G}, \mathbf{W}, \mathbf{H}\}$ while the diagonalizer $\mathbf{Q}$ is obtained with maximum likelihood estimation. Since the true posterior distribution $p(\mathbf{\Theta}|\mathbf{X}, \mathbf{Q})$ is intractable, we approximately estimate the following variational posterior distribution $q(\mathbf{\Theta})$ by assuming the independence of the parameters:

$$p(\mathbf{Z}, \mathbf{G}, \mathbf{W}, \mathbf{H} \mid \mathbf{X}, \mathbf{Q}) \approx q(\mathbf{\Theta}) \triangleq q(\mathbf{Z}) q(\mathbf{G}) q(\mathbf{W}) q(\mathbf{H}). \tag{15}$$

This approximation is conducted by maximizing a lower bound of the log-marginal likelihood $p(\mathbf{X} \mid \mathbf{Q})$ called an evidence lower bound (ELBO) $\mathcal{L}$ as follows:

$$\mathcal{L} = \mathbb{E}_q[\log p(\mathbf{X} \mid \mathbf{\Theta}, \mathbf{Q})] - \mathcal{D}_{\mathrm{KL}}[q(\mathbf{\Theta}) \mid p(\mathbf{\Theta})], \tag{16}$$

where $\mathbb{E}_q[x]$ is an expectation of x by the variational distribution q, and $\mathcal{D}_{\mathrm{KL}}[q \mid p]$ represents the Kullback-Leibler (KL) divergence between q and p. This maximization corresponds to the minimization of the KL divergence between the variational and true posteriors. The proposed VEM algorithm iteratively and alternately updates the variational posterior $q(\mathbf{\Theta})$ at the E step and the diagonalizer $\mathbf{Q}$ at the M step. After obtaining the parameter estimates, the source images are extracted by the multichannel Wiener filtering [12].

1) Variational E step: By using the Jensen's inequality and the first-order Taylor approximation [19], the E step updates the variational posteriors $q(\mathbf{Z})$, $q(\mathbf{W})$, $q(\mathbf{H})$, and $q(\mathbf{G})$ to minimize $\mathcal{D}_{\mathrm{KL}}[q(\mathbf{\Theta}) \mid p(\mathbf{\Theta} \mid \mathbf{X}, \mathbf{Q})]$ as follows:

$$q(z_n) \leftarrow \mathrm{GIG}\left(a^z/N, \rho_n^z, \tau_n^z\right), \tag{17}$$

$$\rho_n^z = c a^z + \sum_{k,f,t,m} \frac{1}{\omega_{ftm}} \mathbb{E}_q\left[g_{nm} w_{nkf} h_{nkt}\right], \tag{18}$$

$$\tau_n^z = \sum_{k,f,t,m} \tilde{x}_{ftm} \psi_{nkftm}^2 \mathbb{E}_q\left[\frac{1}{g_{nm} w_{nkf} h_{nkt}}\right], \tag{19}$$

$$q(w_{nkf}) \leftarrow \mathrm{GIG}\left(a^w, \rho_{nkf}^w, \tau_{nkf}^w\right), \tag{20}$$

$$\rho_{nkf}^w = a^w + \sum_{t,m} \frac{1}{\omega_{ftm}} \mathbb{E}_q\left[z_n g_{nm} h_{nkt}\right], \tag{21}$$

$$\tau_{nkf}^w = \sum_{t,m} \tilde{x}_{ftm} \psi_{nkftm}^2 \mathbb{E}_q\left[\frac{1}{z_n g_{nm} h_{nkt}}\right], \tag{22}$$

$$q(h_{nkt}) \leftarrow \mathrm{GIG}\left(a^h, \rho_{nkt}^h, \tau_{nkt}^h\right), \tag{23}$$

$$\rho_{nkt}^h = a^h K + \sum_{f,m} \frac{1}{\omega_{ftm}} \mathbb{E}_q\left[z_n g_{nm} w_{nkf}\right], \tag{24}$$

$$\tau_{nkt}^h = \sum_{f,m} \tilde{x}_{ftm} \psi_{nkftm}^2 \mathbb{E}_q\left[\frac{1}{z_n g_{nm} w_{nkf}}\right], \tag{25}$$

$$q(g_{nm}) \leftarrow \mathrm{GIG}\left(a^g, \rho_{nm}^g, \tau_{nm}^g\right), \tag{26}$$

$$\rho_{nm}^g = a^g + \sum_{k,f,t} \frac{1}{\omega_{ftm}} \mathbb{E}_q\left[z_n w_{nkf} h_{nkt}\right], \tag{27}$$

