Bayesian Multichannel Audio Source Separation Based on Integrated Source and Spatial Models

Kousuke Itakura, Yoshiaki Bando, *Member, IEEE*, Eita Nakamura, *Member, IEEE*, Katsutoshi Itoyama, *Member, IEEE*, Kazuyoshi Yoshii, *Member, IEEE*, and Tatsuya Kawahara, *Fellow, IEEE*

***Abstract*—This paper presents new statistical methods of multichannel audio source separation based on unified source and spatial models that, respectively, represent the generative process of latent source spectrograms and that of observed mixture spectrograms. One possibility of the source model is a factor model based on nonnegative matrix factorization that represents each time-frequency (TF) bin as the weighted sum of basis spectra. Another possibility is a mixture model inspired by latent Dirichlet allocation that exclusively classifies each TF bin into one of basis spectra. Similarly, the spatial model can either be a factor model that represents each TF bin as the weighted sum of source spectra or a mixture model that classifies each bin into one of those spectra. To unify these models in a principled manner and incorporate prior knowledge of a microphone array, we propose hierarchical Bayesian models of all the source–spatial combinations (factor–factor, mixture–factor, factor–mixture, and mixture–mixture models) and derive efficient Gibbs sampling algorithms for posterior inference. Experimental results showed that the proposed unified models outperformed the state-of-the-art method using only the spatial mixture model. Among the four unified models, the spatial factor model tended to work better than the spatial mixture model in exchange for larger computational cost, and the choice of source models had a little impact on the performance and computational cost.**

***Index Terms*—Multichannel source separation, latent Dirichlet allocation, nonnegative matrix factorization, Bayesian models.**

I. Introduction

MICROPHONE array processing forms the basis of computational auditory scene analysis that aims to understand individual auditory events in a sound mixture. A promising approach to multichannel source separation is to represent a hierarchical generative process of the time-frequency spectrogram of an observed mixture signal by considering both a source model representing a generative process of source spectrograms and a spatial model representing a mixing process of those sources. There are two major kinds of probabilistic models that can be used for representing those generative processes. One is a *mixture model* based on the *sum of probability distributions* used for describing random variables. The other is a *factor model* based on the *sum of random variables* described by probability distributions.

A typical approach to multichannel source separation is to formulate a mixture model as a spatial model for time-frequency (TF) clustering [1]–[8]. If the frequency components of each source are sparsely distributed, as is often the case with pitched 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 [1], is reasonable because the additivity of source spectrograms does not hold exactly and at each TF bin a loud sound masks smaller sounds. Under this assumption, Otsuka *et al.* [8] proposed a Bayesian mixture model inspired by latent Dirichlet allocation (LDA) [9] for classifying each TF bin into one of sources at the same time as classifying each source into one of directions. Such unified source separation and localization can circumvent the permutation problem of conventional frequency-domain separation methods such as independent component analysis (ICA) [10].

In single-channel source separation, a factor model called nonnegative matrix factorization (NMF) has gained popularity [11]. It approximates the power spectrogram of an observed mixture signal as a low-rank matrix given by the product of a basis matrix (a set of basis spectra) and an activation matrix (a set of temporal activations). Multichannel extensions of NMF using factor models as source and spatial models have recently been proposed using both the low-rankness and spatial characteristics of sources [12]–[17]. The complex spectrograms of observed multichannel signals are modeled by basis spectra, temporal activations, and full-rank or rank-1 spatial covariance matrices. A Bayesian model was proposed to use prior knowledge on a microphone array (e.g., impulse responses recorded in an anechoic room) [17]. A further extension based on nonparametric Bayesian modeling would be feasible to estimate the number of sources as in [18].

In this paper we propose and evaluate unified Bayesian models corresponding to all the four combinations of source and

Manuscript received January 9, 2017; revised June 19, 2017, September 30, 2017, and December 14, 2017; accepted December 16, 2017. Date of publication January 31, 2018; date of current version February 8, 2018. This work was supported in part by Grants-in-Aid for Scientific Research (KAKENHI) (No. 24220006, No. 26700020, and No. 15K12063). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Simon Doclo. *(Corresponding author: Kazuyoshi Yoshii.)*

K. Itakura, Y. Bando, E. Nakamura, K. Itoyama, and T. Kawahara are with the Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan (e-mail: itakura@sap.ist.i.kyoto-u.ac.jp; yoshiaki@sap.ist.i.kyoto-u.ac.jp; enakamura@sap.ist.i.kyoto-u.ac.jp; itoyama@i.kyoto-u.ac.jp; kawahara@i.kyoto-u.ac.jp).

K. Yoshii is with the Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan, and also with the RIKEN Center for Advanced Intelligence Project, Tokyo 103-0027, Japan (e-mail: yoshii@i.kyoto-u.ac.jp).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2017.2789320

TABLE I
COMBINATIONS OF SPATIAL MODELS AND SOURCE MODELS

Source \ Spatial	Factor model	Mixture model
Factor model	Factor-Factor [18]	Factor-Mixture [20]
Mixture model	Mixture-Factor	Mixture-Mixture
Not applicable (NA)	NA-Factor	NA-Mixture [8]

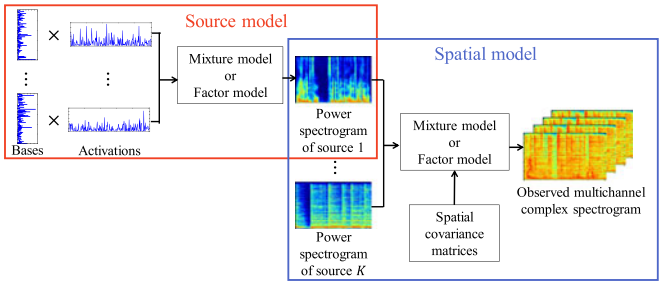


Fig. 1. The hierarchical generative process underlying the proposed models. The power spectrogram of each source signal is generated by a source model represented as a mixture or factor model, and the complex spectrograms of observed multichannel signals are generated by a spatial model represented as a mixture or factor model.

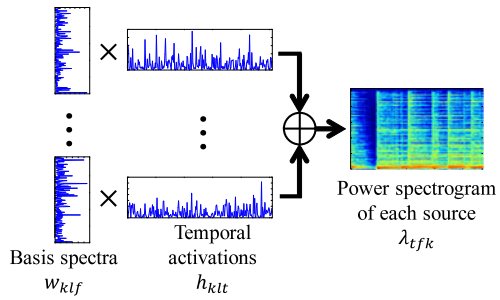


Fig. 2. The generative process of the power spectrogram of each source k in the source factor model. Each TF bin of the source spectrogram is given by the *sum* of the L products of basis spectra and their activations.

spatial models (factor-factor [17], mixture-factor, factor-mixture [19], and mixture-mixture models). Since multichannel source separation is feasible by using only a spatial model (mixture [8] or factor model), there are six variants shown in Table I. As illustrated in Fig. 1, source power spectrograms are stochastically generated based on a source model and observed mixture spectrograms are stochastically generated based on a spatial model. Either a mixture or factor model can be used for formulating a source or spatial model. Given the complex spectrograms of observed multichannel signals, we try to solve the *inverse* problem to jointly estimate all the parameters of those models using Gibbs sampling.

Figs. 2–5 illustrate the source and spatial models. In a source factor model, the power spectrogram of each source is represented as the sum of rank-1 basis spectrograms (Fig. 2). In a source mixture model, it is represented as a patchwork of rank-1 basis spectrograms (Fig. 3). In a spatial factor model, the spatial covariance matrix corresponding to each source is given by the weighted sum of those corresponding to directions, and the observed multichannel complex spectrogram is the sum

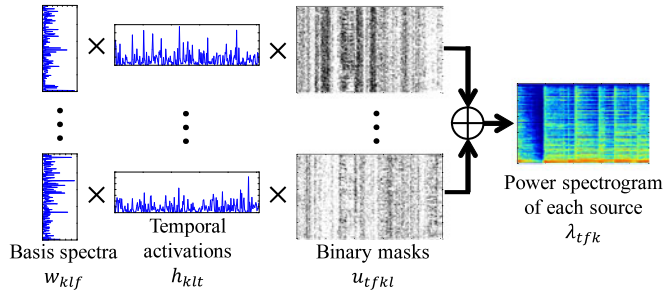


Fig. 3. The generative process of the power spectrogram of each source k in the source mixture model. Each TF bin of the source spectrogram is given by *one* of the L products of basis spectra and their activations.

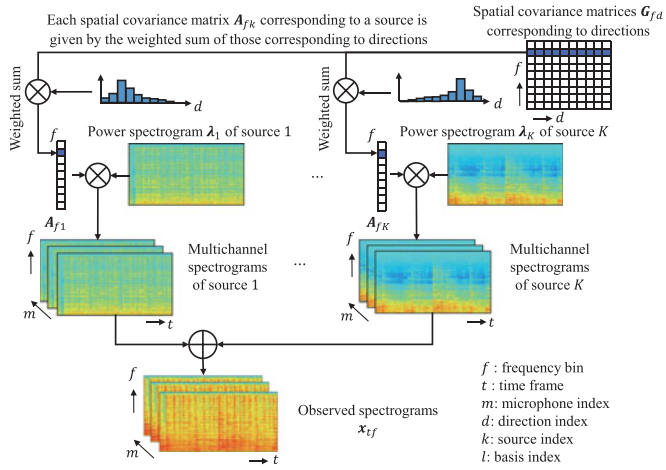


Fig. 4. The generative process of the complex spectrograms of multichannel signals in the spatial factor model. The spatial covariance matrix of each source k is given by the weighted *sum* of D spatial covariance matrices corresponding to different directions and the *sum* of the complex spectra of K sources is observed at each TF bin.

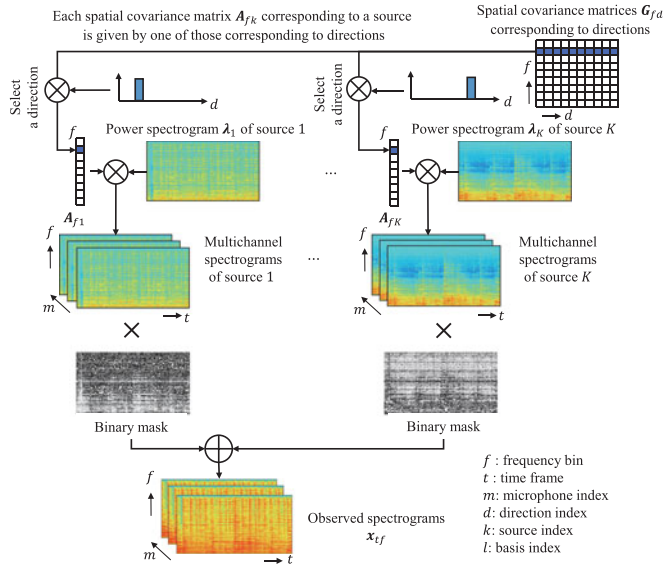


Fig. 5. The generative process of the complex spectrograms of multichannel signals in the spatial mixture model. The spatial covariance matrix of each source k is given by *one* of D spatial covariance matrices corresponding to different directions and *one* of the complex spectra of K sources is observed at each TF bin.

of source images[1] (Fig. 4). In a spatial mixture model, the spatial covariance matrix corresponding to each source is one of those corresponding to directions, and the observed multichannel complex spectrogram is a patchwork of source images (Fig. 5). We experimentally found that the spatial factor model tended to work better than the spatial mixture model in exchange for larger computational cost, and the choice of source models had little impact on the performance and computational cost.

The rest of the paper is organized as follows: Section II reviews related work on multichannel source separation in terms of source and spatial modeling. Section III explains the four proposed models based on source-spatial combinations and Section IV derives Bayesian inference algorithms. Section V reports the results of comparative experiments and Section VI summarizes the key findings.

II. Related Work

This section introduces conventional methods of audio source separation using source models and/or spatial models.

A. Source Models

In single channel source separation without spatial information, it is necessary to focus on some structural characteristics of sources. One of the most popular methods is non-negative matrix factorization (NMF) and various extensions of NMF have recently been proposed [18], [20]. Assuming the low-rankness of source power spectrograms, NMF approximates an observed power spectrogram as the sum of products of basis spectra (spectral templates of individual sources) and activations (temporal volumes of those templates).

In the context of speech enhancement, robust principal component analysis (RPCA) [21] and robust NMF (RNMF) [22] have often been used. While NMF assumes the low-rankness of all sources, RPCA and RNMF assume the sparseness of speech sounds and the low-rankness of noise sounds. The observed noisy spectrogram is decomposed into the sum of a sparse matrix and a low-rank matrix corresponding to speech sounds and noise sounds.

B. Spatial Models

In multichannel source separation, sounds are commonly separated by using spatial information such as the phase differences between microphones. 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 [10], [23]–[27]. Mixture signals are usually represented 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 [10] estimates a linear unmixing filter for each frequency bin. The permutation of separated source spectra, however, is not aligned between different frequency bins. One way to resolve this permutation ambiguity is to use the directions and inter-frequency correlations of the sources [23]. IVA [24], [25] is an extension of ICA that can jointly deal with all frequency components in a vectorial manner. Several variants of beamformers based on a minimum variance distortionless response (MVDR) [26] and generalized eigenvalue decomposition (GEV) [27] have widely been used for multichannel speech enhancement. MVDR puts a constraint that the source signals are not distorted by linear filtering and GEV maximizes the signal-to-noise ratio (SNR) at each TF bin.

