A Unified Bayesian Model of Time-frequency Clustering and Low-rank Approximation for Multi-channel Source Separation

Kousuke Itakura, Yoshiaki Bando, Eita Nakamura, Katsutoshi Itoyama, Kazuyoshi Yoshii
Graduate School of Informatics, Kyoto University, Japan

***Abstract*—This paper presents a statistical method of multi-channel source separation, called NMF-LDA, that unifies nonnegative matrix factorization (NMF) and latent Dirichlet allocation (LDA) in a hierarchical Bayesian manner. If the frequency components of sources are sparsely distributed, the source spectrograms can be considered to be disjoint with each other in most time-frequency bins. Under this assumption, LDA has been used for clustering time-frequency bins into individual sources using spatial information. A way to improve LDA-based source separation is to consider the empirical fact that source spectrograms tend to have low-rank structure. To leverage both the sparseness and low-rankness of source spectrograms, our method iterates an LDA-step (hard clustering of time-frequency bins) that gives deficient source spectrograms and an NMF-step (low-rank matrix approximation) that completes the deficient bins of those spectrograms. Experimental results showed the proposed method outperformed conventional methods.**

I. Introduction

Microphone array processing forms the basis of computational auditory scene analysis that aims to understand individual auditory events in a sound mixture. One promising approach to multi-channel source separation is time-frequency (TF) clustering [1]–[4]. If the frequency components of each source are sparsely distributed, as is often the case with harmonic sounds, the source spectrograms can be considered to be disjoint with each other in most TF bins, *i.e.*, one of the sources is dominant at each bin. This assumption, called W-disjoint orthogonality [5], is reasonable because the additivity of source spectrograms does not hold exactly and a loud sound masks softer sounds at each TF bin. Under this assumption, Otsuka *et al.* [4] proposed a Bayesian mixture model inspired by latent Dirichlet allocation (LDA) [6] for clustering TF bins into sources at the same time as clustering those sources into different directions. Such unified source separation and localization can circumvent the permutation problem of conventional frequency-domain separation methods such as independent component analysis (ICA) [7].

Mainly in the context of single-channel source separation, nonnegative matrix factorization (NMF) has gained a lot of attention [8]. It approximates the power spectrogram of an observed mixture signal as the product of a basis matrix (a set of basis spectra) and an activation matrix (a set of temporal activations). For multi-channel source separation, multi-channel extensions of NMF (MNMF) were proposed [9], [10]. MNMF decomposes the complex spectrograms of mixture signals into basis spectra, temporal activations, and spatial information. To reduce the sensitivity of MNMF to parameter initialization,

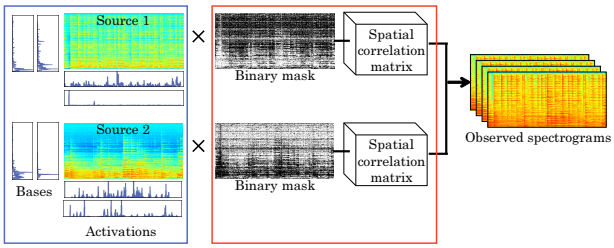

Fig. 1. The generative story of NMF-LDA. Source spectrograms are generated from bases and activations. Masks denote which source is dominant in each TF bin. The masked source spectrograms are given phases according to spatial correlation matrices and then gathered to yield observed spectrograms.

Kitamura *et al.* [11] proposed a method that restricts the spatial correlation matrices to rank-1 matrices. This rank-1 MNMF can be considered a model unifying NMF and independent vector analysis (IVA) [12], [13]. Although MNMF can leverage the low-rankness of source spectrograms, their sparseness is not taken into account because the sources in every TF bin are allowed to be active simultaneously.

In this paper we propose a hierarchical Bayesian model, called NMF-LDA, that improves multi-channel source separation by leveraging both the sparseness and low-rankness of source spectrograms. As illustrated in Fig. 1, the complex spectrograms of observed multi-channel mixture signals are generated as follows: the power spectrogram of each source is stochastically determined as the product of a basis matrix and an activation matrix and then the phases of the mixture signals at each TF bin are stochastically determined according to the spatial correlation matrix of the dominant source at that bin. Using Gibbs sampling, the basis and activation matrices are estimated in the framework of NMF at the same time as a dominant source is identified at each TF bin and the spatial correlation matrix is estimated in the framework of LDA.