$$\tau_{nm}^g = \sum_{k,f,t} \tilde{x}_{ftm} \psi_{nkftm}^2 \mathbb{E}_q\left[\frac{1}{z_n w_{nkf} h_{nkt}}\right], \tag{28}$$

where $\mathrm{GIG}\left(x;\gamma,\rho,\tau\right) \propto x^{\gamma-1}\exp(-\rho x - \tau/x)$ is the generalized inverse Gaussian (GIG) distribution, $\tilde{x}_{ftm}$ is the transformed power spectrogram $|\mathbf{q}_{fm}^{\mathsf{H}}\mathbf{x}_{ft}|^2 \in \mathbb{R}_{+}^{M}$, and ω_{ftm} and ψ_{nkftm} ($\sum_{n,k}\psi_{nkftm} = 1$) are auxiliary parameters obtained as follows:

$$\omega_{ftm} = \mathbb{E}_q\left[z_n g_{nm} w_{nkf} h_{nkt}\right], \tag{29}$$

$$\psi_{nkftm} \propto \mathbb{E}_q\left[\frac{1}{z_n g_{nm} w_{nkf} h_{nkt}}\right]^{-1}. \tag{30}$$

The expectations $\mathbb{E}_q[x]$ and $\mathbb{E}_q[\frac{1}{x}]$ with a random variable x following a GIG distribution can be calculated as described in [19].

2) Variational M step: By using the IP algorithm [7], the M step updates the diagonalizer $\mathbf{q}_{fm}$ such that the log-marginal likelihood is maximized:

$$\mathbf{V}_{fm} = \frac{1}{T}\sum_{n,k,t}\mathbf{x}_{ft}\mathbf{x}_{ft}^{\mathsf{H}}\psi_{nkftm}^2\mathbb{E}_q\left[\frac{1}{z_n g_{nm} w_{nkf} h_{nkt}}\right], \tag{31}$$

$$\mathbf{q}_{fm} \leftarrow \left(\mathbf{Q}_f\mathbf{V}_{fm}\right)^{-1}\mathbf{e}_m, \tag{32}$$

$$\mathbf{q}_{fm} \leftarrow \mathbf{q}_{fm}(\mathbf{q}_{fm}^{\mathsf{H}}\mathbf{V}_{fm}\mathbf{q}_{fm})^{-\frac{1}{2}}, \tag{33}$$

where $\mathbf{e}_m$ is a one-hot vector whose m-th element is one. Note that, in fact, the diagonalizer $\mathbf{Q}_f$ gradually converges to $\mathbf{0}$ because the other parameters, $\mathbf{Z}$, $\mathbf{G}$, $\mathbf{W}$, and $\mathbf{H}$, are assumed to be sparse, and the diagonalizer is updated with no prior distribution. As similar to the Bayesian ILRMA [20], [21], we solve this problem by a heuristic manner that $\mathbf{Q}_f$ is normalized such that the average value of $|\mathbf{Q}_f\mathbf{x}_{ft}|^2$ equals that of $|\mathbf{x}_{ft}|^2$. Although this normalization breaks the monotonically non-decreasing property of the VEM algorithm, we empirically confirmed that the proposed method separated an unknown number of sources properly.

IV. Experimental Evaluation

We report experimental results with multichannel speech mixture signals generated with simulated room impulse responses (RIRs).

A. Dataset

We generated multichannel signals by mixing $L \in \{2, 3, 4\}$ speech signals provided by WSJ0 English speech corpus. For each condition of L, we generated 100 mixture signals by using RIRs simulated with the image method [22]. The speech signals were mixed at random powers uniformly chosen between -2.5 dB and $+2.5$ dB. As an overdetermined condition, we assumed an $M = 8$ channel circular microphone array with a diameter of 8 cm. The array was placed on the center of the simulated room whose dimensions were 10 m × 10 m × 3 m. The sound sources were placed randomly around the array such that the horizontal angle differences between two sources from the array had at least 60°. The reverberation time (RT$_{60}$) was set to 200, 300, or 400 ms. We added Gaussian noise with a signal-to-noise ratio (SNR) of 30 dB. These signals were generated with a sampling rate of 16 kHz.

B. Experimental conditions

The hyperparameters of the proposed method were determined experimentally; a^z was set to 0.01, a^g and a^w were set to 10.0, and a^h was set to 5.0. The multichannel spectrograms were obtained by performing the short-time Fourier transform (STFT) with the window length of 1024 samples and a shifting interval of 256 samples. As inspired by [12], we iterated 50 times with $K = 2$ and then 150 times with $K = 32$ to avoid the frequency permutation problem. The maximum number of sources N (truncation level) was set to 8. The parameters were randomly initialized to be estimated.