Another popular approach to multichannel source separation is nonlinear TF *hard* masking based on the sparseness (disjointness) of source spectrograms [1]–[8]. TF bins can be classified by using the complex Gaussian mixture model [6]–[8]. The complex Gaussian distribution uses the absolute phases and signal energies of sources. The complex Watson mixture distribution is used to describe the phase and level differences between microphones [2]–[5]. If each TF bin is assigned to one of the sources independently, the permutation ambiguity arises as in ICA. To avoid this problem, Otsuka *et al.* [8] proposed a Bayesian *mixture* method inspired by latent Dirichlet allocation (LDA) in which each TF bin is exclusively assigned to one of the sources, each of which is further exclusively assigned to one of the directions. The impulse responses measured in an anechoic room can be used as prior knowledge for joint source separation and localization.

C. Integrated Source and Spatial Models

In multichannel extensions of NMF, the source model and spatial models are both described as factor models [12]–[17], i.e., the power spectrogram of each source signal is given by the sum of products of basis spectra and their activations and the complex spectrograms of observed multichannel signals are given by the sum of those of propagated source signals. Ozerov *et al.* [12] pioneered the use of NMF for multichannel source separation, where the spatial covariance matrices are restricted to rank-1 matrices and the EM or multiplicative update (MU) algorithm is used for minimizing the cost function based on the Itakura-Saito (IS) divergence. This model was extended to have full-rank spatial covariance matrices [13]. Sawada *et al.* [14] introduced partitioning parameters to have a set of basis spectra shared by all sources and derived a majorization-minimization (MM) algorithm. Nikunen and Virtanen [15] proposed a similar model that represents the spatial covariance matrix of each source as the weighted sum of all possible direction-dependent covariance matrices and used the MM algorithm for minimizing the cost function based on the Euclidean distance. Kitamura *et al.* [16] modified Sawada's model by restricting spatial covariance matrices to rank-1 matrices, resulting in a unified model of NMF and IVA.

Some studies have investigated a hybrid of factor and mixture models corresponding to source and spatial models (factor-mixture) [19], [28]. More specifically, the power spectrogram of each source signals is given by the sum of products

[1]The multichannel complex spectrogram of source signals captured by a microphone array is called an *image*.

of basis spectra and activations, and each TF bin of the complex spectrograms of observed multichannel signals are given by one of sources. Our main contribution is to investigate and experimentally compare Bayesian versions of all the four combinations of source and spatial models in terms of source separation performance and computation time.

III. Proposed Methods

In this section we formulate the four possible unified models (factor-factor, mixture-factor, factor-mixture, and mixture-mixture models) because the source and spatial models can each be represented as either a factor or mixture model.

A. Problem Specification

All signals are represented in the TF domain using short-time Fourier transform (STFT). Suppose that K sources are observed with M microphones. Each TF bin of the complex spectrograms of observed signals and that of the complex spectrograms of source signals are defined as

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

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

Similarly, each TF bin of the complex spectrograms of the *latent* signals corresponding to source k is defined as

$$\boldsymbol{x}_{tfk} = [x_{tfk1}, \ldots, x_{tfkM}]^T \in \mathbb{C}^M, \tag{3}$$

Note that $\boldsymbol{x}_{tfkm}$ can be directly observed by microphone m at time t and frequency f if only source k exists in a recording environment. Given observed data $\{\boldsymbol{x}_{tf}\}_{t=1,f=1}^{T,F}$, our goal is to estimate $\{\boldsymbol{x}_{tfk}\}_{t=1,f=1,k=1}^{T,F,K}$.

B. Source Models

We formulate two source models representing the generative process of source power spectrograms $\boldsymbol{\lambda} = \{\lambda_{tfk}\}_{t,f,k=1}^{T,F,K}$. $\boldsymbol{\lambda}$ is generated using basis spectra $\boldsymbol{W} = \{w_{klf}\}_{k,l,f=1}^{K,L,F}$ and activations $\boldsymbol{H} = \{h_{klt}\}_{k,l,t=1}^{K,L,T}$.

1) Source Factor Model: The power spectrogram of each source signal is decomposed into basis spectra and temporal activations via low-rank factorization as follows:

$$\lambda_{tfk} = \sum_{l=1}^{L} w_{klf} h_{klt}, \tag{4}$$

where $\boldsymbol{w}_{kl} = [w_{kl1}, \ldots, w_{klF}]^T$ is the l-th basis spectrum of source k and $\boldsymbol{h}_{kl} = [h_{kl1}, \ldots, h_{klT}]^T$ is the activation of basis l of source k at each time.

2) Source Mixture Model: The power spectrogram of each source signal is decomposed into basis spectra, temporal activations, and binary masks. Assuming the sparseness of the power spectrogram of each basis, one of the bases is considered to be dominant at each TF bin as follows:

$$\lambda_{tfk} = \prod_{l=1}^{L} (w_{klf} h_{klt})^{u_{tfkl}}, \tag{5}$$

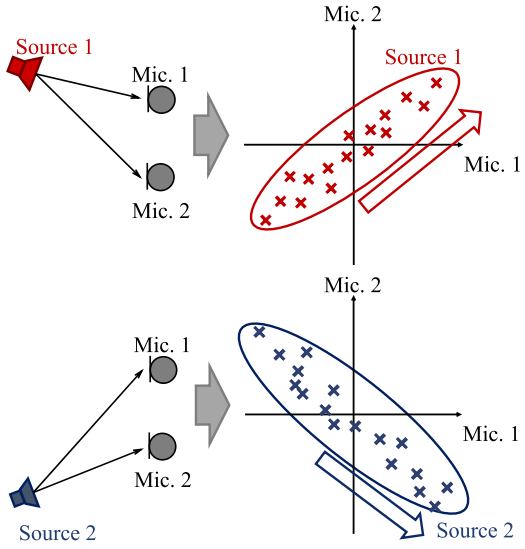


Fig. 6. The distribution of multichannel complex spectra $\{\boldsymbol{x}_{tfk} \in \mathbb{C}^M\}_{t=1}^{T}$ of each source k in frequency f ($M = 2$ and $K = 2$). Each axis indicates one of M channels, an ellipse indicates a spatial covariance matrix $\boldsymbol{A}_{fk}$, and an arrow indicates the direction of a steering vector $\boldsymbol{a}_{fk}$.

where $\boldsymbol{u}_{tfk} = [u_{tfk1}, \ldots, u_{tfkL}]^T$ is a one-hot vector for binary masking, i.e., u_{tfkl} takes 1 when basis l is selected (dominant) as the frequency component of source k at time t and frequency f and otherwise takes 0. We assume $\boldsymbol{u}_{tfk}$ to follow a categorical distribution as follows:

$$\boldsymbol{u}_{tfk} | \boldsymbol{\psi}_{tk} \sim \text{Categorical}(\boldsymbol{\psi}_{tk}), \tag{6}$$

where $\boldsymbol{\psi}_{tk} \in \mathbb{R}_+^L$ is a parameter such that $\sum_{l=1}^{L} \psi_{tkl} = 1$.

C. Spatial Models

We formulate two spatial models representing the generative process of observed spectrograms $\boldsymbol{X} = \{\boldsymbol{x}_{tf}\}_{t,f=1}^{T,F}$. Assuming an instantaneous mixing process in the frequency domain, the observation $\boldsymbol{x}_{tfk}$ is represented using a source spectrum $y_{tfk} \in \mathbb{C}$ of source k at time t and frequency f and a steering vector $\boldsymbol{a}_{fk} \in \mathbb{C}^M$ of source k at frequency f as follows:

$$\boldsymbol{x}_{tfk} = \boldsymbol{a}_{fk} y_{tfk}. \tag{7}$$

As in [8], [29], [30], we assume that y_{tfk} to follow a zero-mean complex Gaussian distribution as follows:

$$y_{tfk} \mid \lambda_{tfk} \sim \mathcal{N}_{\mathbb{C}}(0, \lambda_{tfk}), \tag{8}$$

where λ_{tfk} is a power spectrum density of source k at time t and frequency f. Using (7) and (8), the observation $\boldsymbol{x}_{tfk}$ is found to follow a multivariate complex Gaussian distribution as follows (Fig. 6):

$$\boldsymbol{x}_{tfk} \mid \lambda_{tfk}, \boldsymbol{A}_{fk} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \lambda_{tfk} \boldsymbol{A}_{fk}), \tag{9}$$

where $\boldsymbol{A}_{fk} = \boldsymbol{a}_{fk} \boldsymbol{a}_{fk}^H + \epsilon \boldsymbol{I}$ is a spatial covariance matrix of source k at frequency f, $*^H$ denotes Hermitian conjugate, $\boldsymbol{I}$ is the identity matrix, and $\epsilon > 0$ is a small number to make $\boldsymbol{A}_{fk}$ a positive definite matrix to avoid the degenerate distribution.

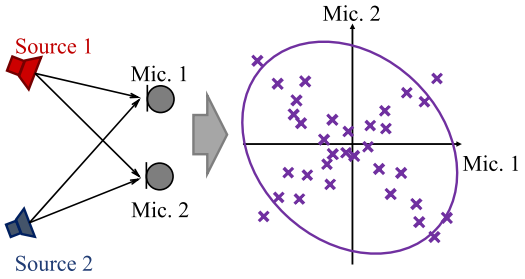


Fig. 7. Factor modeling of multichannel complex spectra $\{\boldsymbol{x}_{tf} \in \mathbb{C}^M\}_{t=1}^T$ in frequency f. Each spectrum $\boldsymbol{x}_{tf}$ is stochastically generated from a Gaussian distribution whose covariance matrix is the sum of spatial covariance matrices given by $\sum_{k=1}^K \lambda_{tfk}\boldsymbol{A}_{fk}$.

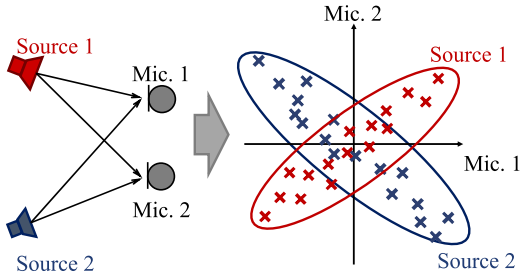


Fig. 8. Mixture modeling of multichannel complex spectra $\{\boldsymbol{x}_{tf} \in \mathbb{C}^M\}_{t=1}^T$ in frequency f. Each spectrum $\boldsymbol{x}_{tf}$ is stochastically generated from a mixture of K Gaussian distributions whose covariance matrices are given by $\{\lambda_{tfk}\boldsymbol{A}_{fk}\}_{k=1}^K$, respectively.

1) Spatial Factor Model: Assuming that an observed spectrum at each TF bin is given by the sum (instantaneous mixture) of latent source spectra, the observation $\boldsymbol{x}_{tf}$ is given by

$$\boldsymbol{x}_{tf} = \sum_{k=1}^K \boldsymbol{x}_{tfk} = \sum_{k=1}^K \boldsymbol{a}_{fk} y_{tfk}. \tag{10}$$

Using (9) and (10) and the reproductive property of the Gaussian distribution, the observation $\boldsymbol{x}_{tf}$ follows a complex Gaussian distribution as follows (Fig. 7):

$$\boldsymbol{x}_{tf} \mid \boldsymbol{\lambda}, \boldsymbol{A} \sim \mathcal{N}_{\mathbb{C}}\left(\boldsymbol{0}, \sum_{k=1}^K \lambda_{tfk}\boldsymbol{A}_{fk}\right). \tag{11}$$

To estimate the direction of each source k, the spatial covariance matrix $\boldsymbol{A}_{fk}$ is further factorized as follows [15]:

$$\boldsymbol{A}_{fk} = \sum_{d=1}^D r_{kd}\boldsymbol{G}_{fd}, \tag{12}$$

where $\boldsymbol{G}_{fd}$ is a spatial covariance matrix for direction d at frequency f and r_{kd} is the weight of $\boldsymbol{G}_{fd}$ in $\boldsymbol{A}_{fk}$. Although each source k has a particular direction and $\boldsymbol{r}_k = [r_{k1}, \ldots, r_{kD}]^T$ should be a one-hot vector in theory, in factor modeling, r_{kd} is allowed to take continuous values. As in [15], source localization can be performed by representing each source as the weighted sum of different directions. We let $\boldsymbol{R} = \{r_{kd}\}_{k,d=1}^{K,D}$. Plugging (12) into (11), $\boldsymbol{x}_{tf}$ is represented as a *unified* factor model given by

$$\boldsymbol{x}_{tf} \mid \boldsymbol{\lambda}, \boldsymbol{R}, \boldsymbol{G} \sim \mathcal{N}_{\mathbb{C}}\left(\boldsymbol{0}, \sum_{k=1}^K \sum_{d=1}^D \lambda_{tfk} r_{kd}\boldsymbol{G}_{fd}\right). \tag{13}$$