II. Related Work

A standard approach to multi-channel source separation is to estimate a linear "unmixing" filter that separates the complex spectra of mixture signals into those of source signals in the frequency domain [7], [12]–[14]. Mixture signals are usually modeled as the sum of source signals convolved with the impulse responses of the corresponding source directions. This is equivalent to an instantaneous mixing process in the frequency domain, *i.e.,* the complex spectra of mixture signals are the sum of source spectra multiplied with impulse-response spectra. Using such linearity between mixture and source spectra,

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

frequency-domain ICA can estimate a linear unmixing filter at each frequency bin [7]. The permutation of separated source spectra, however, is not aligned between different frequency bins. One possibility to solve the permutation ambiguity is to focus on the directions and inter-frequency correlations of sources [14]. IVA [12], [13] is an extension of ICA that can jointly deal with all frequency components in a vectorial manner. These methods can be used only under a determined or overdetermined condition that the number of sources is equal to or less than that of microphones. A promising way to cope with the underdetermined condition is to focus on the sparseness of source spectra [15].

Another popular approach to source separation is to perform nonlinear time-frequency masking based on the sparseness (disjointness) of source spectra [1]–[4]. If each TF bin is independently classified into one of sound sources [3], the permutation ambiguity arises as in ICA. To solve this problem, Otsuka *et al.* [4] proposed a statistical method that jointly clusters each TF bin into different sources and directions. This method can work well under an underdetermined condition.

III. Conventional Methods

Let us consider that K sources are observed with M microphones. Each TF bin in the complex spectrograms of observed and source signals is defined as follows:

$$\boldsymbol{x}_{tf} = [x_{tf1}, \cdots, x_{tfM}]^T \in \mathbb{C}^M, \tag{1}$$

$$\boldsymbol{y}_{tf} = [y_{tf1}, \cdots, y_{tfK}]^T \in \mathbb{C}^K. \tag{2}$$

Assuming an instantaneous mixing process in the frequency domain, the observation $\boldsymbol{x}_{tf}$ is represented as

$$\boldsymbol{x}_{tf} = \sum_{k=1}^{K} \boldsymbol{a}_{fk} y_{tfk}, \tag{3}$$

where $\boldsymbol{a}_{fk}$ is a steering vector of source k at frequency f.

A. Mixture Modeling Approach: Latent Dirichlet Allocation

The LDA-based method [4] simultaneously performs source separation and localization. Let D be the number of possible directions (angles), *e.g.*, $D = 72$ if 360 degrees are discretized with an interval of 5 degrees. Using the sparseness of source spectrograms, Eq. (3) is replaced with

$$\boldsymbol{x}_{tf} = \sum_{k,d=1}^{K,D} z_{tfk} s_{kd} \boldsymbol{a}_{fd} y_{tfk}, \tag{4}$$

where $\boldsymbol{a}_{fd} \in \mathbb{C}^M$ is a steering vector of direction d at frequency f. z_{tfk} takes 1 when source k is dominant at frame t and frequency f and otherwise takes 0. Similarly, s_{kd} takes 1 when source k exists in direction d and otherwise takes 0.

A standard way to represent the complex spectrum y_{tfk} of source k is to use a complex Gaussian distribution as follows:

$$y_{tfk} \mid \lambda_{tfk} \sim \mathcal{N}_{\mathbb{C}}(y_{tfk} \mid 0, \lambda_{tfk}), \tag{5}$$

where λ_{tfk} is the power spectrum density of source k. Using Eq. (4) and Eq. (5), the complex spectrum $\boldsymbol{x}_{tf}$ is found to follow a mixture of complex Gaussian distributions as follows:

$$\boldsymbol{x}_{tf} \mid \boldsymbol{\lambda}, \boldsymbol{G}, \boldsymbol{Z}, \boldsymbol{S} \sim \prod_{k,d=1}^{K,D} \mathcal{N}_{\mathbb{C}}\left(\boldsymbol{x}_{tf} \;\middle|\; \boldsymbol{0}, \lambda_{tfk} \boldsymbol{G}_{fd}^{-1}\right)^{z_{tfk} s_{kd}}, \tag{6}$$

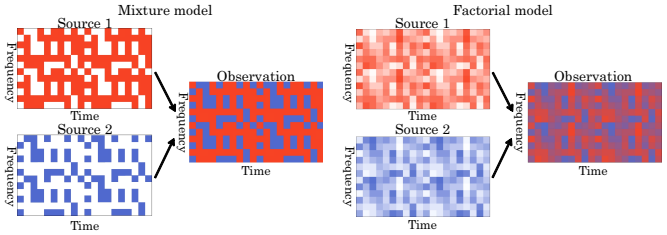

Fig. 2. Comparison of a mixture model and a factorial model. In the mixture model, one of sources is stochastically selected at each TF bin. In the factorial model, all sources are accumulated in a weighed manner at each TF bin.

where $\boldsymbol{G}_{fd}^{-1} = \boldsymbol{a}_{fd} \boldsymbol{a}_{fd}^H$ is a spatial correlation matrix for direction d at frequency f.

