Unified approach for underdetermined BSS, VAD, dereverberation and DOA estimation with multichannel factorial HMM

Takuya Higuchi* and Hirokazu Kameoka*†
*Graduate School of Information Science and Technology,
The University of Tokyo, Hongo 7-3-1, Bunkyo, Tokyo, 113-8656, Japan
† NTT Communication Science Laboratories, NTT Corporation,
Morinosato Wakamiya 3-1, Atsugi, Kanagawa, 243-0198, Japan

***Abstract*—This paper proposes a novel method for simultaneously solving the problems of underdetermined blind source separation (BSS), source activity detection, dereverberation and direction-of-arrival (DOA) estimation by introducing an extension of the "multichannel factorial hidden Markov model (MFHMM)." The MFHMM is an extension of the multichannel non-negative matrix factorization (NMF) model, in which the basis spectra are allowed to vary over time according to the transitions of the hidden states. This model has allowed us to perform source separation, source activity detection and dereverberation in a unified manner. In our previous model, the spatial covariance of each source has been treated as a model parameter. This has led the entire generative model to have an unnecessarily high degree of freedom, and thus the parameter inference has been prone to getting trapped into undesired local optima. To reasonably restrict the solution space of the spatial covariance matrix of each source, we propose to describe it as a weighted sum of the fixed spatial covariance matrix corresponding to the discrete set of DOAs. Through the parameter inference, the proposed model allows us to simultaneously solve the problems of underdetermined BSS, source activity detection, dereverberation and DOA estimation. Experimental results revealed that the proposed method was superior to a previous method in terms of the signal-to-distortion ratios of separated signals.**

Index Terms: source separation, dereverberation, hidden Markov model, non-negative matrix factorization, DOA

I. Introduction

Blind source separation (BSS) refers to a technique for separating out individual source signals from microphone array inputs when the transfer characteristics between the sources and microphones are unknown. To solve BSS problems, it is generally necessary to make some assumptions about the sources, and formulate an appropriate optimization problem based on criteria designed according to those assumptions. For example, if the observed signals outnumber the sources, we can employ independent component analysis (ICA) [1] by assuming that the sources are statistically independent of each other. However, in an underdetermined case, the independence assumption is too weak to allow us to determine a unique solution and so directly applying ICA will not work well. For monaural source separation, one successful approach involves applying non-negative matrix factorization (NMF) to the magnitude spectrogram of a mixture signal, interpreted as a non-negative matrix [2][3]. Up to now, several attempts have been made to extend this approach to a multichannel case in order to allow for the use of spatial information as an additional clue for separation, which have opened a door to a new promising approach for underdetermined BSS [4][5]. Moreover, we previously proposed the multichannel factorial hidden Markov model [6], where non-stationary spectra of a source are modeled by using a HMM. Furthermore, we extended the MFHMM for dereverberation [7], by approximating the multichannel observed signal recorded in a reverberant condition as a form of a convolution of the frequency array response and the source signal in the time-frequency domain. Thus, we modeled the impulse response out of the frame of STFT by introducing a time sequence of frequency response arrays in the time-frequency domain. However, in the model we mentioned above, we do not make any assumption for a spatial correlation matrix and the degree of freedom of the spatial correlation matrix is very high. This sometimes makes the spatial correlation matrix trapped into undesirable local optima. Generally, it is known that the spatial correlation matrix of the direct sound of a point source has a certain structure described by the source ' s direction. One way to model the spatial correlation matrix would be to use a mixture model of matrices parametrized by all possible direction of arrivals (DOAs), as in [8][9]. Another way would be to use a linear sum of those matrices, as in [10]. For reverberant conditions, the latter model would be more relevant than the former, since the spatial correlation matrix could be contaminated by the components of early reflections. In this paper, we incorporate this model of a spatial correlation matrix into our previous model, and approximate the impulse response both within and out of the frame of the STFT in the time-frequency domain. Through the parameter inference of our new generative model, we can simultaneously perform source separation, source activity detection, dereverberation and DOA estimation based on a unified maximum likelihood criterion.

