BAYESIAN MULTICHANNEL NONNEGATIVE MATRIX FACTORIZATION FOR AUDIO SOURCE SEPARATION AND LOCALIZATION

Kousuke Itakura, Yoshiaki Bando, Eita Nakamura,
Katsutoshi Itoyama, Kazuyoshi Yoshii, Tatsuya Kawahara

Graduate School of Informatics, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan

ABSTRACT

This paper presents a Bayesian extension of multichannel nonnegative matrix factorization (MNMF) that decomposes the complex spectrograms of mixture signals recorded by a microphone array into basis spectra, their temporal activations, and the spatial correlation matrices of sources (directions) in the time-frequency-channel domain. Although the original MNMF can be used in a blind setting, prior knowledge of a microphone array is useful for improving source separation. The impulse response (spatial correlation matrix) of each direction can be measured in an anechoic room, however, it differs from that in a real environment where the microphone array is used. To solve this, we propose a unified Bayesian model of source separation and localization by introducing a prior distribution determined by an anechoic spatial correlation matrix on a real spatial correlation matrix with respect to each direction. This enables us to adaptively estimate a real spatial correlation matrix and the direction of each source. Experimental results showed that our method outperformed the original MNMF and the state-of-the-art methods with prior knowledge in terms of signal-to-distortion ratio (SDR) even when the method was used in an unknown environment with acoustic characteristics different from those of the anechoic room.

Index Terms— Source separation, source localization, multichannel nonnegative matrix factorization, Bayesian modeling

1. INTRODUCTION

Microphone array processing forms the basis of computational auditory scene analysis that aims to recognize individual auditory events in a sound mixture. In multichannel source separation, phase differences between microphones play a key role. For example, frequency-domain independent component analysis (ICA) [1] and independent vector analysis (IVA) [2, 3] can separate mixture sounds in such a way that source signals are statistically independent from each other. On the other hand, separation methods based on time-frequency (TF) clustering [4, 5] assign each TF bin of mixture sounds exclusively to one of the sources by focusing on the phase information.

The power spectrograms of sources as well as phase differences between microphones have recently been modeled for multichannel source separation. Nonnegative matrix factorization (NMF) [6] is a well-known technique of single-channel source separation that approximates the power spectrogram of each source as a rank-1 matrix. Various approaches have been proposed to separate sounds by NMF [7–9]. To deal with spatial correlation matrices over microphones, Sawada *et al.* [10] proposed a multichannel extension of NMF (MNMF). Kitamura *et al.* [11] proposed a rank-1 MNMF that restricts the spatial correlation matrices to rank-1 matrices and this model was shown to be an NMF-integrated version of IVA. TF clustering-based method [5] can also be integrated with NMF [12]. Deep neural networks have been used to estimate the power densities of sources [13]. In general, most of multichannel separation methods have been designed so that they can be used in a blind setting.

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

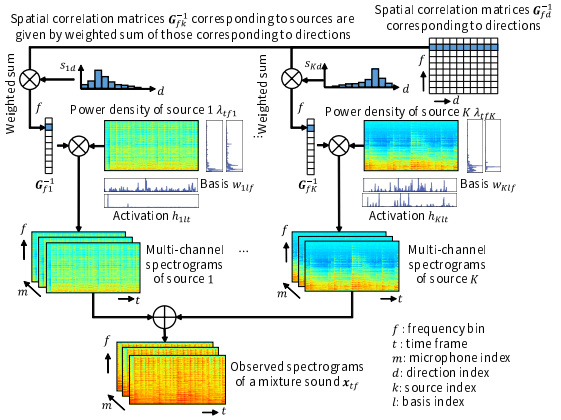

Fig. 1. The generative story of multichannel NMF.

In this paper we propose a Bayesian extension of MNMF that can incorporate various kinds of prior knowledge in a principled manner. When prior knowledge about an environment, microphones, and/or sources (*e.g.*, microphone array geometry, impulse responses measured in an anechoic room, and the template spectra of sources) are available, for example, the parameters of MNMF can be converged to reasonable values. Furthermore, a nonparametric Bayesian extension of MNMF would be feasible to automatically estimate the number of sources according to observed data in a similar way to a nonparametric Bayesian model of TF clustering [5].