B. Factorial Modeling Approach: Multichannel NMF

Multichannel NMF (MNMF) [9] can perform blind source separation. Unlike the LDA-based method, all sources are assumed to be activated in a weighted manner at every TF bin. Assuming Eq. (5) as in the LDA-based method and using Eq. (3), the complex spectrum $\boldsymbol{x}_{tf}$ is found to follow a complex Gaussian distribution as follows:

$$\boldsymbol{x}_{tf} \mid \boldsymbol{\lambda}, \boldsymbol{G} \sim \mathcal{N}_{\mathbb{C}}\left(\boldsymbol{x}_{tf} \;\middle|\; \boldsymbol{0}, \textstyle\sum_{k=1}^{K} \lambda_{tfk} \boldsymbol{G}_{fk}^{-1}\right), \tag{7}$$

where $\boldsymbol{G}_{fk}^{-1} = \boldsymbol{a}_{fk} \boldsymbol{a}_{fk}^H$ is a spatial correlation matrix for source k at frequency f.

A key idea of MNMF is to decompose the power spectrum density λ_{tfk} of source k by using a low-rank approximation technique (*e.g.*, NMF and canonical polyadic decomposition [16]) as follows:

$$\lambda_{tfk} = \sum_{l=1}^{L} u_{lk} w_{lf} h_{lt}, \tag{8}$$

where w_{lf} is the power of the l-th basis spectrum at frequency f, h_{lt} is the volume of the l-th basis at frame t, and u_{lk} is a contribution of the l-th basis to source k such that $\sum_{k=1}^{K} u_{lk} = 1$. Comparing the likelihood of LDA given by Eq. (6) with that of MNMF Eq. (7), we can see the clear difference between mixture and factorial modeling approaches (Fig.2).

IV. Proposed Method

This section explains a hybrid of factorial and mixture models called NMF-LDA that integrates TF clustering based on LDA with low-rank approximation of source spectrograms based on NMF (Fig. 1). Our model was inspired by the LDA-based method [4] and MNMF [9]. The sparseness of source spectrograms justifies an assumption that only one of the sources is activated at each TF bin. This calls for mixture modeling that stochastically chooses one of sources at each bin. Using the low-rankness of source spectrograms, the power spectrogram of each source is approximated by the product of basis spectra and temporal activations in a way of factorial modeling.

A. Model Formulation

The proposed model consists of LDA and NMF parts that are integrated in a hierarchical Bayesian manner.

1) LDA Part: According to the LDA-based model described in Section III-A, the likelihood function for the observation $\boldsymbol{x}_{tf}$ is given by Eq. (6). The total likelihood function over all frames and all frequencies is thus given by

$$\boldsymbol{X}|\boldsymbol{\lambda},\boldsymbol{G},\boldsymbol{Z},\boldsymbol{S} \sim \prod_{t,f,k,d=1}^{T,F,K,D} \mathcal{N}_{\mathbb{C}}(\boldsymbol{x}_{tf}|\boldsymbol{0},\lambda_{tfk}\boldsymbol{G}_{fd}^{-1})^{z_{tfk}s_{kd}}. \quad (9)$$

As in the LDA-based method [4], the latent variables $\boldsymbol{Z}$ and $\boldsymbol{S}$ are drawn from the following categorical distributions:

$$\boldsymbol{z}_{tf} \mid \boldsymbol{\pi}_t \sim \mathrm{Categorical}(\boldsymbol{z}_{tf} \mid \boldsymbol{\pi}_t), \quad (10)$$

$$\boldsymbol{s}_k \mid \boldsymbol{\phi} \sim \mathrm{Categorical}(\boldsymbol{s}_k \mid \boldsymbol{\phi}). \quad (11)$$

For mathematical convenience, conjugate prior distributions are put on the model parameters $\boldsymbol{\pi}$, $\boldsymbol{\phi}$, and $\boldsymbol{G}$ as follows:

$$\boldsymbol{\pi}_t \sim \mathrm{Dirichlet}(\boldsymbol{\pi}_t \mid a_0^{\pi}\boldsymbol{1}_K), \quad (12)$$

$$\boldsymbol{\phi} \sim \mathrm{Dirichlet}(\boldsymbol{\phi} \mid a_0^{\phi}\boldsymbol{1}_D), \quad (13)$$

$$\boldsymbol{G}_{fd} \sim \mathcal{W}_{\mathbb{C}}(\boldsymbol{G}_{fd} \mid \nu, \boldsymbol{G}_{fd}^0), \quad (14)$$

where $\boldsymbol{1}_N$ is a N-dimensional vector with all entries one and $\mathcal{W}_{\mathbb{C}}$ is a complex Wishart distribution(see Appendix). Note that estimation of source indicators $\boldsymbol{Z}$ over all TF bins and estimation of direction indicators $\boldsymbol{S}$ over K sources correspond to source separation and localization, respectively.