II. Multichannel factorial hidden Markov model [7]

A. Mixing model

First we consider a situation where I source signals are recorded by M microphones. Here, let $y_m(t) \in \mathbb{R}$ be the observed signal at the m-th microphone, and $s_i(t) \in \mathbb{R}$ be the signal of the i-th source. The observed signal can be written in the time domain:

$$\boldsymbol{y}(t) = \sum_{i=1}^{I} \sum_{\tau=0}^{\infty} \boldsymbol{a}_i(\tau) s_i(t-\tau), \tag{1}$$

where $\boldsymbol{y}(t) = (y_1(t), \dots, y_M(t))^{\mathsf{T}} \in \mathbb{R}^M$ and $\boldsymbol{a}_i(t) = (a_{1,1}(t), \dots, a_{1,M}(t))^{\mathsf{T}} \in \mathbb{R}^M$. $a_{i,m}(t)$ denotes the impulse response between source i and microphone m. In a reverberant condition, the length of the impulse responses are relatively long and so an instantaneous mixture approximation is not always true. Therefore we approximately express the observed signals as a form of a convolution of the frequency array response and the source signal in the time-frequency domain.

$$\boldsymbol{y}(\omega_k, t_l) \approx \sum_{i=1}^{I} \sum_{\tau=0}^{T} \boldsymbol{a}_i(\omega_k, \tau) s_i(\omega_k, t_l - \tau). \tag{2}$$

Here, let $y_m(\omega_k, t_l) \in \mathbb{C}$ be the short-time Fourier transform (STFT) component observed at the m-th microphone, and $s_i(\omega_k, t_l) \in \mathbb{C}$ be the STFT component of the i-th source. $1 \le k \le K$ and $1 \le l \le L$ are the frequency and time indices, respectively. $\boldsymbol{a}_i(\omega_k, \tau)$ denotes the frequency array response for source i at frequency ω_k and time τ. $0 \le \tau \le T$ is the time index of the frequency array response in the time-frequency domain. Note that $\boldsymbol{a}_i(\omega_k, 1 : T)$ denote the frequency array responses which correspond to the impulse responses out of the frames of the STFT. This approximation is useful for dereverberation [11] and the validity of the approximation is experimentally shown. For convenience of notation, we hereafter use subscripts k and l to indicate ω_k and t_l respectively.

B. Generative process of observed signals

Here we describe the generative process of an observed signal based on Eq. (2). We assume that the source signal $s_{i,k,l}$ follows a complex normal distribution with mean 0 and covariance $\sigma^2_{i,k,l}$.

$$s_{i,k,l}|\sigma_{i,k,l} \sim \mathcal{N}_{\mathbb{C}}(s_{i,k,l}; 0, \sigma^2_{i,k,l}), \tag{3}$$

Then, from Eq. (2), $\boldsymbol{y}_{k,l}$ is also normally distributed such that

$$\begin{aligned} &\boldsymbol{y}_{k,l}|\boldsymbol{a}_{1:I,k,0:T}, s_{1:I,k,l-T:l} \\ &\quad \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{y}_{k,l}; 0, \sum_{i,\tau} \boldsymbol{C}_{i,k,\tau}\sigma^2_{i,k,l-\tau}), \end{aligned} \tag{4}$$

conditioned on $\boldsymbol{a}_{1:I,k,0:T}$ and $s_{1:I,k,l-T:l}$ where $\boldsymbol{C}_{i,k,\tau} = \boldsymbol{a}_{i,k,\tau}\boldsymbol{a}^{\mathsf{H}}_{i,k,\tau}$ and $\mathcal{N}_{\mathbb{C}}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) \propto \exp(-(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{H}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))$. $\boldsymbol{C}_{i,k,\tau}$ is often called a spatial correlation matrix.

C. Generative modeling of source signals using HMMs

We now describe the generative process of the source signal $s_{i,k,l}$ using a HMM. First, we assume each signal has a specific spectrum and utilize NMF for $\sigma^2_{i,k,l}$, which is an expected value of power of $s_{i,k,l}$. $\sigma^2_{i,k,l}$ is factorized as

$$\sigma^2_{i,k,l} = w_{i,k}h_{i,l}, \tag{5}$$

where $w_{i,k}$ and $h_{i,l}$ are non-negative variables. $w_{i,1:K}$ represents a specific structure of spectrum that the source signal has. $h_{i,l}$ means the activation of the spectrum in time l. The generative model of $s_{i,k,l}$ is rewritten as

$$\boldsymbol{s}_{i,k,l}|w_{i,k}, h_{i,l} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{s}_{i,k,l}; 0, w_{i,k}h_{i,l}), \tag{6}$$

conditioned on $w_{i,k}$ and $h_{i,l}$. Under the condition of Eq. (6), the generative model of the single-channel observed signal is equivalent to that proposed in [12].