As illustrated in Fig. 1, we design the generative process of the complex spectrograms of multi-channel mixture signals and then try to solve the *inverse* problem. More specifically, the power spectrogram densities of each source are determined by the product of a basis matrix and an activation matrix. The spatial correlation matrices of each source, on the other hand, are given by the weighted sum of those corresponding to different directions. Using these two types of variables, the complex spectrogram of each source is stochastically generated. The mixture spectrograms are given by the sum of the source spectrograms. The proposed method uses spatial correlation matrices (impulse responses) measured in an *anechoic* room to determine prior distributions on *real* spatial correlation matrices. Given the mixture spectrograms as observed data, we optimize all the parameters iteratively using Gibbs sampling. Finally, the source spectrograms are obtained by multichannel Wiener filtering and the source directions are determined from the direction weights.

2. RELATED WORK

A conventional approach to multichannel source separation is to estimate a linear unmixing filter that decomposes the complex spectra of mixture signals into those of source signals in the frequency domain [1–3, 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 by the impulse-response spectra. Using such linearity between mixture and source spectra, frequency-domain ICA can estimate a linear unmixing filter for each frequency bin [1]. The permutation of separated source spectra, however, is not aligned between different frequency bins. One way to resolve this permutation ambiguity is to focus on the directions and inter-frequency correlations of the sources [14]. IVA [2, 3] is an extension of ICA that can jointly deal with all frequency components in a vectorial manner.

Another popular approach to multichannel source separation is nonlinear time-frequency *hard* masking based on the sparseness (disjointness) of source spectrograms [4, 5, 12, 15–17]. If each TF bin is assigned to one of sources independently [16], the permutation ambiguity arises as in ICA. To avoid this problem, Otsuka *et al.* [5] proposed a Bayesian *mixture* method inspired by latent Dirichlet allocation (LDA) in which each TF bin is exclusively assigned to one of sources, each of which is further exclusively assigned to one of directions. The impulse responses measured in an anechoic room can be used as prior knowledge for joint source separation and localization. This method was extended to a Bayesian *factor-mixture* model called NMF-LDA [12] that approximates the power spectrogram of each source as a low-rank matrix (weighted sum of rank-1 basis matrices) for completing missing TF bins assigned to other sources. An alternative extension is a Bayesian *mixture-mixture* model called LDA-LDA [17] that exclusively assigns each TF bin of the power spectrogram of each source to one of bases.

MNMF [10,18] is nonlinear time-frequency *soft* masking method. More specifically, MNMF decomposes the complex spectrograms of mixture signals into basis spectra, temporal activations, and spatial correlation matrices. This is a *factor-factor* model because the multichannel mixture spectrum at each TF bin is modeled as a weighted sum of source spectra (*i.e.*, each TF bin is not assigned to one of sources), each of which is further modeled as a weighted sum of basis spectra. To reduce the initialization sensitivity of MNMF, various restrictions on the spatial correlation matrix have been proposed. Kitamura *et al.* [11] restricted those to rank-1 matrices. Nikunen *et al.* [19] calculated those using the geometry of the microphone array. The main contribution of this paper is to formulate a Bayesian *factor-factor* model to incorporate prior knowledge of a microphone array into the framework of MNMF as in [5]. A Bayesian model enables prior knowledge to adapt to the recorded environment.

3. BAYESIAN MNMF

This section explains a Bayesian model of MNMF that decomposes the complex spectrograms of mixture signals into basis spectra, temporal activations, and spatial correlation matrices in a statistical manner. Bayesian estimation of real spatial correlation matrices based on those measured in an anechoic room leads to accurate joint source separation and localization in an arbitrary environment.

3.1. Model formulation

When K sources are observed with M microphones, each TF bin of the complex spectrograms of observed mixture signals and that of the complex spectrograms of source signals are defined as follows:

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

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

where x_{tfm} and y_{tfk} respectively denote the complex spectrum of microphone m and that of source k at time frame t and frequency f. As is often the case with related work, we assume that the source spectrum y_{tfk} is complex Gaussian distributed as follows:

$$y_{tfk} \mid \lambda_{tfk} \sim \mathcal{N}_{\mathbb{C}}(0, \lambda_{tfk}), \tag{3}$$

where λ_{tfk} is the power spectrum density of source k at time frame t and frequency f.