2) Spatial Mixture Model: If source spectra are sparse, only one of the sources tends to be observed at each TF bin. To specify the observed source at time t and frequency f, we define a one-hot vector $\boldsymbol{z}_{tf} = [z_{tf1}, \ldots, z_{tfK}]^T$ such that z_{tfk} takes 1 when source k is observed at time t and frequency f and otherwise takes 0. We let $\boldsymbol{Z} = \{\boldsymbol{z}_{tf}\}_{t,f=1}^{T,F}$. Using this assumption, the observation $\boldsymbol{x}_{tf}$ is given by

$$\boldsymbol{x}_{tf} = \prod_{k=1}^K \boldsymbol{x}_{tfk}^{z_{tfk}} = \prod_{k=1}^K (\boldsymbol{a}_{fk} y_{tfk})^{z_{tfk}}. \tag{14}$$

Equations (9) and (14) indicate that $\boldsymbol{x}_{tf}$ is generated by one of K Gaussian distributions, i.e., $\boldsymbol{x}_{tf}$ follows a mixture of multivariate complex Gaussian distributions as follows (Fig. 8):

$$\boldsymbol{x}_{tf} \mid \boldsymbol{\lambda}, \boldsymbol{A}, \boldsymbol{Z} \sim \prod_{k=1}^K \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \lambda_{tfk}\boldsymbol{A}_{fk})^{z_{tfk}}. \tag{15}$$

To associate $\boldsymbol{A}_{fk}$ with one of D directions, (15) can be extended to a *unified* mixture model given by

$$\boldsymbol{x}_{tf} \mid \boldsymbol{\lambda}, \boldsymbol{G}, \boldsymbol{Z}, \boldsymbol{S} \sim \prod_{k=1}^K \prod_{d=1}^D \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \lambda_{tfk}\boldsymbol{G}_{fd})^{z_{tfk}s_{kd}}, \tag{16}$$

where $\boldsymbol{s}_k = [s_{k1}, \ldots, s_{kD}]^T$ is a one-hot vector such that s_{kd} takes 1 when source k exists in direction d. We assume $\boldsymbol{z}_{tf}$ and $\boldsymbol{s}_k$ to follow categorical distributions as follows:

$$\boldsymbol{z}_{tf} | \boldsymbol{\pi}_t \sim \text{Categorical}(\boldsymbol{\pi}_t), \tag{17}$$

$$\boldsymbol{s}_k | \boldsymbol{\phi} \sim \text{Categorical}(\boldsymbol{\phi}), \tag{18}$$

where $\boldsymbol{\pi}_t \in \mathbb{R}_+^K$ and $\boldsymbol{\phi} \in \mathbb{R}_+^D$ are parameters to be estimated such that $\sum_{k=1}^K \pi_{tk} = 1$ and $\sum_{d=1}^D \phi_d = 1$.

D. Combinations of Source and Spatial Models

We formulate four unified models and explain its Bayesian treatment based on prior distributions.

1) Factor–Factor Model: Both the source and spatial models are formulated as factor models [17]. This model is an extension of [13] that further decomposes the spatial covariance matrix of each source into the weighted sum of direction-dependent matrices for joint source separation and localization as in [15]. Note that [15] is based on the minimization of the Euclidean distance while the other variants [12]–[14], [16], [17] including our method are based on the minimization of the IS divergence (the maximization of the complex Gaussian likelihood). In addition, our model is different from MNMF [14] and its rank-1 version [16] in that each source k is forced to have a unique set of L basis spectra. Substituting (4) into (13), the likelihood of the model parameters $\boldsymbol{W}$, $\boldsymbol{H}$, $\boldsymbol{R}$, and $\boldsymbol{G}$ for the observation $\boldsymbol{x}_{tf}$

is given by

$$\boldsymbol{x}_{tf} \mid \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{R}, \boldsymbol{G} \sim \mathcal{N}_{\mathbb{C}}\left(\boldsymbol{0}, \sum_{k=1}^{K}\sum_{d=1}^{D}\sum_{l=1}^{L} w_{klf} h_{klt} r_{kd} \boldsymbol{G}_{fd}\right). \quad (19)$$

2) Mixture–Factor Model: The source model is represented as a mixture model and the spatial model as a factor model. Substituting (5) into (13), the complete likelihood of the model parameters $\boldsymbol{W}$, $\boldsymbol{H}$, $\boldsymbol{R}$, $\boldsymbol{U}$, $\boldsymbol{G}$ for $\boldsymbol{x}_{tf}$ is given by

$$\boldsymbol{x}_{tf} \mid \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{R}, \boldsymbol{U}, \boldsymbol{G} \sim \mathcal{N}_{\mathbb{C}}\left(\boldsymbol{0}, \sum_{k=1}^{K}\sum_{d=1}^{D}\left(\prod_{l=1}^{L}(w_{klf} h_{klt})^{u_{tfkl}}\right) r_{kd} \boldsymbol{G}_{fd}\right). \quad (20)$$

In addition, the likelihood of $\boldsymbol{\psi} = \{\boldsymbol{\psi}_{tk}\}_{t=1,k=1}^{T,K}$ for the latent variables $\boldsymbol{U}$ given by (6) should also be considered.

3) Factor–Mixture Model: As in [19], substituting (4) into (16), the likelihood of the model parameters $\boldsymbol{W}$, $\boldsymbol{H}$, $\boldsymbol{Z}$, $\boldsymbol{S}$, and $\boldsymbol{G}$ for the observation $\boldsymbol{x}_{tf}$ is given by

$$\boldsymbol{x}_{tf} \mid \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{Z}, \boldsymbol{S}, \boldsymbol{G} \sim \prod_{k=1}^{K}\prod_{d=1}^{D}\mathcal{N}_{\mathbb{C}}\left(\boldsymbol{0}, \sum_{l=1}^{L} w_{klf} h_{klt} \boldsymbol{G}_{fd}\right)^{z_{tfk} s_{kd}}. \quad (21)$$

In addition, the likelihoods of $\boldsymbol{\pi} = \{\boldsymbol{\pi}_t\}_{t=1}^{T}$ and $\boldsymbol{\phi}$ for the latent variables $\boldsymbol{Z}$ and $\boldsymbol{S}$ given by (17) and (18) should also be considered.

4) Mixture–Mixture Model: Substituting (5) into (16), the likelihood of the model parameters $\boldsymbol{W}$, $\boldsymbol{H}$, $\boldsymbol{Z}$, $\boldsymbol{S}$, $\boldsymbol{U}$, and $\boldsymbol{G}$ for the observation $\boldsymbol{x}_{tf}$ is given by

$$\begin{aligned}\boldsymbol{x}_{tf} \mid \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{Z}, \boldsymbol{S}, \boldsymbol{U}, \boldsymbol{G} &\sim \prod_{k=1}^{K}\prod_{d=1}^{D}\mathcal{N}_{\mathbb{C}}\left(\boldsymbol{0}, \prod_{l=1}^{L}(w_{klf} h_{klt})^{u_{tfkl}} \boldsymbol{G}_{fd}\right)^{z_{tfk} s_{kd}} \\ &= \prod_{k=1}^{K}\prod_{d=1}^{D}\prod_{l=1}^{L}\mathcal{N}_{\mathbb{C}}\left(\boldsymbol{0}, w_{klf} h_{klt} \boldsymbol{G}_{fd}\right)^{z_{tfk} s_{kd} u_{tfkl}}. \quad (22)\end{aligned}$$

In addition, the likelihoods of $\boldsymbol{\pi}$, $\boldsymbol{\phi}$, and $\boldsymbol{\psi}$ for the latent variables $\boldsymbol{Z}$,$\boldsymbol{S}$, and $\boldsymbol{U}$ given by (17), (18), and (6) should also be considered.

E. Bayesian Extensions

For mathematical convenience, we put conjugate prior distributions on model parameters as follows (some of the prior distributions might not be used):

$$w_{klf} \sim \mathrm{Gamma}(a_0^w, b_0^w), \quad (23)$$

$$h_{klt} \sim \mathrm{Gamma}(a_0^h, b_0^h), \quad (24)$$

$$r_{kd} \sim \mathrm{Gamma}(a_0^r, b_0^r), \quad (25)$$

$$\boldsymbol{G}_{fd} \sim \mathcal{IW}_{\mathbb{C}}(\nu_0, \boldsymbol{G}_{fd}^0), \quad (26)$$

$$\boldsymbol{\pi}_t \sim \mathrm{Dirichlet}(a_0^{\pi} \boldsymbol{1}_K), \quad (27)$$

$$\boldsymbol{\phi} \sim \mathrm{Dirichlet}(a_0^{\phi} \boldsymbol{1}_D), \quad (28)$$

$$\boldsymbol{\psi}_{tk} \sim \mathrm{Dirichlet}(a_0^{\psi} \boldsymbol{1}_L), \quad (29)$$

where $a_0^* > 0$, $b_0^* > 0$, $\nu_0 > 0$, $\boldsymbol{G}_{fd}^0 \succ \boldsymbol{0}$ are hyperparameters that should be set in advance. a_0^* and b_0^* denote the shape and rate parameters of the gamma distribution, $\boldsymbol{1}_N$ denotes an N-dimensional vector with all elements one, $\mathcal{IW}_{\mathbb{C}}$ indicates the complex inverse Wishart distribution [31] given by

$$\mathcal{IW}_{\mathbb{C}}(\boldsymbol{G}|\nu, \boldsymbol{G}^0) = \frac{|\boldsymbol{G}^0|^{\nu} \exp(-\mathrm{tr}(\boldsymbol{G}^0 \boldsymbol{G}^{-1}))}{\pi^{M(M-1)/2} |\boldsymbol{G}|^{\nu - M} \prod_{m=0}^{M-1} \Gamma(\nu - m)}, \quad (30)$$

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

Note that source separation can be performed by using only a spatial model given by (13) or (16) (called a NA-factor or NA-mixture model). In this case, the power spectrum density λ_{tfk} is not decomposed and the conjugate prior distribution is put as follows:

$$\lambda_{tfk} \sim \mathrm{Gamma}(a_0^{\lambda}, b_0^{\lambda}), \quad (31)$$

where $a_0^{\lambda} > 0$ and $b_0^{\lambda} > 0$ are hyperparameters.

F. Source Separation

We estimate all parameters in a Bayesian manner and then recover source signals using the estimated parameters. When the spatial model is represented as a mixture model, sound sources are restored by applying a soft TF mask corresponding to a certain direction. Each mask is estimated using samples of latent variables $\boldsymbol{Z}$ and $\boldsymbol{S}$ given by Gibbs sampling. Multichannel source spectrum $\tilde{\boldsymbol{x}}_{tfd}$ is restored as follows:

$$\tilde{\boldsymbol{x}}_{tfd} = \frac{1}{I}\sum_{i=1}^{I}\sum_{k=1}^{K} z_{tfk}^{(i)} s_{kd}^{(i)} \boldsymbol{x}_{tf}, \quad (32)$$

where $z_{tfk}^{(i)}$ and $s_{kd}^{(i)}$ are the i-th samples of z_{tfk} and s_{kd} and I is the number of samples obtained by Gibbs sampling. The factor $\frac{1}{I}\sum_{i=1}^{I}\sum_{k=1}^{K} z_{tfk}^{(i)} s_{kd}^{(i)}$ is a contribution of direction d at time t and frequency f.

When the spatial model is represented as a factor model, sound sources are restored by using a multichannel Wiener filter [14]. The multichannel mixture spectrum $\boldsymbol{x}_{tf}$ at time frame t and frequency bin f is decomposed into the sum of multichannel

source spectra $\{\tilde{\boldsymbol{x}}_{tfk}\}_{k=1}^{K}$ as follows:

$$\tilde{\boldsymbol{x}}_{tfk} = \boldsymbol{Y}_{tfk}\left(\sum_{k'}\boldsymbol{Y}_{tfk'}\right)^{-1}\boldsymbol{x}_{tf}, \tag{33}$$

where we let $\boldsymbol{Y}_{tfk} = \sum_{ld} w_{klf}h_{klt}r_{kd}\boldsymbol{G}_{fd}$ in the factor-factor model or $\boldsymbol{Y}_{tfk} = \sum_{d}\prod_{l}(w_{klf}h_{klt})^{u_{tfkl}}r_{kd}\boldsymbol{G}_{fd}$ in the mixture-factor model.

IV. Bayesian Inference

Our goal is to calculate the posterior distribution and find optimal parameters that maximize the posterior distribution in practice. Since the posterior distribution is analytically intractable, but the posterior distribution of each parameter conditioned on the remaining parameters e.g., $p(\boldsymbol{W}|\boldsymbol{H},\boldsymbol{G},\boldsymbol{R},\boldsymbol{X})$, is tractable, we can use a Gibbs sampling algorithm [32] that alternately and iteratively updates one of the parameters according to the conditional posterior distribution by fixing the other parameters.