2) NMF Part: The power spectrum density λ_{tfk} in Eq. (9) is factorized in a similar way to MNMF described in Section III-B except for two ways. Unlike Eq. (8), λ_{tfk} is considered as a random variable drawn from a gamma distribution with a factorized scale parameter as follows:

$$\lambda_{tfk} \mid \boldsymbol{W}_k, \boldsymbol{H}_k \sim \mathrm{Gamma}\left(\lambda_{tfk} \,\middle|\, \alpha, \frac{\alpha}{\sum_{l=1}^{L} w_{klf}h_{klt}}\right), \quad (15)$$

where $\mathbb{E}[\lambda_{tfk}|\boldsymbol{W}_k,\boldsymbol{H}_k] = \sum_{l=1}^{L} w_{klf}h_{klt}$ and α is a hyperparameter controlling how likely λ_{tfk} is to be close to an exact low-rank structure ($\lambda_{tfk} \to \sum_{l=1}^{L} w_{klf}h_{klt}$ when $\alpha \to \infty$). Another modification is that the power spectrum density of each source k is represented with a unique set of basis spectra unlike Eq. (8). These modifications enable our model to flexibly represent a wide variety of sound spectrograms.

For mathematical convenience, conjugate prior distributions are put on the model parameters $\boldsymbol{W}$ and $\boldsymbol{H}$ as follows:

$$w_{klf} \sim \mathrm{Gamma}(w_{klf} \mid a_0^w, b_0^w), \quad (16)$$

$$h_{klt} \sim \mathrm{Gamma}(h_{klt} \mid a_0^h, b_0^h), \quad (17)$$

where a_0^* and b_0^* are hyperparameters.

B. Posterior Inference

Our goal is to find optimal model parameters such that the posterior probability $p(\boldsymbol{G},\boldsymbol{Z},\boldsymbol{S},\boldsymbol{\pi},\boldsymbol{\phi},\boldsymbol{\lambda},\boldsymbol{W},\boldsymbol{H}|\boldsymbol{X})$ is maximized. Since the true posterior is analytically intractable, we use partially-collapsed Gibbs sampling after marginalizing out $\boldsymbol{\pi}$ and $\boldsymbol{\phi}$. As shown in Fig. 3, LDA and NMF are iterated until the likelihood converges. $\boldsymbol{G}$, $\boldsymbol{Z}$, and $\boldsymbol{S}$ are updated with LDA and $\boldsymbol{\lambda}$, $\boldsymbol{W}$, and $\boldsymbol{H}$ are updated with NMF. Given the optimal parameters, frequency-domain source signals can be recovered using multichannel wiener filtering [9].

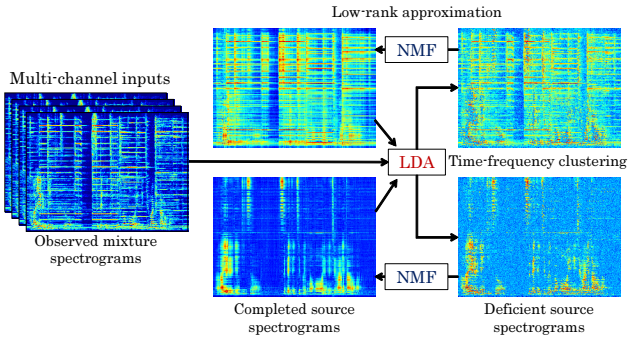

Fig. 3. Overview of the proposed iterative optimization algorithm.

1) Updating LDA Part: $\boldsymbol{G}$, $\boldsymbol{Z}$, and $\boldsymbol{S}$ are alternately sampled from conditional posterior distributions given by

$$\boldsymbol{G}_{fd} \mid \boldsymbol{X}, \boldsymbol{\Theta}_{\neg \boldsymbol{G}_{fd}} \sim \mathcal{W}_{\mathbb{C}}(\boldsymbol{G}_{fd} \mid \nu'_{fd}, \boldsymbol{G}'_{fd}), \quad (18)$$

$$\boldsymbol{z}_{tf} \mid \boldsymbol{X}, \boldsymbol{\Theta}_{\neg \boldsymbol{z}_{tf}} \sim \mathrm{Categorical}(\boldsymbol{z}_{tf} \mid \boldsymbol{\pi}'_{tf}), \quad (19)$$

$$\boldsymbol{s}_k \mid \boldsymbol{X}, \boldsymbol{\Theta}_{\neg \boldsymbol{s}_k} \sim \mathrm{Categorical}(\boldsymbol{s}_k \mid \boldsymbol{\phi}'_k), \quad (20)$$