Assuming an instantaneous mixing process in the frequency domain, the observation $\boldsymbol{x}_{tf}$ is represented using source spectra and steering vectors as follows:

$$\boldsymbol{x}_{tf} = \sum_{k=1}^{K} \boldsymbol{a}_{fk} y_{tfk}, \tag{4}$$

where $\boldsymbol{a}_{fk} \in \mathbb{C}^M$ is a steering vector of source k at frequency f. We represent spatial directions with D discretized values indexed by $d = 1, \ldots, D$. Note that a steering vector $\boldsymbol{a}_{fd}$ depends on both direction d and frequency f and it should theoretically be equivalent to a steering vector $\boldsymbol{a}_{fk}$ of a particular source k because the source exists at one of the directions. Relaxing this constraint, the steering vector of the source is represented as the weighted sum of steering vectors of all directions as follows:

$$\boldsymbol{a}_{fk} = \sum_{d=1}^{D} u_{kd} \boldsymbol{a}_{fd}, \tag{5}$$

where u_{kd} is the weight of direction d for source k and $\{u_{kd}\}_{d=1}^{D}$ is expected to be sparse to associate source k with one direction.

Using Eqs. (3), (4), and (5), the observation spectrum $\boldsymbol{x}_{tf}$ is found to be multivariate complex Gaussian distributed as follows:

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

where $s_{kd} = u_{kd}^2$ and $\boldsymbol{G}_{fd}^{-1} = \boldsymbol{a}_{fd} \boldsymbol{a}_{fd}^H$ is a spatial correlation matrix for direction d at frequency f.

If the power spectrogram densities $\{\lambda_{tfk}\}_{t=1,f=1}^{T,F}$ of each source k are assumed to have a low-rank structure, they are decomposed as

$$\lambda_{tfk} = \sum_{l=1}^{L} w_{klf} h_{klt}, \tag{7}$$

where w_{klt} is the power spectrum density of basis l at frequency f and h_{klt} is the volume of basis l at time frame t.

Plugging Eq. (7) into Eq. (6), the likelihood of the unknown parameters $\boldsymbol{W}$, $\boldsymbol{H}$, $\boldsymbol{S}$, and $\boldsymbol{G}$ for the observed data $\boldsymbol{X}$ is given by

$$\boldsymbol{x}_{tf} \mid \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{S}, \boldsymbol{G} \sim \mathcal{N}_{\mathbb{C}} \left(\boldsymbol{0}, \sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{d=1}^{D} w_{klf} h_{klt} s_{kd} \boldsymbol{G}_{fd}^{-1} \right), \tag{8}$$

For mathematical convenience, the conjugate prior distributions are put on those model parameters as follows:

$$w_{klf} \sim \text{Gamma}(a_0^w, b_0^w), \tag{9}$$

$$h_{klt} \sim \text{Gamma}(a_0^h, b_0^h), \tag{10}$$

$$s_{kd} \sim \text{Gamma}(a_0^s, b_0^s), \tag{11}$$

$$\boldsymbol{G}_{fd} \sim \mathcal{W}_{\mathbb{C}}(\nu_0, \boldsymbol{G}_{fd}^0), \tag{12}$$

where $\mathcal{W}_{\mathbb{C}}$ is the complex Wishart distribution defined by

$$\mathcal{W}_{\mathbb{C}}(\boldsymbol{X}|\nu,\boldsymbol{\Lambda}) = \frac{|\boldsymbol{X}|^{\nu-M}\exp\{-\mathrm{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{X})\}}{|\boldsymbol{\Lambda}|^{\nu}\pi^{M(M-1)/2}\prod_{m=1}^{M-1}\Gamma(\nu-m)}, \quad (13)$$

where $\nu \geq M$ is a degree of freedom and $\boldsymbol{\Lambda} \succ \boldsymbol{0} \in \mathbb{C}^{M\times M}$ is a scale matrix. To use prior knowledge about a microphone array, the steering vectors $\{\boldsymbol{a}_{fd}^0\}_{f=1}^F$ are measured for each direction d in an anechoic room and $\boldsymbol{G}_{fd}^0$ is set as $\boldsymbol{G}_{fd}^0 = (\boldsymbol{a}_{fd}^0(\boldsymbol{a}_{fd}^0)^H + \epsilon\boldsymbol{I})$, where $\epsilon > 0$ is a small number to make $\boldsymbol{G}_{fd}^0$ positive definite.