A. Factor–Factor Model

The factor-factor model estimates four model parameters $\boldsymbol{W}$, $\boldsymbol{H}$, $\boldsymbol{R}$ and $\boldsymbol{G}$. Since it is hard to directly draw samples from the conditional posterior distributions, all parameters are updated using a variational approach [33]. More specifically, the log-likelihood function defined by (19) 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. The log-likelihood is given by

$$\log p(\boldsymbol{X}|\boldsymbol{W},\boldsymbol{H},\boldsymbol{R},\boldsymbol{G}) = \sum_{tf}\left(-\log|\boldsymbol{Y}_{tf}| - \mathrm{tr}(\boldsymbol{X}_{tf}\boldsymbol{Y}_{tf}^{-1})\right) + C_1, \tag{34}$$

where $\boldsymbol{Y}_{tfkld} = w_{klf}h_{klt}r_{kd}\boldsymbol{G}_{fd}$, $\boldsymbol{Y}_{tf} = \sum_{kld}\boldsymbol{Y}_{tfkld}$, and C_1 is a constant. To derive a lower bound from (34), we use two inequalities proposed in [33]. 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:

$$-\log|\boldsymbol{Z}| \geq -\log|\boldsymbol{\Omega}| - \mathrm{tr}(\boldsymbol{\Omega}^{-1}\boldsymbol{Z}) + M, \tag{35}$$

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 an inequality given by

$$-\mathrm{tr}\left(\left(\textstyle\sum_{k=1}^{K}\boldsymbol{Z}_k\right)^{-1}\boldsymbol{A}\right) \geq -\sum_{k=1}^{K}\mathrm{tr}\left(\boldsymbol{Z}_k^{-1}\boldsymbol{\Phi}_k\boldsymbol{A}\boldsymbol{\Phi}_k^{H}\right), \tag{36}$$

where $\{\boldsymbol{Z}_k\}_{k=1}^{K}$ 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 (35) and (36), the log-likelihood function given by (34) is lower-bounded by $\mathcal{L}_1$ as follows:

$$\begin{aligned}&\log p(\boldsymbol{X}|\boldsymbol{W},\boldsymbol{H},\boldsymbol{R},\boldsymbol{G})\\ &\geq \sum_{tf}\left(-\mathrm{tr}\left(\boldsymbol{Y}_{tf}\boldsymbol{\Omega}_{tf}^{-1}\right) - \log|\boldsymbol{\Omega}_{tf}| + M\right)\\ &\quad - \sum_{tfkld}\mathrm{tr}\left(\boldsymbol{Y}_{tfkld}^{-1}\boldsymbol{\Phi}_{tfkld}\boldsymbol{X}_{tf}\boldsymbol{\Phi}_{tfkld}^{H}\right) + C_1 \stackrel{\mathrm{def}}{=} \mathcal{L}_1,\end{aligned} \tag{37}$$

where $\boldsymbol{\Omega}_{tf}$ and $\boldsymbol{\Phi}_{tfkld}$ are newly introduced auxiliary variables. The auxiliary function $\mathcal{L}_1$ is maximized (i.e., the equality holds) when $\boldsymbol{\Omega}_{tf}$ and $\boldsymbol{\Phi}_{tfkld}$ are given by

$$\boldsymbol{\Omega}_{tf} = \boldsymbol{Y}_{tf} \quad \text{and} \quad \boldsymbol{\Phi}_{tfkld} = \boldsymbol{Y}_{tfkld}\boldsymbol{Y}_{tf}^{-1}. \tag{38}$$

The parameters w_{klf}, h_{klt}, r_{kd}, and $\boldsymbol{G}_{fd}$ can be sampled from conditional distributions proportional to the product of (23)–(26) and (37) as follows:

$$w_{klf} \mid \boldsymbol{X},\boldsymbol{\Theta}_{\neg w_{klf}} \sim \mathrm{GIG}(\gamma_{klf}^{w},\rho_{klf}^{w},\tau_{klf}^{w}), \tag{39}$$

$$h_{klt} \mid \boldsymbol{X},\boldsymbol{\Theta}_{\neg h_{klt}} \sim \mathrm{GIG}(\gamma_{klt}^{h},\rho_{klt}^{h},\tau_{klt}^{h}), \tag{40}$$

$$r_{kd} \mid \boldsymbol{X},\boldsymbol{\Theta}_{\neg r_{kd}} \sim \mathrm{GIG}(\gamma_{kd}^{r},\rho_{kd}^{r},\tau_{kd}^{r}), \tag{41}$$

$$\boldsymbol{G}_{fd} \mid \boldsymbol{X},\boldsymbol{\Theta}_{\neg \boldsymbol{G}_{fd}} \sim \mathrm{MGIG}_{\mathbb{C}}(\nu_{fd},\boldsymbol{Q}_{fd},\boldsymbol{V}_{fd}), \tag{42}$$

where $\boldsymbol{\Theta}$ is the set of all parameters, $\boldsymbol{\Theta}_{\neg *}$ indicates the set of all parameters except $*$, and GIG and $\mathrm{MGIG}_{\mathbb{C}}$ indicate the generalized inverse Gaussian distribution [34] and the complex matrix GIG distribution [35], defined by

$$\mathrm{GIG}(x|\gamma,\rho,\tau) = \frac{\exp\{(\gamma-1)\log x - \rho x - \tau/x\}\rho^{\gamma/2}}{2\tau^{\gamma/2}\mathcal{K}_{\gamma}(2\sqrt{\rho\tau})}, \tag{43}$$

$$\begin{aligned}&\mathrm{MGIG}_{\mathbb{C}}(\boldsymbol{X}|\gamma,\boldsymbol{Q},\boldsymbol{V})\\ &\qquad\propto |\boldsymbol{X}|^{\gamma-M}\exp\{-\mathrm{tr}(\boldsymbol{Q}\boldsymbol{X}+\boldsymbol{V}\boldsymbol{X}^{-1})\},\end{aligned} \tag{44}$$

where $\mathcal{K}_{\gamma}$ is the modified Bessel function of the second kind, γ is a real number, $\rho>0$, $\tau>0$, $\boldsymbol{Q}\succ\boldsymbol{0}$ and $\boldsymbol{V}\succ\boldsymbol{0}$. To draw samples from the GIG, we use a rejection sampling method [36]. To draw samples from the complex MGIG distributions, we use a Metropolis-Hastings (MH) sampler [35] that uses as a proposal distribution a complex Wishart distribution having the same mode as a target MGIG distribution (the mode of the MGIG distribution can be calculated by using an algebraic Riccati equation). The conditional posterior parameters γ_*^{*}, ρ_*^{*}, τ_*^{*}, $\boldsymbol{Q}_{fd}$, and $\boldsymbol{V}_{fd}$ are given by

$$\gamma_{klf}^{w} = a_0^{w}, \tag{45}$$

$$\rho_{klf}^{w} = b_0^{w} + \sum_{td}h_{klt}r_{kd}\mathrm{tr}(\boldsymbol{G}_{fd}\boldsymbol{\Omega}_{tf}^{-1}), \tag{46}$$

$$\tau^w_{klf} = \sum_{td} h^{-1}_{klt} r^{-1}_{kd} \mathrm{tr}\left(\boldsymbol{G}^{-1}_{fd} \boldsymbol{\Phi}_{tfkld} \boldsymbol{X}_{tf} \boldsymbol{\Phi}^H_{tfkld}\right), \tag{47}$$

$$\gamma^h_{klt} = a^h_0, \tag{48}$$

$$\rho^h_{klt} = b^h_0 + \sum_{fd} w_{klf} r_{kd} \mathrm{tr}(\boldsymbol{G}_{fd} \boldsymbol{\Omega}^{-1}_{tf}), \tag{49}$$

$$\tau^h_{klt} = \sum_{fd} w^{-1}_{klf} r^{-1}_{kd} \mathrm{tr}\left(\boldsymbol{G}^{-1}_{fd} \boldsymbol{\Phi}_{tfkld} \boldsymbol{X}_{tf} \boldsymbol{\Phi}^H_{tfkld}\right), \tag{50}$$

$$\gamma^r_{kd} = a^r_0, \tag{51}$$

$$\rho^r_{kd} = b^r_0 + \sum_{tfl} w_{klf} h_{klt} \mathrm{tr}(\boldsymbol{G}_{fd} \boldsymbol{\Omega}^{-1}_{tf}), \tag{52}$$

$$\tau^r_{kd} = \sum_{tfl} w^{-1}_{klf} h^{-1}_{klt} \mathrm{tr}\left(\boldsymbol{G}^{-1}_{fd} \boldsymbol{\Phi}_{tfkld} \boldsymbol{X}_{tf} \boldsymbol{\Phi}^H_{tfkld}\right), \tag{53}$$

$$\nu_{fd} = \nu_0, \tag{54}$$

$$\boldsymbol{Q}_{fd} = \sum_{tkl} w_{klf} h_{klt} r_{kd} \boldsymbol{\Omega}^{-1}_{tf}, \tag{55}$$

$$\boldsymbol{V}_{fd} = \boldsymbol{G}^0_{fd} + \sum_{tkl} w^{-1}_{klf} h^{-1}_{klt} r^{-1}_{kd} \boldsymbol{\Phi}_{tfkld} \boldsymbol{X}_{tf} \boldsymbol{\Phi}^H_{tfkld}. \tag{56}$$

B. Mixture–Factor Model

The mixture-factor model estimates five model parameters $\boldsymbol{U}$, $\boldsymbol{W}$, $\boldsymbol{H}$, $\boldsymbol{R}$, and $\boldsymbol{G}$. While $\boldsymbol{U}$ can be directly sampled from the conditional posterior distributions, $\boldsymbol{W}$, $\boldsymbol{H}$, $\boldsymbol{R}$, and $\boldsymbol{G}$ are updated using a variational approach [33] like that used in the factor-factor model. The conditional posterior distribution of $\boldsymbol{U}$ is calculated by the product of the likelihood function given by (13) and the prior distributions given by (6) and (29) after the parameter $\boldsymbol{\psi}_{tk}$ is marginalized out as follows:

$$\boldsymbol{u}_{tfk} \mid \boldsymbol{X}, \boldsymbol{\Theta}_{\neg \boldsymbol{u}_{tfk}} \sim \mathrm{Categorical}(\boldsymbol{\psi}'_{tfk}), \tag{57}$$

where ψ'_{tfkl} is given by

$$\psi'_{tfkl} \propto (a_{0^\psi} + n^{\neg f}_{tkl}) \left| \boldsymbol{\Lambda}_{tfk} + \sum_d w_{klf} h_{klt} r_{kd} \boldsymbol{G}_{fd} \right| \exp\left(-\boldsymbol{x}^H_{tf} \left(\boldsymbol{\Lambda}_{tfk} + \sum_d w_{klf} h_{klt} r_{kd} \boldsymbol{G}_{fd} \right)^{-1} \boldsymbol{x}_{tf} \right), \tag{58}$$

where $\boldsymbol{\Lambda}_{tfk}$ is given by

$$\boldsymbol{\Lambda}_{tfk} = \sum_{d, k' \neq k} \left(\prod_l (w_{k'lf} h_{k'lt})^{u_{tfk'l}} \right) r_{kd} \boldsymbol{G}_{fd}. \tag{59}$$

Although the conditional posterior of each parameter $\boldsymbol{W}$, $\boldsymbol{H}$, $\boldsymbol{R}$ or $\boldsymbol{G}$ is proportional to the complete joint likelihood given by the product of (20), (23)–(26), it is difficult to directly get samples from the conditional posterior because of the complicated form of (20). The log-likelihood function defined by (20) is therefore lower-bounded by a tractable auxiliary function having auxiliary variables. Letting $\boldsymbol{Y}_{tfkd} = \prod_l (w_{klf} h_{klt})^{u_{tfkl}} r_{kd} \boldsymbol{G}_{fd}$ and $\boldsymbol{Y}_{tf} = \sum_{kd} \boldsymbol{Y}_{tfkd}$, the log-likelihood is given by

$$\log p(\boldsymbol{X} | \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{R}, \boldsymbol{G}) = \sum_{tf} \left(-\log |\boldsymbol{Y}_{tf}| - \mathrm{tr}(\boldsymbol{X}_{tf} \boldsymbol{Y}^{-1}_{tf}) \right) + C_2, \tag{60}$$

where C_2 is a constant. Using Inequalities (35) and (36), (60) is lower-bounded by $\mathcal{L}_2$ as follows:

$$\begin{aligned} &\log p(\boldsymbol{X} | \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{R}, \boldsymbol{G}, \boldsymbol{U}) \\ &\geq \sum_{tf} \left(-\mathrm{tr}\left(\boldsymbol{Y}_{tf} \boldsymbol{\Omega}^{-1}_{tf} \right) - \log |\boldsymbol{\Omega}_{tf}| + M \right) \\ &\quad - \sum_{tfkd} \mathrm{tr}\left(\boldsymbol{Y}^{-1}_{tfkd} \boldsymbol{\Phi}_{tfkd} \boldsymbol{X}_{tf} \boldsymbol{\Phi}^H_{tfkd} \right) + C_2 \stackrel{\mathrm{def}}{=} \mathcal{L}_2. \end{aligned} \tag{61}$$

The auxiliary function $\mathcal{L}_2$ is maximized, i.e., the equality holds, when $\boldsymbol{\Omega}_{tf}$ and $\boldsymbol{\Phi}_{tfkld}$ are given by

$$\boldsymbol{\Omega}_{tf} = \boldsymbol{Y}_{tf} \quad \text{and} \quad \boldsymbol{\Phi}_{tfkd} = \boldsymbol{Y}_{tfkd} \boldsymbol{Y}^{-1}_{tf}. \tag{62}$$