where $\boldsymbol{\Theta}$ is a set of all parameters and $\boldsymbol{\Theta}_{\neg *}$ indicates a set of all parameters excluding $*$. The conditional posterior parameters ν'_{fd}, $\boldsymbol{G}'_{fd}$, $\boldsymbol{\pi}'_{tf}$, and $\boldsymbol{\phi}'_k$ are given by

$$\nu'_{fd} = \nu + \textstyle\sum_{t,k=1}^{T,K} z_{tfk}s_{kd}, \quad (21)$$

$$\boldsymbol{G}'^{-1}_{fd} = (\boldsymbol{G}^0_{fd})^{-1} + \textstyle\sum_{t,k=1}^{T,K} \frac{\boldsymbol{x}_{tf}\boldsymbol{x}_{tf}^H}{\lambda_{tfk}} z_{tfk}s_{kd}, \quad (22)$$

$$\pi'_{tfk} = (a_0^{\pi} + n_{tk}^{\neg tf}) \textstyle\prod_{d=1}^{D} \left\{ \left| \frac{\boldsymbol{G}_{fd}}{\lambda_{tfk}} \right| \exp\left(-\frac{\boldsymbol{x}_{tf}^H \boldsymbol{G}_{fd} \boldsymbol{x}_{tf}}{\lambda_{tfk}} \right) \right\}^{s_{kd}}, \quad (23)$$

$$\phi'_{kd} = (a_0^{\phi} + c_d^{\neg k}) \textstyle\prod_{t,f=1}^{T,F} \left\{ \left| \frac{\boldsymbol{G}_{fd}}{\lambda_{tfk}} \right| \exp\left(-\frac{\boldsymbol{x}_{tf}^H \boldsymbol{G}_{fd} \boldsymbol{x}_{tf}}{\lambda_{tfk}} \right) \right\}^{z_{tfk}}, \quad (24)$$

where $n_{tk}^{\neg tf}$ indicates the number of TF bins assigned to source k without a sample at frame t and frequency f, and $c_d^{\neg k}$ is the number of sources assigned to direction d without source k

2) Updating NMF Part: $\boldsymbol{\lambda}$, $\boldsymbol{W}$, and $\boldsymbol{H}$ are alternately sampled by using a non-collapsed Gibbs sampler and a Metropolis-Hastings (MH) algorithm [17]. Since the gamma distribution (Eq. (15)) is a special case of the generalized inverse Gaussian (GIG) distribution [18] (see Appendix) and the GIG distribution is a conjugate prior for the Gaussian distribution (Eq. (9)), $\boldsymbol{\lambda}$ can be analytically sampled as follows:

$$\lambda_{tfk} \mid \boldsymbol{X}, \boldsymbol{\Theta}_{\neg \lambda_{tfk}} \sim \mathrm{GIG}(\lambda_{tfk} \mid \gamma_{tfk}, \rho_{tfk}, \tau_{tfk}), \quad (25)$$

where γ_{tfk}, ρ_{tfk}, and τ_{tfk} are given by

$$\gamma_{tfk} = \alpha - Mz_{tfk}, \quad \rho_{tfk} = \frac{\alpha}{\sum_{l=1}^{L} w_{klf}h_{klt}}, \quad (26)$$

$$\tau_{tfk} = \textstyle\sum_{d=1}^{D} \boldsymbol{x}_{tf}\boldsymbol{G}_{fd}\boldsymbol{x}_{tf} z_{tfk}s_{kd}. \quad (27)$$

Since the true conditional posterior distributions of $\boldsymbol{W}$ and $\boldsymbol{H}$ are analytically intractable, the MH algorithm [17] is used for sampling $\boldsymbol{W}$ and $\boldsymbol{H}$ instead of Gibbs sampling. The key of the MH algorithm is the design of a proposal distribution that stochastically generates a candidate of a next sample based on

a previous sample. If the proposal is close to the true posterior, the candidate is set to the next sample with a high acceptance ratio. Otherwise, the previous value is set to the next sample. In this study, approximated posterior distributions of $\boldsymbol{W}$ and $\boldsymbol{H}$ are calculated by variational Bayesian (VB) inference and then used as proposal distributions for MH sampling.