We compared the proposed method (GaP-FastMNMF) with ILRMA having the rank-1 spatial model and FastMNMF having the JD full-rank spatial model. The number of bases vectors K for both ILRMA and FastMNMF was set to 2, which gave the best performance in $K \in \{2, 4, 8, 16, 32\}$. The parameters were updated for 200 times. The FastMNMF was evaluated with the number of sources $N \in \{2, 3, \ldots, M\}$.

The separation performance was evaluated by using the scale-invariant source-to-distortion ratio (SI-SDR) [23]. The SI-SDRs for a mixture were evaluated with top L sources having large powers in the N separated signals. We also evaluated the source counting accuracy (the percentage of correct estimates). The source counting was performed by counting the number of estimated sources that had more power than -15 dB from the average power of the observed signal.

C. Experimental results

The separation performances are summarized in Fig. 2. First, we can see that the SI-SDR of FastMNMF changed drastically according to the number of sources N. The best values of N for FastMNMF were between $L + 2$ and $L + 4$. It is difficult to appropriately configure this parameter N in advance because the best configuration is affected by not only the actual number of sources but also the reverberation and diffuse noise [12], [24]. In contrast, the proposed GaP-FastMNMF robustly separated the unknown number of sound sources. We can see that the SI-SDR of the GaP-FastMNMF was comparable to that of the FastMNMF having the best configuration of N.

The performances of source counting are summarized in Fig. 3. GaP-FastMNMF achieved the best counting performance when the RT$_{60}$ was 200 ms. The counting accuracy of GaP-FastMNMF at this condition was more than 90 %. On the other hand, the counting performance deteriorated as the reverberation increased. This could be overcome by combining the proposed GaP-FastMNMF with dereverberation frameworks such as autoregressive models [25].

V. Conclusion

This paper presented a BSS method that can separate an unknown number of sound sources in a Bayesian manner. The proposed method is based on FastMNMF, which has the JD full-rank spatial model and the NMF source model. We handle the unknown number of sound sources by introducing latent source activation variables for making redundant source

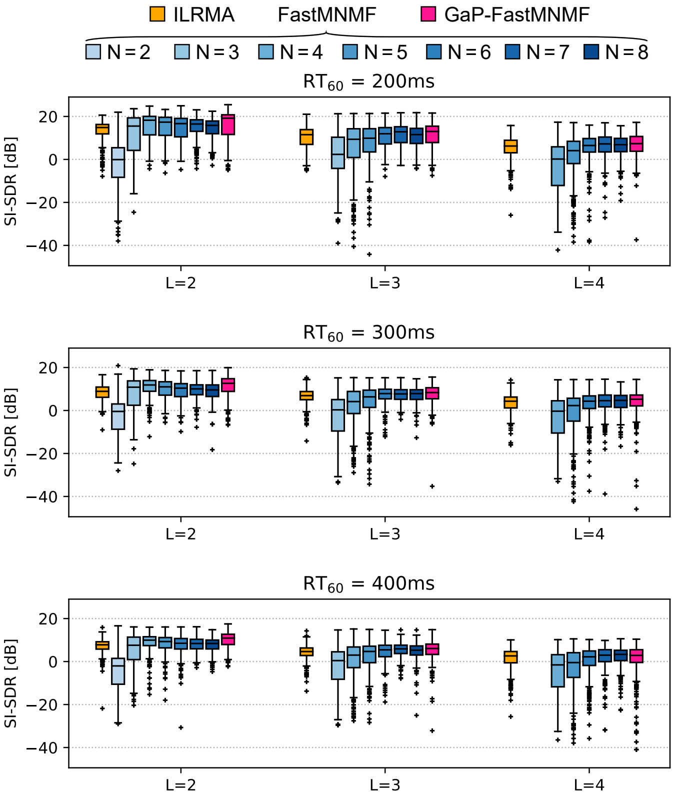

Fig. 2. Box plots of SI-SDRs for the separated signals by ILRMA, FastMNMF ($N = 2, \ldots, 8$), and GaP-FastMNMF.

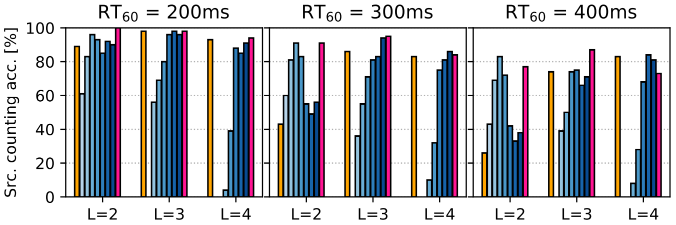

Fig. 3. Source counting accuracy of ILRMA, FastMNMF ($N = 2, \ldots, 8$), and GaP-FastMNMF.

classes shrink. The inference is conducted by a VEM algorithm to jointly estimate the diagonalizer and the posterior distributions of the latent variables. We demonstrated that the proposed GaP-FastMNMF robustly separated an unknown number of sound sources while the conventional FastMNMF required careful tuning depending on the actual number of sources in a mixture.