3.2. Bayesian inference

Our goal is to calculate the posterior distribution $p(\boldsymbol{W},\boldsymbol{H},\boldsymbol{S},\boldsymbol{G}|\boldsymbol{X})$ using the Bayes' theorem $p(\boldsymbol{W},\boldsymbol{H},\boldsymbol{S},\boldsymbol{G}|\boldsymbol{X}) = \frac{p(\boldsymbol{X},\boldsymbol{W},\boldsymbol{H},\boldsymbol{S},\boldsymbol{G})}{p(\boldsymbol{X})}$ and find optimal parameters that maximize the posterior in practice. Since $p(\boldsymbol{W},\boldsymbol{H},\boldsymbol{S},\boldsymbol{G}|\boldsymbol{X})$ is analytically intractable, but a posterior distribution of each parameter conditioned on the remaining parameters (*e.g.*, $p(\boldsymbol{W}|\boldsymbol{H},\boldsymbol{G},\boldsymbol{S},\boldsymbol{X})$) is tractable, we can use Gibbs sampling [20] that alternately and iteratively updates one of the parameters $\boldsymbol{W}$, $\boldsymbol{H}$, $\boldsymbol{S}$, and $\boldsymbol{G}$ according to the conditional posterior distribution by fixing the other parameters.

To derive a tractable conditional posterior of each parameter, we use a variational approach [21]. Although the conditional posterior is proportional to the complete joint likelihood given by the product of Eqs. (8)–(12), it is difficult to directly get samples from the conditional posterior because of the complicated form of Eq. (8). Therefore, the log-likelihood function defined by Eq. (8) is lower bounded by a tractable auxiliary function having auxiliary variables. The auxiliary function should become equal to the log-likelihood function when it is maximized with respect to the auxiliary variables. Such an auxiliary function can be used as a proxy for the log-likelihood function. More specifically, letting $\boldsymbol{Y}_{tfkld} = w_{klf}h_{klt}s_{kd}\boldsymbol{G}_{fd}^{-1}$, the log-likelihood is given by

$$\log p(\boldsymbol{X}|\boldsymbol{W},\boldsymbol{H},\boldsymbol{S},\boldsymbol{G}) \overset{\mathrm{c}}{=} \sum_{tf}\left(-\log|\boldsymbol{Y}_{tf}| - \mathrm{tr}(\boldsymbol{X}_{tf}\boldsymbol{Y}_{tf}^{-1})\right), \quad (14)$$

where $\boldsymbol{X}_{tf} = \boldsymbol{x}_{tf}^H\boldsymbol{x}_{tf}$ and $\boldsymbol{Y}_{tf} = \sum_{kld}\boldsymbol{Y}_{tfkld}$. To derive a lower bound $\mathcal{L}$ from Eq. (14), we use two inequalities used in [21]. First, for a convex function $f(\boldsymbol{Z}) = -\log|\boldsymbol{Z}|$ ($\boldsymbol{Z} \succeq \boldsymbol{0} \in \mathbb{C}^{M\times M}$), we calculate a tangent plane at arbitrary $\boldsymbol{\Omega} \succeq \boldsymbol{0}$ by using a first-order Taylor expansion as follows:

$$-\log|\boldsymbol{Z}| \geq -\log|\boldsymbol{\Omega}| - \mathrm{tr}(\boldsymbol{\Omega}^{-1}\boldsymbol{Z}) + M, \quad (15)$$

where the equality holds when $\boldsymbol{\Omega} = \boldsymbol{Z}$. Second, for a concave function $g(\boldsymbol{Z}) = -\mathrm{tr}(\boldsymbol{Z}^{-1}\boldsymbol{A})$ with any matrix $\boldsymbol{A} \succeq \boldsymbol{0}$, we use the following inequality:

$$-\mathrm{tr}\left(\left(\textstyle\sum_{k=1}^K \boldsymbol{Z}_k\right)^{-1}\boldsymbol{A}\right) \geq -\sum_{k=1}^K \mathrm{tr}\left(\boldsymbol{Z}_k^{-1}\boldsymbol{\Phi}_k\boldsymbol{A}\boldsymbol{\Phi}_k^H\right), \quad (16)$$

where $\{\boldsymbol{Z}_k \succeq \boldsymbol{0}\}_{k=1}^K$ is a set of arbitrary matrices, $\{\boldsymbol{\Phi}_k\}_{k=1}^K$ is a set of auxiliary matrices that sum to the identity matrix ($\sum_k \boldsymbol{\Phi}_k = \boldsymbol{I}$), and the equality holds when $\boldsymbol{\Phi}_k = \boldsymbol{Z}_k(\sum_{k'}\boldsymbol{Z}_{k'})^{-1}$.