The parameters w_{klf}, h_{klt}, r_{kd}, and $\boldsymbol{G}_{fd}$ can be sampled from conditional distributions proportional to the product of (23)–(26) and (61) as follows:

$$w_{klf} \mid \boldsymbol{X}, \boldsymbol{\Theta}_{\neg w_{klf}} \sim \mathrm{GIG}(\gamma^w_{klf}, \rho^w_{klf}, \tau^w_{klf}), \tag{63}$$

$$h_{klt} \mid \boldsymbol{X}, \boldsymbol{\Theta}_{\neg h_{klt}} \sim \mathrm{GIG}(\gamma^h_{klt}, \rho^h_{klt}, \tau^h_{klt}), \tag{64}$$

$$r_{kd} \mid \boldsymbol{X}, \boldsymbol{\Theta}_{\neg r_{kd}} \sim \mathrm{GIG}(\gamma^r_{kd}, \rho^r_{kd}, \tau^r_{kd}), \tag{65}$$

$$\boldsymbol{G}_{fd} \mid \boldsymbol{X}, \boldsymbol{\Theta}_{\neg \boldsymbol{G}_{fd}} \sim \mathrm{MGIG}_{\mathbb{C}}(\nu_{fd}, \boldsymbol{Q}_{fd}, \boldsymbol{V}_{fd}). \tag{66}$$

The conditional posterior parameters γ^*_*, ρ^*_*, τ^*_*, $\boldsymbol{Q}_{fd}$, and $\boldsymbol{V}_{fd}$ are given by

$$\gamma^w_{klf} = a^w_0, \tag{67}$$

$$\rho^w_{klf} = b^w_0 + \sum_{td} u_{tfkl} h_{klt} r_{kd} \mathrm{tr}(\boldsymbol{G}_{fd} \boldsymbol{\Omega}^{-1}_{tf}), \tag{68}$$

$$\tau^w_{klf} = \sum_{td} u_{tfkl} h^{-1}_{klt} r^{-1}_{kd} \mathrm{tr}\left(\boldsymbol{G}^{-1}_{fd} \boldsymbol{\Phi}_{tfkd} \boldsymbol{X}_{tf} \boldsymbol{\Phi}^H_{tfkd}\right), \tag{69}$$

$$\gamma^h_{klt} = a^h_0, \tag{70}$$

$$\rho^h_{klt} = b^h_0 + \sum_{fd} u_{tfkl} w_{klf} r_{kd} \mathrm{tr}(\boldsymbol{G}_{fd} \boldsymbol{\Omega}^{-1}_{tf}), \tag{71}$$

$$\tau^h_{klt} = \sum_{fd} u_{tfkl} w^{-1}_{klf} r^{-1}_{kd} \mathrm{tr}\left(\boldsymbol{G}^{-1}_{fd} \boldsymbol{\Phi}_{tfkld} \boldsymbol{X}_{tf} \boldsymbol{\Phi}^H_{tfkld}\right), \tag{72}$$

$$\gamma^r_{klf} = a^r_0, \tag{73}$$

$$\rho^r_{kd} = b^r_0 + \sum_{tf} \left(\prod_l (w_{klf} h_{klt})^{u_{tfkl}} \right) \mathrm{tr}(\boldsymbol{G}_{fd} \boldsymbol{\Omega}^{-1}_{tf}), \tag{74}$$

$$\tau^r_{kd} = \sum_{tf} \left(\prod_l (w_{klf} h_{klt})^{u_{tfkl}} \right)^{-1} \mathrm{tr}\left(\boldsymbol{G}^{-1}_{fd} \boldsymbol{\Phi}_{tfkld} \boldsymbol{X}_{tf} \boldsymbol{\Phi}^H_{tfkd}\right), \tag{75}$$

$$\nu_{fd} = \nu_0, \tag{76}$$

$$\boldsymbol{Q}_{fd} = \sum_{tk} \left(\prod_l (w_{klf} h_{klt})^{u_{tfkl}} \right) r_{kd} \boldsymbol{\Omega}_{tf}^{-1}, \tag{77}$$

$$\boldsymbol{V}_{fd} = \sum_{tk} \left(\prod_l (w_{klf} h_{klt})^{u_{tfkl}} \right)^{-1} r_{kd}^{-1} \boldsymbol{\Phi}_{tfkd} \boldsymbol{X}_{tf} \boldsymbol{\Phi}_{tfkd}^H + \boldsymbol{G}_{fd}^0. \tag{78}$$

C. Factor–Mixture Model

The factor-mixture model estimates five model parameters $\boldsymbol{G}$, $\boldsymbol{Z}$, $\boldsymbol{S}$, $\boldsymbol{W}$ and $\boldsymbol{H}$. While $\boldsymbol{G}$, $\boldsymbol{Z}$, and $\boldsymbol{S}$ are updated by sampling from the conditional posterior distributions, $\boldsymbol{W}$ and $\boldsymbol{H}$ are updated using a variational approach [18]. The conditional posterior distributions of model parameters $\boldsymbol{G}$, $\boldsymbol{Z}$, and $\boldsymbol{S}$ are calculated by using the likelihood function ((16)) and the prior distributions ((18)–(28)) by marginalizing out parameters $\boldsymbol{\pi}_t$ and $\boldsymbol{\phi}$ as follows:

$$\boldsymbol{G}_{fd} \mid \boldsymbol{X}, \boldsymbol{\Theta}_{\neg \boldsymbol{G}_{fd}} \sim \mathcal{IW}_{\mathbb{C}}(\nu'_{fd}, \boldsymbol{G}'_{fd}), \tag{79}$$

$$\boldsymbol{z}_{tf} \mid \boldsymbol{X}, \boldsymbol{\Theta}_{\neg \boldsymbol{z}_{tf}} \sim \text{Categorical}(\boldsymbol{\pi}'_{tf}), \tag{80}$$

$$\boldsymbol{s}_k \mid \boldsymbol{X}, \boldsymbol{\Theta}_{\neg \boldsymbol{s}_k} \sim \text{Categorical}(\boldsymbol{\phi}'_k), \tag{81}$$

The conditional posterior parameters ν'_{fd}, $\boldsymbol{G}'_{fd}$, $\boldsymbol{\pi}'_{tf}$, and $\boldsymbol{\phi}'_k$ are given by

$$\nu'_{fd} = \nu_0 + \sum_{tk} z_{tfk} s_{kd}, \tag{82}$$

$$\boldsymbol{G}'_{fd} = \boldsymbol{G}^0_{fd} + \sum_{tk} \frac{\boldsymbol{x}_{tf} \boldsymbol{x}_{tf}{}^H}{\sum_l w_{klf} h_{klt}} z_{tfk} s_{kd}, \tag{83}$$

$$\pi'_{tfk} = \prod_d \left(\frac{1}{|\sum_l w_{klf} h_{klt} \boldsymbol{G}_{fd}|} \exp\left(- \frac{\boldsymbol{x}_{tf}^H \boldsymbol{G}_{fd}^{-1} \boldsymbol{x}_{tf}}{\sum_l w_{klf} h_{klt}} \right) \right)^{s_{kd}} \times (a_0^{\pi} + n_{tk}^{\neg f}), \tag{84}$$

$$\phi'_{kd} = \prod_{tf} \left(\frac{1}{|\sum_l w_{klf} h_{klt} \boldsymbol{G}_{fd}|} \exp\left(- \frac{\boldsymbol{x}_{tf}^H \boldsymbol{G}_{fd}^{-1} \boldsymbol{x}_{tf}}{\sum_l w_{klf} h_{klt}} \right) \right)^{z_{tfk}} \times (a_0^{\phi} + c_d^{\neg k}), \tag{85}$$

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

Although the conditional posterior of each parameter $\boldsymbol{W}$ or $\boldsymbol{H}$ is proportional to the complete joint likelihood given by the product of (21), (23), and (24), it is difficult to directly get samples from the conditional posterior because of the complicated form of (21). Therefore, the log-likelihood function defined by (21) is lower-bounded by a tractable auxiliary function having auxiliary variables. More specifically, letting $\lambda_{tfkl} = w_{klf} h_{klt}$, the log-likelihood is given by

$$\log p(\boldsymbol{X} \mid \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{Z}, \boldsymbol{S}, \boldsymbol{G}) = \sum_{tf} \sum_{kd} \left(- \log \left| \sum_l \lambda_{tfkl} \boldsymbol{G}_{fd} \right| - \text{tr}\left(\frac{\boldsymbol{X}_{tf} \boldsymbol{G}_{fd}^{-1}}{\sum_l \lambda_{tfkl}} \right) \right)^{z_{tfk} s_{kd}} + C_3, \tag{86}$$

where $\boldsymbol{X}_{tf} = \boldsymbol{x}_{tf}^H \boldsymbol{x}$ and C_3 is a constant. To derive a lower bound of (86), we use two inequalities used in [18]. First, for a convex function $f(z) = -\log|z|$, we use a first-order Taylor approximation around an arbitrary point α as follows:

$$- \log|z| \geq - \log|\alpha| - \frac{z}{\alpha} + 1, \tag{87}$$

where the equality holds when $\alpha = z$. Second, for a concave function $g(z) = -\frac{1}{z}$, we use Jensen's inequality that for any vector $\boldsymbol{\beta}$ such that $\beta_l \geq 0$ and $\sum_l \beta_l = 1$ as follows:

$$- \frac{1}{\sum_l z_l} = - \frac{1}{\sum_l \beta_l \frac{z_l}{\beta_l}} \geq - \sum_l \beta_l \frac{1}{\frac{z_l}{\beta_l}} = - \sum_l \beta_l^2 \frac{1}{z_l}, \tag{88}$$

where the equality holds when $\beta_l = \frac{z_l}{\sum_{l'} z_{l'}}$.

Using Inequalities (87) and (88), the log-likelihood function given by (86) is lower-bounded by $\mathcal{L}_3$ as follows:

$$\log p(\boldsymbol{X} \mid \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{Z}, \boldsymbol{S}, \boldsymbol{G}) \geq \sum_{tf} \sum_{kd} \left(M \left(- \log|\alpha_{tfk}| - \frac{\sum_l \lambda_{tfkl}}{\alpha_{tfk}} + 1 \right) - \log|\boldsymbol{G}_{fd}| - \text{tr}\left(\boldsymbol{X}_{tf} \boldsymbol{G}_{fd}^{-1} \right) \sum_l \beta_{tfkl}^2 \frac{1}{\lambda_{tfkl}} \right)^{z_{tfk} s_{kd}} + C_3 \stackrel{\text{def}}{=} \mathcal{L}_3, \tag{89}$$

where α_{tfk} and β_{tfkl} are newly introduced auxiliary variables. The auxiliary function $\mathcal{L}_3$ is maximized (i.e., the equality holds) when α_{tfk} and β_{tfkl} are given by

$$\alpha_{tfk} = \sum_l \lambda_{tfkl} \quad \text{and} \quad \beta_{tfkl} = \frac{\lambda_{tfkl}}{\sum_{l'} \lambda_{tfkl'}}. \tag{90}$$

The parameters w_{klf} and h_{klt} can be sampled from the following conditional distributions proportional to the product of (23), (24), and (89):

$$w_{klf} \mid \boldsymbol{X}, \boldsymbol{\Theta}_{\neg w_{klf}} \sim \text{GIG}(\gamma_{klf}^w, \rho_{kfl}^w, \tau_{kfl}^w), \tag{91}$$

$$h_{klt} \mid \boldsymbol{X}, \boldsymbol{\Theta}_{\neg h_{klt}} \sim \text{GIG}(\gamma_{klt}^h, \rho_{klt}^h, \tau_{klt}^h). \tag{92}$$

The conditional posterior parameters γ_*^*, ρ_*^*, and τ_*^* are

$$\gamma_{klf}^w = a_0^w, \tag{93}$$

$$\rho_{klf}^w = b_0^w + \sum_{td} \frac{M h_{klt}}{\alpha_{tfk}}, \tag{94}$$

$$\tau_{klf}^w = \sum_{td} \text{tr}(\boldsymbol{X}_{tf} \boldsymbol{G}_{fd}^{-1}) \frac{\beta_{tfkl}^2}{h_{klt}}, \tag{95}$$

$$\gamma_{klt}^h = a_0^h, \tag{96}$$

$$\rho_{klt}^h = b_0^h + \sum_{fd} \frac{M w_{klf}}{\alpha_{tfk}}, \tag{97}$$

$$\tau_{klt}^h = \sum_{fd} \text{tr}(\boldsymbol{X}_{tf} \boldsymbol{G}_{fd}^{-1}) \frac{\beta_{tfkl}^2}{w_{klf}}. \tag{98}$$

D. Mixture–Mixture Model

The mixture-mixture model estimates six model parameters $\boldsymbol{G}$, $\boldsymbol{Z}$, $\boldsymbol{S}$, $\boldsymbol{U}$, $\boldsymbol{W}$, and $\boldsymbol{H}$. All the parameters can be updated efficiently by directly sampling from the conditional posterior distributions. The conditional posterior of each parameter is proportional to the complete joint likelihood given by the product of (22)–(24), and (26)–(29), and samples are taken from each conditional posterior after parameters $\boldsymbol{\pi}_t$, $\boldsymbol{\phi}$, and $\boldsymbol{\psi}_{tk}$ are marginalized out as follows:

$$w_{klf} \mid \boldsymbol{X}, \boldsymbol{\Theta}_{\neg w_{klf}} \sim \mathrm{GIG}(\gamma^w_{klf}, \rho^w_{klf}, \tau^w_{klf}), \tag{99}$$

$$h_{klt} \mid \boldsymbol{X}, \boldsymbol{\Theta}_{\neg h_{klt}} \sim \mathrm{GIG}(\gamma^h_{klt}, \rho^h_{klt}, \tau^h_{klt}), \tag{100}$$

$$\boldsymbol{G}_{fd} \mid \boldsymbol{X}, \boldsymbol{\Theta}_{\neg \boldsymbol{G}_{fd}} \sim \mathcal{IW}_{\mathbb{C}}(\nu'_{fd}, \boldsymbol{G}'_{fd}), \tag{101}$$

$$\boldsymbol{z}_{tf} \mid \boldsymbol{X}, \boldsymbol{\Theta}_{\neg \boldsymbol{z}_{tf}} \sim \mathrm{Categorical}(\boldsymbol{\pi}'_{tf}), \tag{102}$$

$$\boldsymbol{s}_k \mid \boldsymbol{X}, \boldsymbol{\Theta}_{\neg \boldsymbol{s}_k} \sim \mathrm{Categorical}(\boldsymbol{\phi}'_k), \tag{103}$$

$$\boldsymbol{u}_{tfk} \mid \boldsymbol{X}, \boldsymbol{\Theta}_{\neg \boldsymbol{u}_{tfkl}} \sim \mathrm{Categorical}(\boldsymbol{\psi}'_{tfk}). \tag{104}$$