Although the GIG distribution can be a conjugate prior on the scale parameter of the gamma distribution, neither Eq. (16) nor Eq. (17) has no direct conjugacy with Eq. (15) because the sum operation is involved in the scale parameter. The complete log-likelihood given by Eq. (15), Eq. (16), and Eq. (17) is thus lower bounded by an auxiliary function $\mathcal{L}$, using Jensen's inequality and a first-order Taylor approximation in the same way as Bayesian NMF [19] as follows:

$$\begin{aligned}\mathcal{L} = &- \textstyle\sum_{t,f,k=1}^{T,F,K} \lambda_{tfk} \sum_{l=1}^{L} \frac{\psi_{lktf}}{w_{kfl}h_{klt}} - \log(\omega_{ktf}) \\ &- \textstyle\sum_{l=1}^{L} \frac{w_{kfl}h_{klt}}{\omega_{ktf}} - \log \frac{q(w_{kfl})q(h_{klt})}{p(w_{kfl})p(h_{klt})},\end{aligned} \tag{28}$$

where ϕ_{lktf} and ω_{ktf} are auxiliary variables. The original log-likelihood can be recovered by maximizing the lower bound $\mathcal{L}$ with respect to ϕ_{lktf} and ω_{ktf} as follows:

$$\psi_{lktf} = \frac{w_{kfl}h_{klt}}{\sum_{l=1}^{L} w_{kfl}h_{klt}}, \quad \omega_{ktf} = \textstyle\sum_{l=1}^{L} w_{kfl}h_{klt}. \tag{29}$$

Since the GIG-gamma conjugacy is recovered in $\mathcal{L}$, the variational posterior distributions of $\boldsymbol{W}$ and $\boldsymbol{H}$ used as proposal distributions for MH sampling are obtained as follows:

$$q(w_{kfl}) = \mathrm{GIG}(w_{kfl} \mid a_0^w, \rho_{kfl}^w, \tau_{kfl}^w), \tag{30}$$

$$q(h_{klt}) = \mathrm{GIG}(h_{klt} \mid a_0^h, \rho_{klt}^h, \tau_{klt}^h), \tag{31}$$

where ρ_{kfl}^w, τ_{kfl}^w, ρ_{klt}^h and τ_{klt}^h are given by

$$\rho_{kfl}^w = b_0^w + \textstyle\sum_{t=1}^{T} \frac{h_{klt}}{\omega_{ktf}}, \ \tau_{kfl}^w = \sum_{t=1}^{T} \frac{\psi_{lktf}^2 \lambda_{tfk}}{h_{klt}}, \tag{32}$$

$$\rho_{klt}^h = b_0^h + \textstyle\sum_{f=1}^{F} \frac{w_{kfl}}{\omega_{ktf}}, \ \tau_{klt}^h = \sum_{f=1}^{F} \frac{\psi_{lktf}^2 \lambda_{tfk}}{w_{kfl}}. \tag{33}$$

V. Evaluation

This section presents source separation results obtained with simulated convolutive mixture signals. The experiments were conducted in underdetermined conditions (the number of microphones $M <$ the number of sources K) and overdetermined conditions ($M > K$). The proposed NMF-LDA was compared with IVA [12], MNMF [9] and the LDA-based method (simply written as LDA in this paper) [4] in overdetermined conditions and with MNMF and LDA in underdetermined conditions because IVA can not be used in underdetermined conditions. Although the original model of LDA [4] can estimate the number of sources, for fair evaluation we conducted experiments under the condition that the number of sources is fixed. More specifically, Eq. (15) is replaced with:

$$p(\lambda_{tfk}) = \mathrm{Gamma}(\lambda_{tfk}|1, 1). \tag{34}$$

A. Experimental Conditions

Figure 4 shows the locations of microphones and sources. Three sources were convoluted using impulse responses measured in a room where the reverberation time RT_{60} was 400 ms. The number of microphones M was 2 or 4; mic. 2 and mic. 4 (shown in black) were used when $M = 2$ and all microphones were used when $M = 4$. 30 mixtures were used for evaluation; 10 were mixtures of music signals (including guitar, bass, vocal, hi-hat, piano sounds), 10 were mixtures of speech signals, and 10 were mixtures of music and speech signals. The music and speech signals were selected from the SiSEC data set [20] and the JNAS phonetically balanced Japanese utterances [21], respectively. The audio signals were sampled at 16000 Hz and a short-time Fourier transformation was carried out with a 512 pt Hamming window and a 256 pt shift size. The steering vectors $\hat{\boldsymbol{a}}_{fd}$ were measured in an anechoic room such that $D = 72$ with 5° resolution. Hyperparameters were set as follows: $\nu = M$, $\boldsymbol{G}_{fd}^0 = (\hat{\boldsymbol{a}}_{fd}\hat{\boldsymbol{a}}_{fd}^H + 0.01 \times \boldsymbol{I})^{-1}$, $L = 20$, $a_0^\pi = a_0^\phi = 10$, $a_0^w = b_0^w = a_0^h = 1$, $b_0^h = L$, $\alpha = 10$. Signal-to-distortion ratio (SDR) [22] was used to evaluate separation performances. The larger the SDR, the better the separation performance.

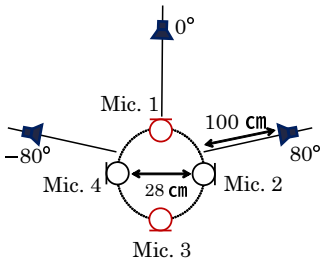

Fig. 4. Locations of microphones and sources.