Using Inequalities (15) and (16), the log-likelihood function given by Eq. (14) is lower bounded by $\mathcal{L}$ as follows:

$$\begin{aligned}
&\log p(\boldsymbol{X}|\boldsymbol{W},\boldsymbol{H},\boldsymbol{S},\boldsymbol{G}) \\
&\overset{\mathrm{c}}{\geq} \sum_{tf}\left(-\mathrm{tr}\left(\boldsymbol{Y}_{tf}\boldsymbol{\Omega}_{tf}^{-1}\right) - \log|\boldsymbol{\Omega}_{tf}| + M\right) \\
&\quad - \sum_{tfkld}\mathrm{tr}\left(\boldsymbol{Y}_{tfkld}^{-1}\boldsymbol{\Phi}_{tfkld}\boldsymbol{X}_{tf}\boldsymbol{\Phi}_{tfkld}\right) \overset{\mathrm{def}}{=} \mathcal{L} \quad (17)
\end{aligned}$$

where $\boldsymbol{\Omega}_{tf}$ and $\boldsymbol{\Phi}_{tfkld}$ are newly-introduced auxiliary variables. The auxiliary function $\mathcal{L}$ is maximized, *i.e.*, the equality holds, when $\boldsymbol{\Omega}_{tf}$ and $\boldsymbol{\Phi}_{tfkld}$ are given by

$$\boldsymbol{\Omega}_{tf} = \boldsymbol{Y}_{tf}, \quad (18)$$

$$\boldsymbol{\Phi}_{tfkld} = \boldsymbol{Y}_{tfkld}\boldsymbol{Y}_{tf}^{-1}. \quad (19)$$

The parameters w_{klf}, h_{klt}, s_{kd}, and $\boldsymbol{G}_{fd}$ can be sampled from the following conditional distributions proportional to the product of Eqs. (9)–(12) and (17):

$$w_{klf} \mid \boldsymbol{X},\boldsymbol{\Theta}_{\neg w_{klf}} \sim \mathrm{GIG}(a_0^w,\rho_{klf}^w,\tau_{klf}^w), \quad (20)$$

$$h_{klt} \mid \boldsymbol{X},\boldsymbol{\Theta}_{\neg h_{klt}} \sim \mathrm{GIG}(a_0^h,\rho_{klt}^h,\tau_{klt}^h), \quad (21)$$

$$s_{kd} \mid \boldsymbol{X},\boldsymbol{\Theta}_{\neg s_{kd}} \sim \mathrm{GIG}(a_0^s,\rho_{kd}^s,\tau_{kd}^s), \quad (22)$$

$$\boldsymbol{G}_{fd} \mid \boldsymbol{X},\boldsymbol{\Theta}_{\neg \boldsymbol{G}_{fd}} \sim \mathrm{MGIG}_{\mathbb{C}}(\nu_0,\boldsymbol{R}_{fd},\boldsymbol{U}_{fd}), \quad (23)$$

where $\boldsymbol{\Theta}_{\neg *}$ is a set of all parameters excluding $*$. GIG indicates the generalized inverse Gaussian distribution [22] and $\mathrm{MGIG}_{\mathbb{C}}$ indicates the complex matrix GIG distribution [23], defined by

$$\mathrm{GIG}(x|\gamma,\rho,\tau) = \frac{\exp\{(\gamma-1)\log x - \rho x - \tau/x\}\rho^{\gamma/2}}{2\tau^{\gamma/2}\mathcal{K}_{\gamma}(2\sqrt{\rho\tau})}, \quad (24)$$

$$\mathrm{MGIG}_{\mathbb{C}}(\boldsymbol{X}|\gamma,\boldsymbol{R},\boldsymbol{U}) \propto |\boldsymbol{X}|^{\gamma-M}\exp\{-\mathrm{tr}(\boldsymbol{R}\boldsymbol{X}+\boldsymbol{U}\boldsymbol{X}^{-1})\}, \quad (25)$$

where $\mathcal{K}_{\gamma}$ is the modified Bessel function of the second kind, γ is a real number, $\rho > 0$, $\tau > 0$, $\boldsymbol{R} \succ \boldsymbol{0}$, and $\boldsymbol{U} \succ \boldsymbol{0}$. To draw samples from the GIG and complex MGIG distributions, we use a rejection sampling method [24] and a Metropolis-Hastings (MH) method [25], respectively. In the MH method, we use as a proposal distribution a complex Wishart distribution whose mode equals to that of a target complex MGIG distribution (the mode of an MGIG distribution can be calculated by using an algebraic Riccati equation [23]). The conditional posterior parameters ρ_*^*, τ_*^*, $\boldsymbol{R}_{fd}$, and $\boldsymbol{U}_{fd}$ are given by