The conditional posterior parameters γ^*_*, ρ^*_*, τ^*_*, ν'_{fd}, $\boldsymbol{G}'_{fd}$, $\boldsymbol{\pi}'_{tf}$, $\boldsymbol{\phi}'_k$, and $\boldsymbol{\psi}'_{tfk}$ are given by

$$\gamma^w_{klf} = a^w_0 - M n_{fkl}, \tag{105}$$

$$\rho^w_{klf} = b^w_0, \tag{106}$$

$$\tau^w_{klf} = \sum_{td} \frac{\boldsymbol{x}^H_{tf} \boldsymbol{G}^{-1}_{fd} \boldsymbol{x}_{tf}}{h_{klt}} z_{tfk} s_{kd} u_{tfkl}, \tag{107}$$

$$\gamma^h_{klt} = a^h_0 - M n_{tkl}, \tag{108}$$

$$\rho^h_{klt} = b^h_0, \tag{109}$$

$$\tau^h_{klt} = \sum_{fd} \frac{\boldsymbol{x}^H_{tf} \boldsymbol{G}^{-1}_{fd} \boldsymbol{x}_{tf}}{w_{klf}} z_{tfk} s_{kd} u_{tfkl}, \tag{110}$$

$$\nu'_{fd} = \nu_0 + \sum_{tk} z_{tfk} s_{kd}, \tag{111}$$

$$\boldsymbol{G}'_{fd} = \boldsymbol{G}^0_{fd} + \sum_{tkl} \frac{\boldsymbol{x}_{tf} \boldsymbol{x}_{tf}{}^H}{w_{klf} h_{klt}} z_{tfk} s_{kd}, \tag{112}$$

$$\pi'_{tfk} = \prod_{ld} \left(\left| \frac{1}{w_{klf} h_{klt} \boldsymbol{G}_{fd}} \right| \exp\left(- \frac{\boldsymbol{x}^H_{tf} \boldsymbol{G}^{-1}_{fd} \boldsymbol{x}_{tf}}{w_{klf} h_{klt}} \right) \right)^{s_{kd} u_{tfkl}} \times (a^\pi_0 + n^{\neg f}_{tk}), \tag{113}$$

$$\phi'_{kd} = \prod_{tfl} \left(\left| \frac{1}{w_{klf} h_{klt} \boldsymbol{G}_{fd}} \right| \exp\left(- \frac{\boldsymbol{x}^H_{tf} \boldsymbol{G}^{-1}_{fd} \boldsymbol{x}_{tf}}{w_{klf} h_{klt}} \right) \right)^{z_{tfk} u_{tfkl}} \times (a^\phi_0 + c^{\neg k}_d), \tag{114}$$

$$\psi'_{tfkl} = \prod_{d} \left(\left| \frac{1}{w_{klf} h_{klt} \boldsymbol{G}_{fd}} \right| \exp\left(- \frac{\boldsymbol{x}^H_{tf} \boldsymbol{G}^{-1}_{fd} \boldsymbol{x}_{tf}}{w_{klf} h_{klt}} \right) \right)^{z_{tfk} s_{kd}} \times (a^\psi_0 + n^{\neg f}_{tkl}), \tag{115}$$

where n_{fkl} (n_{tkl}) indicates the number of TF bins at frequency f (frame t) assigned to source k and basis l, and $n^{\neg f}_{tkl}$ indicates the number of TF bins at frame t assigned to source k and basis l except for frequency f.

E. NA-Factor Model

The NA-factor model estimates three model parameters $\boldsymbol{\lambda}$, $\boldsymbol{R}$, and $\boldsymbol{G}$. r_{kd} and $\boldsymbol{G}_{fd}$ can be sampled from (41), (42), and (51)–(56) where $\sum_l w_{klf} h_{klt}$ is replaced with λ_{tfk}. In the same way as the factor-factor and mixture-factor models, the conditional posterior distribution of λ_{tfk} is given by

$$\lambda_{tfk} \sim \mathrm{GIG}(\gamma^\lambda_{tfk}, \rho^\lambda_{tfk}, \tau^\lambda_{tfk}). \tag{116}$$

The conditional posterior parameters are given by

$$\gamma^\lambda_{tfk} = a^\lambda_0, \tag{117}$$

$$\rho^\lambda_{tfk} = b^\lambda_0 + \sum_d r_{kd} \mathrm{tr}\left(\boldsymbol{G}_{fd} \boldsymbol{\Omega}^{-1}_{tf} \right), \tag{118}$$

$$\tau^\lambda_{tfk} = \sum_d r^{-1}_{kd} \mathrm{tr}\left(\boldsymbol{G}^{-1}_{fd} \boldsymbol{\Phi}_{tfkd} \boldsymbol{X}_{tf} \boldsymbol{\Phi}^H_{tfkd} \right), \tag{119}$$

where $\boldsymbol{\Omega}_{tf}$ and $\boldsymbol{\Phi}_{tfkd}$ are defined by $\boldsymbol{\Omega}_{tf} = \sum_{kd} \lambda_{tfk} r_{kd} \boldsymbol{G}_{fd}$ and $\boldsymbol{\Phi}_{tfkd} = \lambda_{tfk} r_{kd} \boldsymbol{G}_{fd} (\sum_{k'd'} \lambda_{tfk'} r_{k'd'} \boldsymbol{G}_{fd'})^{-1}$.

F. NA-Mixture Model

The NA-mixture model estimates three model parameters $\boldsymbol{\lambda}$, $\boldsymbol{G}$, $\boldsymbol{Z}$, and $\boldsymbol{S}$. $\boldsymbol{G}_{fd}$, z_{tfk} and s_{kd} can be sampled from (79)–(85) where $\sum_l w_{klf} h_{klt}$ is replaced with λ_{tfk} and λ_{tfk} can be sampled from a conditional posterior distribution given by

$$\lambda_{tfk} \sim \mathrm{GIG}(\gamma^\lambda_{tfk}, \rho^\lambda_{tfk}, \tau^\lambda_{tfk}). \tag{120}$$

The conditional posterior parameters are given by

$$\gamma^\lambda_{tfk} = a^\lambda_0 - z_{tfk} M, \tag{121}$$

$$\rho^\lambda_{tfk} = b^\lambda_0, \tag{122}$$

$$\tau^\lambda_{tfk} = \sum_d z_{tfk} s_{kd} \boldsymbol{x}^H_{tf} \boldsymbol{G}^{-1}_{fd} \boldsymbol{x}_{tf}. \tag{123}$$

V. Evaluation

This section presents source separation results obtained with simulated convolutive mixture signals.

A. Experimental Conditions

We synthesized mixture sounds as test data. Fig. 9 shows the locations of microphones and sources. Three sources were convoluted using impulse responses measured with four microphones in a room where the reverberation time RT_{60} was 400 ms. We used music signals (including guitar, bass, vocal, hihat, and piano sounds), speech signals selected from the SiSEC data set [37], and the JNAS phonetically balanced Japanese utterances [38]. Thirty mixture signals were used for evaluation: ten samples of music mixtures, ten samples of speech mixtures, and ten samples of music and speech mixtures. The audio signals sampled at 16 kHz were analyzed with STFT with a 512-pt Hanning window and a 256-pt shift. The average length of mixture signals was 5.22 sec. We also examined the separation

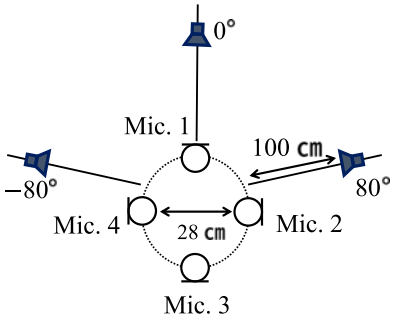


Fig. 9. Locations of microphones and sources.

performance for two times longer mixture signals. In this additional evaluation, to make each source signal, two utterances randomly selected from the JNAS database were concatenated or two times longer signal was extracted from the beginning of each musical piece.

The proposed factor-factor, mixture-factor, factor-mixture, mixture-mixture methods were compared with IVA [24] and MNMF [14] and with the NA-factor and NA-mixture [8] models using only a spatial model. MNMF is similar to the factor-factor model except that the factor-factor model is based on a Bayesian model that can use prior knowledge on steering vectors. When spatial correlation matrices are restricted to rank-1 matrices and the source model is forced to be time-invariant, MNMF reduces to IVA [16]. The number of iterations for IVA and that for MNMF were 100 and 200, respectively. The number of basis spectra for MNMF was 20. The parameters of IVA and MNMF were initialized randomly and estimated by using the majorization-minimization methods.

The steering vectors $\{\boldsymbol{g}^0_{fd}\}^{F,D}_{f=1,d=1}$ for all directions were measured with an angular interval of 5° ($D = 72$) in an anechoic room. This is important to maximize the potential of the proposed models. Once this measurement is done, the proposed models can be used in different echoic conditions without any special treatments (e.g., level-difference compensation). Hyperparameters were determined such that the prior expectation of diagonal elements of the covariance matrix in the likelihood function equals 1 as follows: $a^w_0 = a^h_0 = a^r_0 = b^w_0 = 1, b^r_0 = \frac{1}{D}$, and $\boldsymbol{G}^0_{fd} = \nu(\boldsymbol{g}^0_{fd}(\boldsymbol{g}^0_{fd})^H + \epsilon\boldsymbol{I})$. b^h_0 varied depending on the model: $b^h_0 = 1$ in mixture-mixture, $b^h_0 = \frac{1}{L}$ in factor-mixture, and $b^h_0 = \frac{1}{K \times L}$ in factor-factor and mixture-factor. The other parameters were determined experimentally: $a^\pi_0 = a^\phi_0 = 10$, $a^\psi_0 = 1$, $L = 20$, $\nu_0 = M + 1$, and $\epsilon = 0.01$. All the models were implemented using the C++ language. The parameters of each method were updated or sampled 200 times: 180 samples were abandoned as burn-in and 20 samples were used to estimate sound sources ($I = 20$). The signal-to-distortion ratio (SDR), signal-to-inferences ratio (SIR), and signal-to-artifacts ratio (SAR) [39] were used to evaluate the separation performance.

B. Experimental Results

The experimental results are listed in Tables II–IV, in which the rightmost columns show the SDRs for longer mixture signals. In terms of SDR and SAR, the factor-factor model worked better than almost all of the other methods. In terms of SIR, the mixture-mixture model worked best for all mixtures. Although the separated signals sounded clear because of the clustering nature of mixture modeling, severe distortion was caused. Regarding the spatial model, the factor models achieved higher SDRs than the mixture models. Regarding the source model, there was little difference between the factor and mixture models. Fig. 10 shows the power spectrograms of ground-truth signals and Figs. 11–14 show the those of the sounds separated with the proposed methods. We can see large difference between the spatial factor and mixture models and little difference in the source model.

The SDRs were improved for longer speech mixtures and degraded for longer music mixtures in most methods. The spatial factor models kept the relatively high SDRs for all kinds of longer mixture signals. The spatial mixture models were not suitable to music mixtures because the W-disjoint orthogonality did not hold. MNMF failed to separate longer music mixtures. This indicates that longer signals were not approximated well in the experimental condition because music had larger spectral

TABLE II
EXPERIMENTAL RESULTS FOR MUSIC MIXTURES

	SDR	SIR	SAR	SDR*
IVA [24]	0.3 dB	4.9 dB	5.7 dB	− 2.5 dB
MNMF [14]	1.0 dB	6.2 dB	6.7 dB	− 0.4 dB
Factor-Factor	3.3 dB	7.9 dB	**7.4** dB	3.3 dB
Mixture-Factor	**3.7** dB	9.3 dB	6.7 dB	3.1 dB
Factor-Mixture	0.5 dB	8.3 dB	3.6 dB	− 0.3 dB
Mixture-Mixture	1.2 dB	**9.8** dB	3.3 dB	− 0.4 dB
NA-Mixture [8]	0.5 dB	8.1 dB	3.8 dB	− 0.6 dB

*Separation performance for longer mixture signals.