In many cases, a sound signal consists of several spectra and the source's spectrum varies according to the source's state. Now we introduce latent variable $z_{i,l} \in \{1, \ldots, Q\}$ to denote a state of i-th source in time l. The state sequence $z_{i,1}, \ldots, z_{i,L}$ follows a Markov chain:

$$z_{i,l}|z_{i,l-1} \sim \text{Categorical}(z_{i,l}; \boldsymbol{\rho}_{z_{i,l-1}}), \tag{7}$$

where $\text{Categorical}(x; \boldsymbol{y}) = y_x$, $\boldsymbol{\rho}_q = (\rho_{q,1}, \ldots, \rho_{q,Q})$ denotes the transition probability of state q to each state $1, \ldots, Q$, and $\boldsymbol{\rho} = (\rho_{q,q'})_{Q\times Q}$ denotes the transition matrix. Then, we assume $h_{i,l}$ follows a gamma distribution which has different parameters according to $z_{i,l}$ [6],

$$h_{i,l}|z_{i,l} \sim \text{Gamma}(h_{i,l}; \alpha_{z_{i,l}}\beta_{z_{i,l}}), \tag{8}$$

where $\alpha_{1:Q}$ and $\beta_{1:Q}$ are the shape and scale parameters of a gamma distribution, and $\text{Gamma}(x; \alpha, \beta) = \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}}$. As we want $h_{i,l}$ to take a small value when $z_{i,l}$ is the "inactive" (i.e., silent) state, we set the hyperparameters of the gamma distribution of that state so that it becomes a sparsity-inducing distribution. As regards the gamma distributions of the remaining states, we consider setting the hyperparameters so that they become uniform distributions. The power spectrum of the i-th source at time l is also assumed to be determined according to $z_{i,l}$. Thus, the generative model of $s_{i,k,l}$ is eventually written as

$$\begin{aligned} &s_{i,k,l}|w_{i,k,1:Q}, h_{i,l}, z_{i,l} \\ &\quad \sim \mathcal{N}_{\mathbb{C}}(s_{i,k,l}; 0, w_{i,k,z_{i,l}}h_{i,l}). \end{aligned} \tag{9}$$

With a similar motivation, we have previously proposed modeling single-channel mixture signals using a factorial HMM, where the basis spectrum of each latent source signal is allowed to vary over time as a result of state transitions [13]. Unlike our previous model [13], the amplitudes of the basis spectra also depend on the hidden states in the present model, making it possible to perform audio event detection and source separation simultaneously through parameter inference. Our overall generative model is given by Eqs. (7), Eqs. (8) and

$$\begin{aligned} &\boldsymbol{y}_{k,l}|\boldsymbol{a}_{1:I,k,1:T}, w_{1:I,k,1:Q}, h_{1:I,l-T:l}, z_{1:I,l-T:l} \\ &\quad \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{y}_{k,l}; 0, \sum_{i,\tau} \boldsymbol{C}_{i,k,\tau}w_{i,k,z_{i,l-\tau}}h_{i,l-\tau}), \end{aligned} \tag{10}$$

conditioned on $\boldsymbol{a}_{1:I,k,0:T}$,$w_{1:I,k,1:Q}$, $h_{1:I,l-T:l}$ and $z_{1:I,l-T:l}$.

III. Modeling a spatial correlation matrix based on DOA

If we assume that a source is far from the microphones, the frequency array response has a certain structure in the time-frequency domain depending on Direction of Arrival (DOA). Therefore, we can express a spatial correlation matrix by using the DOAs [8][9][10]. Specifically, with $M = 2$ microphones, the spatial correlation matrix for a source at direction θ such that $0 \leq \theta \leq \pi$ is defined as a function of ω depending on θ

$$\boldsymbol{J}(\theta, \omega) = \begin{bmatrix} 1 \\ e^{\jmath\omega B\cos\theta/C} \end{bmatrix} \begin{bmatrix} 1 & e^{\jmath\omega B\cos\theta/C} \end{bmatrix}^* \tag{11}$$

where $\jmath$ is the imaginary unit, B [m] is the distance between the two microphones, and C [m/s] is the speed of sound. If the DOA θ_i of source i is known, the spatial correlation matrix for the direct wave should be equal to $\boldsymbol{J}(\theta_i, \omega_k)$. However, $\boldsymbol{C}_{i,k,0}$ would not be equal to $\boldsymbol{J}(\theta_i, \omega_k)$ because of reverberant components in the frame of the STFT. Therefore, we introduce a discrete set of O possible directions $\vartheta_1, \ldots, \vartheta_O$ and weight variables $d_{i,1} \ldots d_{i,O}$, then we express a spatial correlation matrix using $\boldsymbol{J}(\theta_i, \omega_k)$ and $d_{i,o}$ as;

$$\boldsymbol{C}_{i,k,0} = \sum_{o} d_{i,o}\boldsymbol{J}(\vartheta_o, \omega_k). \tag{12}$$

$d_{i,1} \ldots d_{i,O}$ are non-negative and satisfy $\sum_o d_{i,o} = 1$ [10]. We expect that an relatively large value in $d_{i,1} \ldots d_{i,O}$ would indicate the DOA of the direct wave of the i-th source.