$$\rho_{klf}^w = b_0^w + \sum_{td} h_{klt}s_{kd}\mathrm{tr}(\boldsymbol{G}_{fd}^{-1}\boldsymbol{\Omega}_{tf}^{-1}), \quad (26)$$

$$\tau_{klf}^w = \sum_{td} h_{klt}^{-1}s_{kd}^{-1}\mathrm{tr}\left(\boldsymbol{G}_{fd}\boldsymbol{\Phi}_{tfkld}\boldsymbol{X}_{tf}\boldsymbol{\Phi}_{tfkld}\right), \quad (27)$$

$$\rho_{klt}^h = b_0^h + \sum_{fd} w_{klf}s_{kd}\mathrm{tr}(\boldsymbol{G}_{fd}^{-1}\boldsymbol{\Omega}_{tf}^{-1}), \quad (28)$$

$$\tau_{klt}^h = \sum_{fd} w_{klf}^{-1}s_{kd}^{-1}\mathrm{tr}\left(\boldsymbol{G}_{fd}\boldsymbol{\Phi}_{tfkld}\boldsymbol{X}_{tf}\boldsymbol{\Phi}_{tfkld}\right), \quad (29)$$

$$\rho_{kd}^s = b_0^s + \sum_{tfl} w_{klf}h_{klt}\mathrm{tr}(\boldsymbol{G}_{fd}^{-1}\boldsymbol{\Omega}_{tf}^{-1}), \quad (30)$$

$$\tau_{kd}^s = \sum_{tfl} w_{klf}^{-1}h_{klt}^{-1}\mathrm{tr}\left(\boldsymbol{G}_{fd}\boldsymbol{\Phi}_{tfkld}\boldsymbol{X}_{tf}\boldsymbol{\Phi}_{tfkld}\right), \quad (31)$$

$$\boldsymbol{R}_{fd} = (\boldsymbol{G}_{fd}^0)^{-1} + \sum_{tkl} w_{klf}^{-1}h_{klt}^{-1}s_{kd}^{-1}\boldsymbol{\Phi}_{tfkld}\boldsymbol{X}_{tf}\boldsymbol{\Phi}_{tfkld}, \quad (32)$$

$$\boldsymbol{U}_{fd} = \sum_{tkl} w_{klf}h_{klt}s_{kd}\boldsymbol{\Omega}_{tf}^{-1}. \quad (33)$$

3.3. Source separation and localization

The multichannel mixture spectrum $\boldsymbol{x}_{tf}$ over microphones at time frame t and frequency bin f is decomposed into the sum of multichannel source spectra $\{\tilde{\boldsymbol{x}}_{tfk}\}_{k=1}^K$ using multichannel Wiener filtering [10] as follows:

$$\tilde{\boldsymbol{x}}_{tfk} = \boldsymbol{Y}_{tfk}\boldsymbol{Y}_{tf}^{-1}\boldsymbol{x}_{tf}, \quad (34)$$

where $\boldsymbol{Y}_{tfk} = \sum_{ld}\boldsymbol{Y}_{tfkld}$.

The source spectrograms of the first channel are transformed into time-domain source signals using the inverse short-time Fourier transform. The direction $d(k)$ of each source k can be estimated by finding the direction such that the weight s_{kd} of direction d for the source k is maximized as follows:

$$d(k) = \mathrm{argmax}_d s_{kd}. \quad (35)$$

4. EVALUATION

This section reports comparative quantitative experiments conducted for evaluating the proposed method.

4.1. Experimental conditions

We synthesized convolutive mixture sounds as test data. Fig. 2 shows the locations of microphones and sources. Three sources were convoluted using impulse responses measured with 4 microphones in a room where the reverberation time RT_{60} was 400 ms. We used music signals (including guitar, bass, vocal, hi-hat, and piano sounds) and speech signals selected from the SiSEC data set [26] and the JNAS phonetically balanced Japanese utterances [27]. 30 mixture signals were used for evaluation: 10 mixtures of music signals, 10 mixtures of speech signals, and 10 mixtures of music and speech signals. The audio signals were sampled at 16 kHz and a short-time Fourier transformation was carried out with a 512-pt Hanning window and a 256-pt shift size. Hyperparameters were determined experimentally as follows: $L = 20$, $a_0^w = a_0^h = a_0^s = b_0^w = b_0^h = b_0^s = 1$, $\nu_0 = M + 1$, and $\epsilon = 0.01$. The steering vectors $\boldsymbol{a}_{fd}^0$ were measured for all directions with an angular interval of 5° ($D = 72$) in an anechoic room and they were different from those used to generate the test data.