TABLE III
EXPERIMENTAL RESULTS FOR SPEECH MIXTURES

	SDR	SIR	SAR	SDR*
IVA [24]	3.4 dB	7.5 dB	7.1 dB	2.6 dB
MNMF [14]	4.8 dB	10.0 dB	7.7 dB	5.2 dB
Factor-Factor	**6.1** dB	12.1 dB	**7.8** dB	7.0 dB
Mixture-Factor	5.7 dB	13.4 dB	6.9 dB	7.3 dB
Factor-Mixture	4.9 dB	15.0 dB	5.6 dB	6.2 dB
Mixture-Mixture	5.0 dB	**15.3** dB	5.9 dB	6.7 dB
NA-Mixture [8]	5.1 dB	14.8 dB	5.9 dB	6.2 dB

*Separation performance for longer mixture signals.

TABLE IV
EXPERIMENTAL RESULTS FOR MUSIC AND SPEECH MIXTURES

	SDR	SIR	SAR	SDR*
IVA [24]	0.1 dB	5.3 dB	5.3 dB	0.3 dB
MNMF [14]	1.8 dB	8.6 dB	6.1 dB	2.8 dB
Factor-Factor	**4.9** dB	12.6 dB	**6.5** dB	4.9 dB
Mixture-Factor	4.6 dB	13.4 dB	5.7 dB	4.8 dB
Factor-Mixture	1.7 dB	11.6 dB	4.2 dB	3.3 dB
Mixture-Mixture	3.1 dB	**14.7** dB	4.1 dB	2.7 dB
NA-Mixture [8]	2.5 dB	12.3 dB	4.4 dB	2.3 dB

*Separation performance for longer mixture signals.

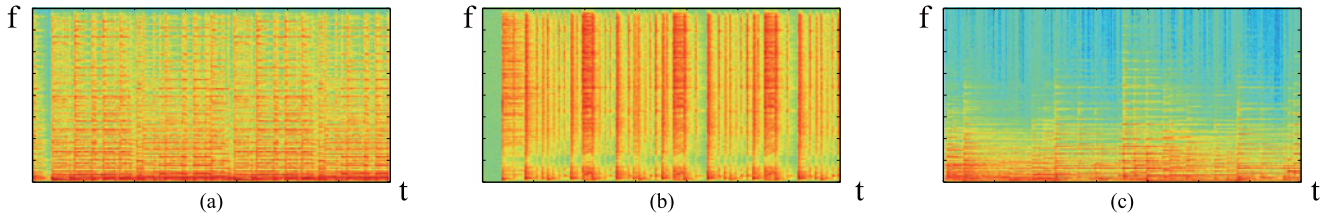


Fig. 10. Spectrograms of musical instrument sounds (ground-truth data). (a) Guitar. (b) Hi-hat. (c) Piano.

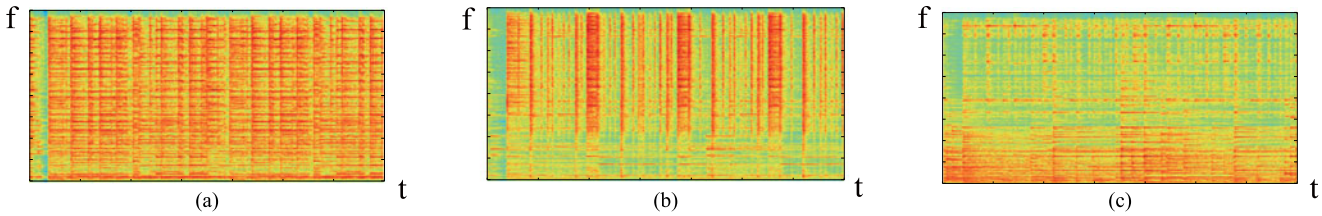


Fig. 11. Spectrograms of musical instrument sounds separated from a music signal by using the factor–factor model. (a) Guitar (SDR: 5.7 dB) (b) Hi-hat (SDR: 5.5 dB) (c) Piano (SDR: 7.7 dB)

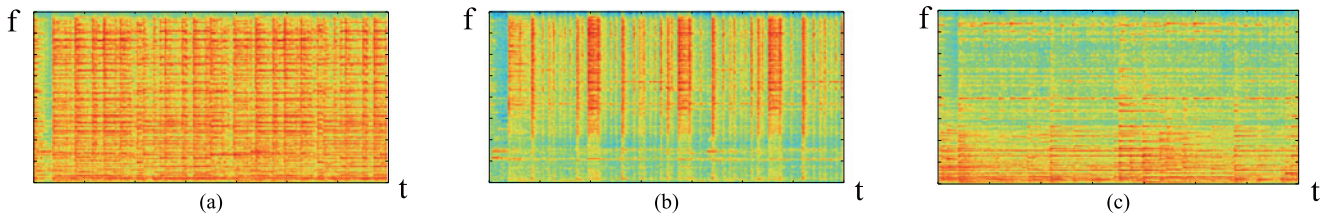


Fig. 12. Spectrograms of musical instrument sounds separated from a music signal by using the mixture–factor model. (a) Guitar (SDR: 6.3 dB). (b) Hi-hat (SDR: 5.0 dB). (c) Piano (SDR: 6.6 dB).

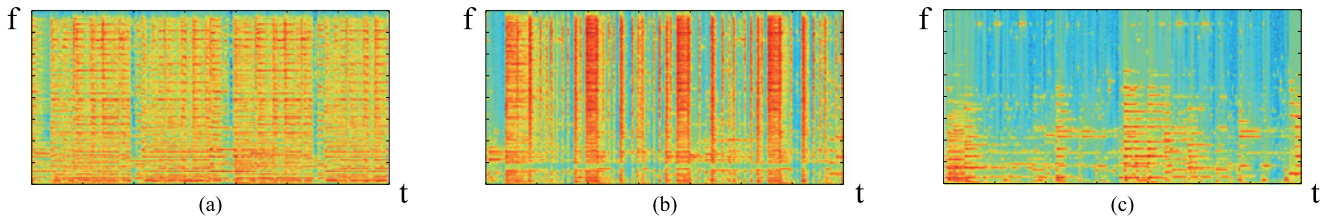


Fig. 13. Spectrograms of musical instrument sounds separated from a music signal by using the factor–mixture model. (a) Guitar (SDR: 1.9 dB). (b) Hi-hat (SDR: 4.1 dB). (c) Piano (SDR: 4.9 dB).

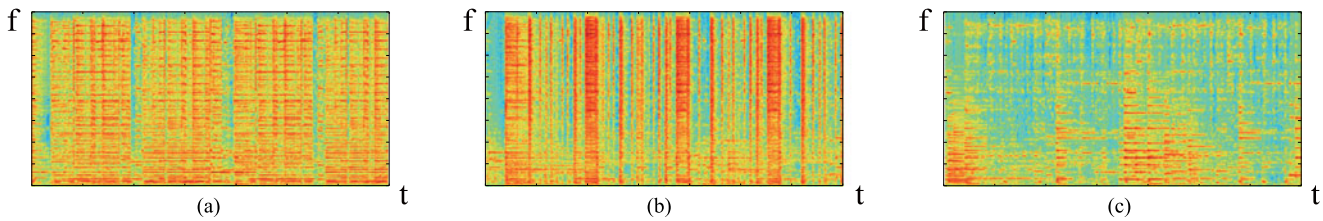


Fig. 14. Spectrograms of musical instrument sounds separated from a music signal by using the mixture–mixture model. (a) Guitar (SDR: 1.8 dB). (b) Hi-hat (SDR: 1.2 dB). (c) Piano (SDR: 4.8 dB).

variations than speech. In this condition, the prior distribution on $\boldsymbol{G}_{fd}$ was helpful, resulting in the difference between MNMF and the spatial factor models.

The performance of the NA-factor model could not be measured because the estimated sound sources could not be associated with ground-truth data (some estimated sources included no sounds). Although how to represent the source model made little difference, the source model plays an important role in combination with the spatial factor model. Since the factor-factor and mixture-factor models are equivalent to the NA-factor model when the number of bases L in a source model becomes infinite, the estimation of L is consider to be essential to the

TABLE V
ELAPSED TIME OF ONE ITERATION FOR A MIXTURE SIGNAL OF 4.6 S

IVA [24]	MNMF [14]		
0.02 s	0.60 s		
Factor-Factor	Mixture-Factor	Factor-Mixture	Mixture-Mixture
57.3 s	64.2 s	3.26 s	1.97 s

TABLE VI
EVALUATION OF MUSIC SIGNALS SEPARATED FROM MUSIC AND SPEECH MIXTURES

Source-Spatial	SDR	SIR	SAR
Factor-Factor	**1.0** dB	8.4 dB	**3.1** dB
Mixture-Factor	0.5 dB	9.5 dB	1.9 dB
Factor-Mixture	− 4.7 dB	4.8 dB	− 0.2 dB
Mixture-Mixture	− 2.1 dB	**10.4** dB	− 0.7 dB

TABLE VII
EVALUATION OF SPEECH SIGNALS SEPARATED FROM MUSIC AND SPEECH MIXTURES

Source-Spatial	SDR	SIR	SAR
Factor-Factor	**8.8** dB	16.7 dB	**9.9** dB
Mixture-Factor	8.7 dB	17.3 dB	9.5 dB
Factor-Mixture	8.1 dB	18.4 dB	8.7 dB
Mixture-Mixture	8.3 dB	**19.0** dB	8.8 dB

mixture-factor and factor-factor models. Indeed, the mixture-factor model did not work when the number of bases L was equal to the number of time frames T.

While there was not much difference of the SDR for speech mixtures between the proposed four models, the factor-factor and mixture-factor models based on the spatial factor model achieved much higher SDRs than the mixture-mixture model for music mixtures. This is because the W-disjoint orthogonality assumption does not hold for music signals. Therefore, the factor-factor and mixture-factor models are suitable to separate music mixtures, and the mixture-mixture model is suitable to separate speech mixtures in terms of SDR.

Which model is best when we want to separate music sounds and speech sounds in music and speech mixtures? To answer this question, we consider the results for music and speech mixtures listed separately for each kind of sounds in Tables VI and VII. For music and speech mixtures, the mixture-mixture model seems to be inferior to the factor-factor model in terms of SDR. Focusing on the speech signals, however, one sees little difference between the mixture-mixture and factor-factor models. Therefore, in terms of SDR, the mixture-mixture model can extract speech sounds almost as well as the factor-factor and mixture-factor models even when the mixture sounds include other kinds of sounds. Further investigations are needed to discuss whether the mixture-mixture model can be used as preprocessing for automatic speech recognition because the mixture-mixture model achieved much lower SARs than the factor-factor and mixture-factor models.

These results show that the proposed models could successfully separate mixture sounds in an unseen environment where the actual steering vectors are significantly different from those measured in an anechoic room. The proposed Bayesian models can flexibly adapt the prior knowledge on the steering vectors to the environment where mixture signals are observed. Open problems to be tackled in the future are to deal with moving sound sources, to achieve real-time source separation, and to estimate the number of sound sources over time.

C. Computational Costs

In our experiments, all the methods sufficiently converged within 180 samples. While the mixture-mixture and factor-mixture models converged within 50 samples, the mixture-factor and factor-factor models needed to draw more than 100 samples before convergence. The proposed models were not sensitive to initialization because the impulse responses measured in an anechoic room were used as prior information. The time elapsed for one iteration of each of the proposed methods with a four-channel input signal (4.6 s) are listed in Table V. The spatial model had larger impact on computational time than the source model, and the mixture models needed less computation time than the factor models except for the mixture-factor model. The computational cost of the mixture-factor model was larger than that of the factor-factor model because of the complicated form of (58). Although the mixture-factor model took 7.8 seconds to sample the latent variable u_{tfkl}, the mixture-factor model could sample other parameters faster than factor-factor. As shown in Tables II–V, we found that the factor-factor and mixture-factor models achieved high SDRs but had large computational costs. On the other hand, we found that the mixture-mixture model achieved slightly lower SDRs than the factor-factor and mixture-factor models but attained remarkably fast computation. Therefore, the mixture-mixture model may be best in terms of SDR and SIR if you want to separate fast; otherwise, you should use the factor-factor or mixture-factor models.

D. Effect of Hyperparameters

We investigated the effect of the hyperparameters on the separation performances of the proposed and conventional models by changing the window length (512, 1024, or 2048 pts), the angular interval between sound sources (20, 50, or 80 degrees), and the number of basis spectra ($L = 5, 10, 20, 50$) if necessary. The SDRs obtained by the compared models with different window lengths under a condition that $L = 20$ and the angular interval was 80 degrees were shown in Fig. 15. In our experiment, the window length of 512 pts was sufficient to attain good SDRs because the SDRs tend to be degraded according to the increase of the window length. The factor-factor and mixture-factor models (spatial factor models) worked well regardless of the sound characteristics.

The SDRs obtained by the compared models with different angular intervals under a condition that the window size was 512 pts and $L = 20$ were shown in Fig. 16. We found that the

Fig. 15. The SDRs obtained by IVA, MNMF, and the factor–factor, mixture–factor, factor–mixture, mixture–mixture, and spatial–mixture models with the window length of 512, 1024, or 2048 pts. (a) Music mixtures. (b) Speech mixtures. (c) Music and speech mixtures.

Fig. 16. The SDRs obtained by IVA, MNMF, and the factor–factor, mixture–factor, factor–mixture, mixture–mixture, and spatial–mixture models for the inter-source interval of 20°, 50°, and 80°. (a) Music mixtures. (b) Speech mixtures. (c) Music and speech mixtures.