B. Experimental Results

Figures 5 and 6 show the SDR improvements. In all conditions, NMF-LDA achieved best separation performance of all compared methods. For speech data, NMF-LDA slightly outperformed LDA. For music data, although LDA was inferior to MNMF, NMF-LDA achieved better performance than MNMF. This indicates that the low-rankness of music spectrograms is much stronger than that of speech spectrograms.

Tables I and II show the average SDR improvements. The average SDR improvement obtained by NMF-LDA was at most 3.3 dB greater than that obtained by LDA. While in the case of two microphones NMF-LDA archived much better performance than LDA, in the case of four microphones the performance difference was small. This indicates that the low-rankness of source spectrograms effectively helps estimate the power spectrogram density when a fewer number of microphones are available.

Although steering vectors are required for setting appropriate prior distributions on spatial correlation matrices in NMF-LDA (Eq. (14)) as in LDA, this is not a problem in practice. The experimental results showed that NMF-LDA works well in an unknown environment whose acoustic characteristics are significantly different from those of an anechoic room.

VI. Conclusion

This paper presented a hierarchical Bayesian model of multichannel source separation that combines LDA and NMF con-

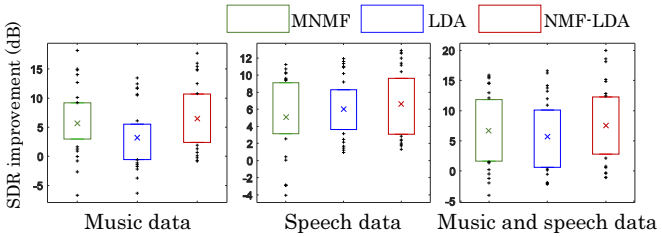

Fig. 5. SDR improvements in the case of two microphones.

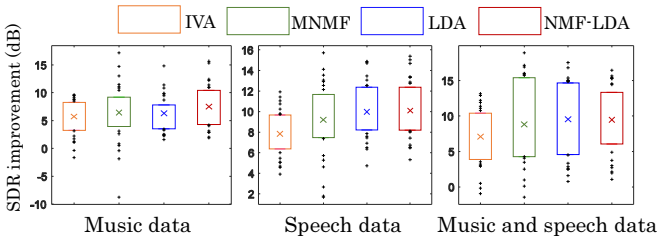

Fig. 6. SDR improvements in the case of four microphones.

TABLE I
AVERAGE SDR IMPROVEMENTS IN THE CASE OF TWO MICROPHONES.

	Music	Speech	Music and speech
MNMF	5.7 dB	5.1 dB	6.6 dB
LDA	3.2 dB	6.0 dB	5.7 dB
NMF-LDA	**6.5** dB	**6.7** dB	**7.5** dB

TABLE II
AVERAGE SDR IMPROVEMENTS IN THE CASE OF FOUR MICROPHONES.

	Music	Speech	Music and speech
IVA	5.7 dB	7.8 dB	7.1 dB
MNMF	6.4 dB	9.2 dB	8.8 dB
LDA	6.3 dB	10.0 dB	**9.5** dB
NMF-LDA	**7.5** dB	**10.1** dB	**9.5** dB

sidering both the sparseness and low-rankness of source spectrograms. A dominant source at each TF bin is identified and the spatial correlation matrix for each source is estimated in the framework of LDA. The power spectrogram of each source is decomposed into basis spectra and temporal activations, which are estimated in the framework of NMF. Experimental results showed that the proposed method achieved better source separation performance than conventional methods.

To estimate the number of sound sources as the same time as optimizing the number of basis spectra, we plan to formulate a nonparametric Bayesian extension of NMF-LDA based on the Dirichlet, gamma, and/or beta processes. Another possible extension is to allow spatial correlation matrices to smoothly vary over time for dealing with moving sound sources. In this case, the NMF part is expected to help because basis power spectra are scarcely affected by the locations of sound sources. To track moving sources in real time, we also plan to develop an efficient algorithm of online Bayesian inference.

APPENDIX

The probability density function of the complex Wishart distribution and that of the GIG distribution are given by:

$$\mathcal{W}_{\mathbb{C}}(\boldsymbol{G}|\nu,\boldsymbol{G}^0)=\frac{|\boldsymbol{G}|^{\nu-M}\exp(-tr(\boldsymbol{G}(\boldsymbol{G}^0)^{-1}))}{|\boldsymbol{G}^0|^{\nu}\pi^{M(M-1)/2}\prod_{m=0}^{M-1}\Gamma(\nu-m)}, \quad (35)$$

$$\mathrm{GIG}(y|\gamma,\rho,\tau)=\frac{\exp\{(\gamma-1)\log y-\tau y-\tau/y\}}{2\tau^{\gamma/2}\mathcal{K}_{\gamma}(2\sqrt{\rho\tau})}, \quad (36)$$

where $\mathcal{K}_{\gamma}$ is a modified Bessel function of the second kind.