We compared our method with the standard and state-of-the-art methods such as IVA [2], MNMF [10], NMF-LDA [12], and LDA-LDA [17]. The parameters of each method were updated or sampled 200 times. The signal-to-distortion ratio (SDR), signal-to-inferences ratio (SIR) and signal-to-artifacts ratio (SAR) [28] were used to evaluate the separation performance. We compared the localization performance of the proposed method with those of NMF-LDA [12] and LDA-LDA [17]. We evaluated the localization performance in terms of the average of the absolute localization errors.

4.2. Experimental results

The experimental results are listed in Tables 1, 2, and 3. With all signal mixtures the SDR was highest for the proposed method, but in terms of SIR the proposed method was inferior to NMF-LDA and LDA-LDA, which are methods based on TF clustering (hard masking of TF bins). In terms of SAR the proposed method was almost the same as the conventional MNMF and was better than any of the other conventional methods. The proposed Bayesian extension was found to work well because the SDR and SIR for it were higher than those of MNMF and the SAR for it was almost the same as that for MNMF.

The average absolute localization errors are listed in Table 4. Although the localization errors for the proposed method were larger than those for the conventional methods, the differences were less than 1°. We can thus say that the localization performance of the proposed method is comparable to that of the conventional methods.

The results show that the proposed model could separate and localize mixture sounds in an environment where the effective steering vectors are different from those measured in an anechoic room. This suggests that the proposed model can not only utilize the prior knowledge on the steering vectors but also adapt them according to the environment where mixture signals are observed.

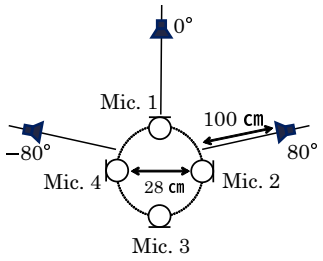

Fig. 2. Positions of microphones and sources.

Table 1. Evaluation on music signal mixtures.

	SDR	SIR	SAR
Bayesian MNMF	**3.2** dB	8.1 dB	**7.5** dB
MNMF [10]	1.0 dB	6.2 dB	6.7 dB
NMF-LDA [12]	0.5 dB	**8.7** dB	3.2 dB
LDA-LDA [17]	0.7 dB	**8.7** dB	3.3 dB
IVA [2]	0.3 dB	4.9 dB	5.7 dB

Table 2. Evaluation on speech signal mixtures.

	SDR	SIR	SAR
Bayesian MNMF	**6.0** dB	12.6 dB	7.5 dB
MNMF [10]	4.8 dB	10.0 dB	**7.7** dB
NMF-LDA [12]	4.2 dB	14.0 dB	5.2 dB
LDA-LDA [17]	5.8 dB	**17.0** dB	6.3 dB
IVA [2]	3.4 dB	7.5 dB	7.1 dB

Table 3. Evaluation on music and speech signal mixtures.

	SDR	SIR	SAR
Bayesian MNMF	**4.9** dB	13.0 dB	**6.6** dB
MNMF [10]	1.8 dB	8.6 dB	6.1 dB
NMF-LDA [12]	1.1 dB	9.8 dB	4.1 dB
LDA-LDA [17]	2.8 dB	**14.2** dB	3.9 dB
IVA [2]	0.1 dB	5.3 dB	5.3 dB

Table 4. Average absolute localization error (degree).

	speech	music	music+speech
Bayesian MNMF	2.00	2.50	2.50
NMF-LDA [12]	1.67	1.67	1.83
LDA-LDA [17]	1.67	1.83	2.00

5. CONCLUSION

This paper presented a Bayesian extension of MNMF for audio source separation and localization that can incorporate prior knowledge about an environment, microphones, and/or sources. Experimental results showed that (1) the proposed method achieved better separation performance than the conventional MNMF, (2) its localization performance was comparable to that of conventional methods, and (3) it worked well in an unknown environment whose acoustic characteristics were significantly different from those of an anechoic room.