SDRs tended to be considerably degraded as the angular interval became small and heavily depended on the characteristics of mixture signals. The spatial factor models kept good performance when the angular interval was 50 or 80 degrees. When the angular interval was 20 degrees, the spatial factor models failed while conventional MNMF relatively worked well. In this condition, the spatial mixture models worked well for speech mixtures while the source mixture models did so for music and speech mixtures.

The SDRs obtained by the compared models with different values of L under a condition that the window size was 512 pts and the angular interval was 80 degrees were shown in Fig. 17. The proposed models were found to work best when $L = 10$ or 20. Too many basis spectra degraded the SDRs.

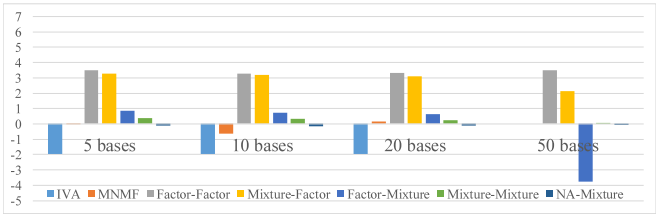
E. Effect of Bayesian Formulation

We investigated the effect of Bayesian formulation in the factor-factor model, which tends to work best in various settings. This model is a Bayesian and direction-aware extension of a variant of multichannel NMF [13]. Note that the direction-aware extension that decomposes the spatial covariance matrix of each source into the weighted sum of direction-dependent covariance matrices was already proposed by [15]. We therefore compared our Bayesian model with its restricted version whose direction-dependent spatial covariance matrix $\boldsymbol{G}_{fd}$ was fixed to $\boldsymbol{G}^0_{fd}$ measured in an anechoic room. In our model, on the other hand, $\boldsymbol{G}^0_{fd}$ is the hyperparameter of the inverse Wishart prior on $\boldsymbol{G}_{fd}$. This enables us to adaptively estimate $\boldsymbol{G}_{fd}$ in posterior inference.

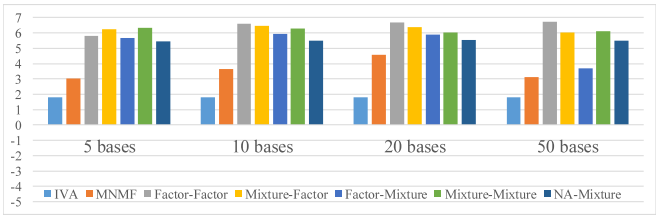
TABLE VIII
EVALUATION OF BAYESIAN INFERENCE OF DIRECTION-DEPENDENT SPATIAL COVARIANCE MATRICES

	Music	Speech	Music & speech
Factor-Factor	3.3 dB	6.7 dB	4.7 dB
Factor-Factor-fixed	3.8 dB	Failed	0.8 dB

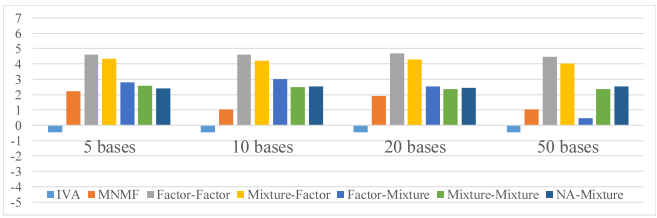
The SDRs obtained by the compared methods under a condition that the window size was 512 pts, the source interval was 80 degrees, and $L = 20$ (described in Section V-A) were shown in Table VIII. When $\boldsymbol{G}_{fd}$ was fixed to $\boldsymbol{G}^0_{fd}$, the speech signals could not be separated. For such signals that violate the low-rankness assumption, spatial information is essential for accurate source separation. On the other hand, the restricted model



(a)



(b)



(c)

Fig. 17. The SDRs obtained by IVA, MNMF, and the factor–factor, mixture–factor, factor–mixture, mixture–mixture, and spatial–mixture models for the numbers of bases, $L = 5, 10, 20, 50$. (a) Music mixtures. (b) Speech mixtures. (c) Music and speech mixtures.

was comparable with the proposed model for music mixtures because the source factor model was effective for the low-rankness of music spectrograms. Although MNMF can be extended to have $\boldsymbol{G}_{fd}$'s as in [15] and each $\boldsymbol{G}_{fd}$ is initialized as $\boldsymbol{G}^0_{fd}$, it has no mechanism to associate $\boldsymbol{G}_{fd}$ with the direction of $\boldsymbol{G}^0_{fd}$ during parameter updating. To prevent $\boldsymbol{G}_{fd}$ and $\boldsymbol{G}_{fd'}$ $(d \neq d')$ from converging to the same value and let $\boldsymbol{G}_{fd}$ and $\boldsymbol{G}_{f'd}$ $(f \neq f')$ indicate the same direction, Bayesian formulation is considered to be effective.

VI. Conclusion

This paper presented multichannel audio source separation methods based on unified source and spatial models. The source model represents the generative process of source power spectrograms and the spatial model represents that of observed multichannel spectrograms. Each model can either be a factor model that represents a generative process as the sum of bases or sources or a mixture model that represents it as the selection of bases or sources. Experimental results showed (1) that the proposed unified models except the factor-mixture model outperformed the conventional spatial model without the source model, (2) that the spatial factor model achieved higher SDRs than the spatial mixture model, (3) that the choice of the source model had little impact on the separation performance, and (4) that the mixture models needed less computational cost than the factor models so that the mixture-mixture model was fastest in the four proposed methods. With further extensions, we plan to estimate the number of the sources and basis spectra in a nonparametric Bayesian manner and to develop an online source separation algorithm.

References

[1] O. Yılmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[2] N. Ito, S. Araki, and T. Nakatani, "Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 3238–3242.

[3] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 516–527, Mar. 2011.

[4] L. Drude, C. Boeddeker, and R. Haeb-Umbach, "Blind speech separation based on complex spherical k-mode clustering," in *IEEE Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 141–145.

[5] D. H. Tran Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 241–244.

[6] I. Jafari, S. Haque, R. Togneri, and S. Nordholm, "On the integration of time-frequency masking speech separation and recognition in underdetermined environments," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, 2012, pp. 1613–1617.

[7] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 382–394, Feb. 2010.

[8] T. Otsuka, K. Ishiguro, H. Sawada, and H. G. Okuno, "Bayesian nonparametrics for microphone array processing," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 2, pp. 493–504, Feb. 2014.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[10] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. San Francisco, CA, USA: Academic, 2010.

[11] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2003, pp. 177–180.

[12] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 550–563, Mar. 2010.

[13] S. Arberet *et al.*, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *Proc. 10th Int. Conf. Inf. Sci. Signal Process. Appl.*, 2010, pp. 1–4.

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

[15] J. Nikunen and T. Virtanen, "Multichannel audio separation by direction of arrival based spatial covariance model and non-negative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 6677–6681.

[16] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1626–1641, Sep. 2016.

[17] K. Itakura, Y. Bando, E. Nakamura, K. Itoyama, K. Yoshii, and T. Kawahara, "Bayesian multichannel nonnegative matrix factorization for audio source separation and localization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 551–555.

[18] M. D. Hoffman, D. M. Blei, and P. R. Cook, "Bayesian nonparametric matrix factorization for recorded music," in *Proc. 27th Int. Conf. Int. Conf. Mach. Learn.*, 2010, pp. 439–446.

[19] 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 *Proc. 24th Eur. Signal Process. Conf.*, 2016, pp. 2280–2284.

[20] J. T. Chien and P. K. Yang, "Bayesian factorization and learning for monaural source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 1, pp. 185–195, Jan. 2016.

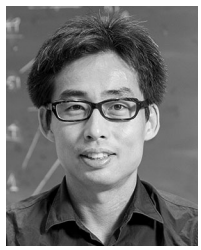
[21] C. Sun, Q. Zhang, J. Wang, and J. Xie, "Noise reduction based on robust principal component analysis," *J. Comput. Inf. Syst.*, vol. 10, no. 10, pp. 4403–4410, 2014.
[22] M. Sun, Y. Li, J. F. Gemmeke, and X. Zhang, "Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank NMF with Kullback-Leibler divergence," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 7, pp. 1233–1242, Jul. 2015.
[23] 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 Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 530–538, Sep. 2004.
[24] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2011, pp. 189–192.
[25] I. Lee, T. Kim, and T. Lee, "Fast fixed-point independent vector analysis algorithms for convolutive blind source separation," *Signal Process.*, vol. 87, no. 8, pp. 1859–1871, 2007.
[26] H. Erdogan, J. R. Hershey, S. Watanabe, M. Mandel, and J. Le Roux, "Improved MVDR beamforming using single-channel mask prediction networks," in *Proc. INTERSPEECH*, 2016, pp. 1981–1985.
[27] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 196–200.
[28] T. Higuchi, T. Yoshioka, and T. Nakatani, "Sparseness-based multichannel nonnegative matrix factorization for blind source separation," in *Proc. IEEE Int. Workshop Acoust. Signal Enhancement*, 2016, pp. 1–5.
[29] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1652–1664, Sep. 2016.
[30] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1830–1840, 2010.
[31] K. Conradsen, A. A. Nielsen, J. Schou, and H. Skriver, "A test statistic in the complex Wishart distribution and its application to change detection in polarimetric SAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 1, pp. 4–19, Jan. 2003.
[32] G. Casella and E. I. George, "Explaining the Gibbs sampler," *Amer. Statist.*, vol. 46, no. 3, pp. 167–174, 1992.
[33] 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 *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 51–55.
[34] B. Jørgensen, *Statistical Properties of the Generalized Inverse Gaussian Distribution*, vol. 9. New York, NY, USA: Springer, 2012.
[35] F. Fazayeli and A. Banerjee, "The matrix generalized inverse Gaussian distribution: Properties and applications," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2016, pp. 648–664.
[36] J. S. Dagpunar, *Simulation and Monte Carlo: With Applications in Finance and MCMC*. New York, NY, USA: Wiley, 2007.
[37] S. Araki *et al.*, "The 2011 signal separation evaluation campaign (SiSEC2011): Audio source separation," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, 2012, pp. 414–422.
[38] K. Itou *et al.*, "The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus," in *Proc. Int. Conf. Spoken Lang. Process.*, 1998, pp. 3261–3264.
[39] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

Kousuke Itakura received the B.E. and M.S. degrees from Kyoto University, Kyoto, Japan, in 2015 and 2017, respectively. He is currently working in an electronics manufacturer in Japan. His research interests include audio signal processing and statistical machine learning.

Yoshiaki Bando (M'17) received the M.S. degree in informatics in 2015 from Kyoto University, Kyoto, Japan, where he is currently working toward the Ph.D. degree. His research interests include microphone array signal processing, rescue robotics, and machine learning. He was a recipient of the Advanced Robotics Best Paper Award in 2016, the Most Innovative Paper Award at the 2015 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR), the Best Student Paper Award at the IEEE SSRR 2014, the IEEE Robotics and Automation Society Japan Chapter Young Award at the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems. He is a member of the Robotics Society of Japan, the Information Processing Society of Japan, and the Japanese Society for Artificial Intelligence.

Eita Nakamura received the Ph.D. degree in physics from the University of Tokyo, Tokyo, Japan, in 2012. After having been a Postdoctoral Researcher at the National Institute of Informatics, Meiji University, and Kyoto University, Kyoto, Japan, he is currently a Japan Society for the Promotion of Science Research Fellow with Kyoto University. His research interests include music modeling and analysis, music information processing, and statistical machine learning.

Katsutoshi Itoyama (M'11) received the M.S. and Ph.D. degrees in informatics from Kyoto University, Kyoto, Japan, in 2008 and 2011, respectively. He is currently an Assistant Professor at the Graduate School of Informatics, Kyoto University. His research interests include musical sound source separation, music listening interfaces, and music information retrieval. He was a recipient of the Information Processing Society of Japan (IPSJ) Digital Courier Funai Young Researcher Encouragement Award. He is a member of the IPSJ and the Acoustical Society of Japan.



Kazuyoshi Yoshii (M'08) received the M.S. and Ph.D. degrees in informatics from Kyoto University, Kyoto, Japan, in 2005 and 2008, respectively. He is currently a Senior Lecturer at the Graduate School of Informatics, Kyoto University, and concurrently the Leader of the Sound Scene Understanding Team, RIKEN Center for Advanced Intelligence Project, Tokyo, Japan. His research interests include music analysis, audio signal processing, and machine learning. He is a member of the Information Processing Society of Japan and the Institute of Electronics, Information and Communication Engineers, Japan.

Tatsuya Kawahara (F'17) received the B.E., M.E., and Ph.D. degrees in information science from Kyoto University, Kyoto, Japan, in 1987, 1989, and 1995, respectively. He is currently a Professor at the Graduate School of Informatics, Kyoto University. He has also been an Invited Researcher at ATR and NICT. He has authored or coauthored more than 300 technical papers on speech recognition, spoken language processing, and spoken dialog systems. He has been conducting several speech-related projects in Japan including a speech recognition software Julius and an automatic transcription system for the Japanese Parliament (Diet).

Prof. Kawahara is an Editorial Board Member of the Elsevier's journal *Computer Speech and Language*, the *APSIPA Transactions on Signal and Information Processing*, and the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING. He is a VP—Publications of the Asia-Pacific Signal and Information Processing Association and a Board Member of the International Speech Communication Association.