ACKNOWLEDGMENT

This study was partially supported by Grant-in-Aid for Scientific Research Nos. 24220006 and 15K12063.

REFERENCES

[1] Ito, N., Araki, S. and Nakatani, T.: Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors, *IEEE ICASSP*, pp. 3238–3242 (2013).
[2] Sawada, H., Araki, S. and Makino, S.: Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment, *IEEE TASLP*, Vol. 19, No. 3, pp. 516–527 (2011).
[3] Mandel, M., Weiss, R. and Ellis, D.: Model-Based Expectation-Maximization Source Separation and Localization, *IEEE TASLP*, Vol. 18, No. 2, pp. 382–394 (2010).
[4] Otsuka, T., Ishiguro, K., Sawada, H. and Okuno, H.: Bayesian nonparametrics for microphone array processing, *IEEE TASLP*, pp. 493–504 (2014).
[5] Yilmaz, O. and Rickard, S.: Blind separation of speech mixtures via time-frequency masking, *IEEE TSP*, Vol. 52, No. 7, pp. 1830–1847 (2004).
[6] Blei, D. M., Ng, A. Y. and Jordan, M. I.: Latent Dirichlet allocation, *J. Mach. Learn. Res.*, Vol. 3, pp. 993–1022 (2003).
[7] Comon, P. and Jutten, C.: *Handbook of Blind Source Separation: Independent component analysis and applications*, Academic press (2010).
[8] Smaragdis, P. and Brown, J.: Non-negative matrix factorization for polyphonic music transcription, *IEEE WASPAA*, pp. 177–180 (2003).
[9] Sawada, H., Kameoka, H., Araki, S. and Ueda, N.: Multichannel extensions of non-negative matrix factorization with complex-valued data, *IEEE TASLP*, Vol. 21, No. 5, pp. 971–982 (2013).
[10] Ozerov, A. and Fevotte, C.: Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation, *IEEE TASLP*, Vol. 18, No. 3, pp. 550–563 (2010).
[11] Kitamura, D., Ono, N., Sawada, H., Kameoka, H. and Saruwatari, H.: Relaxation of rank-1 spatial constraint in overdetermined blind source separation, *EUSIPCO*, pp. 1271–1275 (2015).
[12] Ono, N.: Stable and fast update rules for independent vector analysis based on auxiliary function technique, *IEEE WASPAA*, pp. 189–192 (2011).
[13] Lee, I., Kim, T. and Lee, T.-W.: Fast fixed-point independent vector analysis algorithms for convolutive blind source separation, *Signal Processing*, Vol. 87, No. 8, pp. 1859–1871 (2007).
[14] Sawada, H., Mukai, R., Araki, S. and Makino, S.: A robust and precise method for solving the permutation problem of frequency-domain blind source separation, *IEEE TSAP*, Vol. 12, No. 5, pp. 530–538 (2004).
[15] Feng, F. and Kowalski, M.: An unified approach for blind source separation using sparsity and decorrelation, *EUSIPCO*, pp. 1736–1740 (2015).
[16] Comon, P.: Tensors : A brief introduction, *IEEE Signal Processing Magazine*, Vol. 31, No. 3, pp. 44–53 (2014).
[17] Chib, S. and Greenberg, E.: Understanding the Metropolis-Hastings algorithm, *The american statistician*, Vol. 49, No. 4, pp. 327–335 (1995).
[18] Jørgensen, B.: *Statistical properties of the generalized inverse Gaussian distribution*, Vol. 9, Springer Science & Business Media (2012).
[19] Blei, D. M., Cook, P. R. and Hoffman, M. D.: Bayesian nonparametric matrix factorization for recorded music, *ICML*, pp. 439–446 (2010).
[20] Araki, S., Nesta, F., Vincent, E., Koldovskỳ, Z., Nolte, G., Ziehe, A. and Benichoux, A.: The 2011 signal separation evaluation campaign (SiSEC2011):-audio source separation, *Latent Variable Analysis and Signal Separation*, Springer, pp. 414–422 (2012).
[21] Itou, K., Yamamoto, M., Takeda, K., Takezawa, T., Matsuoka, T., Kobayashi, T., Shikano, K. and Itahashi, S.: The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus, *ICSLP*, pp. 3261–3264 (1998).
[22] Vincent, E., Gribonval, R. and Fevotte, C.: Performance measurement in blind audio source separation, *IEEE TASLP*, Vol. 14, No. 4, pp. 1462–1469 (2006).