With further extensions, it would be possible to estimate the number of the sources in a nonparametric Bayesian manner and to develop an online source separation algorithm. Other future works include examining the effect of prior learning of basis matrices and comparing the proposed method with various methods such as other extensions of MNMF [11, 19].

6. REFERENCES

[1] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent component analysis and applications*, Academic press, 2010.

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

[3] I. Lee, T. Kim, and T. Lee, "Fast fixed-point independent vector analysis algorithms for convolutive blind source separation," *Signal Processing*, vol. 87, no. 8, pp. 1859–1871, 2007.

[4] N. Ito, S. Araki, and T. Nakatani, "Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors," in *IEEE ICASSP*, 2013, pp. 3238–3242.

[5] T. Otsuka, K. Ishiguro, H. Sawada, and H. G. Okuno, "Bayesian nonparametrics for microphone array processing," *IEEE TASLP*, pp. 493–504, 2014.

[6] P. Smaragdis and J.C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *IEEE WASPAA*, 2003, pp. 177–180.

[7] David M Blei, Perry R. Cook, and Matthew D. Hoffman, "Bayesian nonparametric matrix factorization for recorded music," in *ICML*, 2010, pp. 439–446.

[8] K. Adilolu and E. Vincent, "Variational bayesian inference for source separation and robust feature extraction," *IEEE TASLP*, vol. 24, no. 10, pp. 1746–1758, 2016.

[9] J. T. Chien and P. K. Yang, "Bayesian factorization and learning for monaural source separation," *IEEE TASLP*, vol. 24, no. 1, pp. 185–195, 2016.

[10] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE TASLP*, vol. 21, no. 5, pp. 971–982, 2013.

[11] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Relaxation of rank-1 spatial constraint in overdetermined blind source separation," in *EUSIPCO*, 2015, pp. 1271–1275.

[12] K. Itakura, Y. Bando, E. Nakamura, K. Itoyama, and K. Yoshii, "A unified Bayesian model of time-frequency clustering and low-rank approximation for multi-channel source separation," in *EUSIPCO*, 2016, pp. 2280–2284.

[13] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE TASLP*, vol. 24, no. 10, pp. 1652–1664, 2016.

[14] H. Sawada, R. Mukai, S. Araki, and S. Makino, "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] H. Sawada, S Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE TASLP*, vol. 19, no. 3, pp. 516–527, 2011.

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

[17] K. Itakura, Y. Bando, E. Nakamura, K. Itoyama, K. Yoshii, and T. Kawahara, "Time-frequency clustering based on a nested mixture model for multichannel source separation (in Japanese)," in *speech processing society*, 2016, pp. 25–28.

[18] 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, 2010.

[19] J. Nikunen and T. Virtanen, "Multichannel audio separation by direction of arrival based spatial covariance model and nonnegative matrix factorization," in *IEEE ICASSP*, 2014, pp. 6677–6681.

[20] G. Casella and E. I. George, "Explaining the Gibbs sampler," *The American Statistician*, vol. 46, no. 3, pp. 167–174, 1992.

[21] K. Yoshii, K. Itoyama, and M. Goto, "Student's t nonnegative matrix factorization and positive semidefinite tensor factorization for single-channel audio source separation," in *IEEE ICASSP*, 2016, pp. 51–55.

[22] B. Jørgensen, *Statistical properties of the generalized inverse Gaussian distribution*, vol. 9, Springer Science & Business Media, 2012.

[23] F. Fazayeli and A. Banerjee, *The Matrix Generalized Inverse Gaussian Distribution: Properties and Applications*, pp. 648–664, Springer International Publishing, 2016.

[24] J. S. Dagpunar, *Simulation and Monte Carlo: With applications in finance and MCMC*, John Wiley & Sons, 2007.

[25] S. Chib and E. Greenberg, "Understanding the Metropolis-Hastings algorithm," *The american statistician*, vol. 49, no. 4, pp. 327–335, 1995.

[26] S. Araki, F. Nesta, E. Vincent, Z. Koldovskỳ, G. Nolte, A. Ziehe, and A. Benichoux, "The 2011 signal separation evaluation campaign (SiSEC2011):-audio source separation," in *Latent Variable Analysis and Signal Separation*, pp. 414–422. Springer, 2012.

[27] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus," *ICSLP*, pp. 3261–3264, 1998.

[28] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE TASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.