Our future work includes utilizing the beta process [17] that can explicitly count the number of sound sources by introducing a binary activation variable to the source spectrogram. We also plan to extend our method to a nested non-parametric model to jointly estimate the number of sound sources and that of bases vectors for NMF. This extension will enable the method to automatically determine all the model complexity according to the observation.

Acknowledgments

This study was partially supported by JST ACT-X JPMJAX200N and JSPS KAKENHI 19H04137.

References

[1] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE TASLP*, vol. 18, no. 3, pp. 550–563, 2009.

[2] D. Kitamura *et al.*, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM TASLP*, vol. 24, no. 9, pp. 1626–1641, 2016.

[3] A. A. Nugraha *et al.*, "Multichannel audio source separation with deep neural networks," *IEEE/ACM TASLP*, vol. 24, no. 9, pp. 1652–1664, 2016.

[4] N. Ito *et al.*, "Complex angular central Gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *EUSIPCO*, 2016, pp. 1153–1157.

[5] Y. Kubo *et al.*, "Mask-based MVDR beamformer for noisy multisource environments: introduction of time-varying spatial covariance model," in *IEEE ICASSP*, 2019, pp. 6855–6859.

[6] J. Taghia and A. Leijon, "Separation of unknown number of sources," *IEEE SPL*, vol. 21, no. 5, pp. 625–629, 2014.

[7] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *IEEE WASPAA*, 2011, pp. 189–192.

[8] N. Q. K. Duong *et al.*, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE TASLP*, vol. 18, no. 7, pp. 1830–1840, 2010.

[9] H. Sawada *et al.*, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE TASLP*, vol. 21, no. 5, pp. 971–982, 2013.

[10] N. Ito *et al.*, "FastFCA: A joint diagonalization based fast algorithm for audio source separation using a full-rank spatial covariance model," in *EUSIPCO*, 2018, pp. 151–155.

[11] N. Ito and T. Nakatani, "FastMNMF: Joint diagonalization based accelerated algorithms for multichannel nonnegative matrix factorization," in *IEEE ICASSP*, 2019, pp. 371–375.

[12] K. Sekiguchi *et al.*, "Fast multichannel nonnegative matrix factorization with directivity-aware jointly-diagonalizable spatial covariance matrices for blind source separation," *IEEE/ACM TASLP*, vol. 28, pp. 2610–2625, 2020.

[13] H. Kameoka *et al.*, "Bayesian nonparametric approach to blind separation of infinitely many sparse sources," *IEICE TFECCS*, vol. 96, no. 10, pp. 1928–1937, 2013.

[14] J. Taghia and A. Leijon, "Variational inference for watson mixture model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 9, pp. 1886–1900, 2015.

[15] S. Araki *et al.*, "Blind sparse source separation for unknown number of sources using Gaussian mixture model fitting with Dirichlet prior," in *IEEE ICASSP*, 2009, pp. 33–36.

[16] O. Walter *et al.*, "Source counting in speech mixtures by nonparametric bayesian estimation of an infinite gaussian mixture model," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 459–463.

[17] K. Nagira *et al.*, "Nonparametric Bayesian sparse factor analysis for frequency domain blind source separation without permutation ambiguity," *EURASIP JASMP*, vol. 2013, no. 1, p. 4, 2013.

[18] Y. Mitsufuji *et al.*, "Multichannel blind source separation based on non-negative tensor factorization in wavenumber domain," in *IEEE ICASSP*, 2016, pp. 56–60.

[19] M. D. Hoffman *et al.*, "Bayesian nonparametric matrix factorization for recorded music," in *ICML*, 2010, pp. 439–446.

[20] C. Narisetty, "A unified Bayesian source modelling for determined blind source separation." in *INTERSPEECH*, 2019, pp. 1343–1347.

[21] C. Narisetty *et al.*, "Bayesian non-parametric multi-source modelling based determined blind source separation," in *ICASSP*, 2019, pp. 111–115.

[22] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *JASA*, vol. 65, no. 4, pp. 943–950, 1979.

[23] J. Le Roux *et al.*, "SDR–half-baked or well done?" in *IEEE ICASSP*, 2019, pp. 626–630.

[24] K. Sekiguchi *et al.*, "Fast multichannel source separation based on jointly diagonalizable spatial covariance matrices," in *EUSIPCO*, 2019, pp. 1–5.

[25] ——, "Autoregressive fast multichannel nonnegative matrix factorization for joint blind source separation and dereverberation," in *IEEE ICASSP*, 2021, accepted.