IV. Algorithm for parameter estimation

A. Objective function

In this section, we describe a parameter estimation algorithm for our generative model based on an auxiliary function method. The random variables of interest in our model are $\boldsymbol{W} = w_{1:I,1:K}$, $\boldsymbol{H} = h_{1:I,1:L}$,$\boldsymbol{C} = \boldsymbol{C}_{1:I,1:K,1:T}$ $\boldsymbol{D} = d_{1:I,1:O}$ and $\boldsymbol{Z} = z_{1:I,1:L}$. We denote the entire set of the above parameters except $\boldsymbol{Z}$ as Θ. In the following, $\boldsymbol{\rho}$, α and β are constants that is determined experimentally.

The objective function $L(\Theta)$ is defined as $L(\Theta) = \log\sum_{\boldsymbol{Z}} p(\Theta, \boldsymbol{Z}|\boldsymbol{Y})$, where $\boldsymbol{Y} = \boldsymbol{y}_{1:K,1:L}$ is a set consisting of the time-frequency components of observed multichannel signals. Our goal is to obtain $\hat{\Theta}$ such that

$$\hat{\Theta} = \underset{\Theta}{\text{argmax}} \log \sum_{\boldsymbol{Z}} p(\Theta, \boldsymbol{Z}|\boldsymbol{Y}). \tag{13}$$

By using the conditional distributions defined in Sec. II, we can rewrite $\hat{\Theta}$ as

$$
\begin{aligned}
\hat{\Theta} &= \underset{\Theta}{\operatorname{argmax}}\big(\log \sum_{\boldsymbol{Z}} p(\boldsymbol{Y}|\boldsymbol{W},\boldsymbol{H},\boldsymbol{C},\boldsymbol{D},\boldsymbol{Z}) \\
&\quad + \log \sum_{\boldsymbol{Z}} p(\boldsymbol{H}|\boldsymbol{Z}) + \log \sum_{\boldsymbol{Z}} p(\boldsymbol{Z})\big) \\
&= \underset{\Theta}{\operatorname{argmin}}\big(\sum_{k,l}(-\frac{1}{2}\log \sum_{\boldsymbol{Z}} |\hat{\boldsymbol{X}}_{k,l}| \\
&\quad - \frac{1}{2}\log \sum_{\boldsymbol{Z}} \exp \boldsymbol{y}_{k,l}{}^{\mathsf{H}} \hat{\boldsymbol{X}}_{k,l}^{-1} \boldsymbol{y}_{k,l}) \\
&\quad + \log \sum_{\boldsymbol{Z}} p(\boldsymbol{H}|\boldsymbol{Z}) + \log \sum_{\boldsymbol{Z}} p(\boldsymbol{Z})\big), \qquad (14)
\end{aligned}
$$

where $\hat{\boldsymbol{X}}_{k,l} = \sum_{i,\tau} \boldsymbol{C}_{i,k,\tau} w_{i,k,z_{i,l-\tau}} h_{i,l-\tau,z_{i,l-\tau}}$.

B. Optimization algorithm based on an auxiliary function method

The optimization problem of maximizing $L(\Theta)$ with respect to Θ is difficult to solve analytically. However, we can invoke an auxiliary function approach to derive an iterative algorithm that searches for the estimate of Θ, as with [5]. To apply an auxiliary function approach to the current optimization problem, the first step is to construct an auxiliary function $L^+(\Theta, \Lambda)$ satisfying $L(\Theta) = \max_\Lambda L^+(\Theta, \Lambda)$. We refer to Λ as an auxiliary variable. It can then be shown that $L(\Theta)$ is non-decreasing under the updates $\Theta \leftarrow \operatorname{argmax}_\Theta L^+(\Theta, \Lambda)$ and $\Lambda \leftarrow \operatorname{argmax}_\Lambda L^+(\Theta, \Lambda)$. The proof of this shall be omitted owing to space limitations. Thus, $L^+(\Theta, \Lambda)$ should be designed as a function that can be maximized analytically with respect to Θ and Λ. Such a function can be constructed as follows.

$$
\begin{aligned}
& L(\boldsymbol{\Theta}) \\
\geq & L^+(\boldsymbol{\Theta}, \Lambda) \\
= & -\frac{1}{2}\sum_{k,l}\Bigg(\sum_{i,q,o} \lambda_{q,i,l}\Bigg(\frac{\operatorname{tr}(\boldsymbol{y}_{k,l}\boldsymbol{y}_{k,l}{}^{\mathsf{H}}\boldsymbol{R}_{i,k,l,0,q}\boldsymbol{J}_{i,k,o}^{-1}\boldsymbol{R}_{i,k,l,0,q})}{d_{i,o} w_{i,k,q_{i,l}} h_{i,l}} \\
& + \operatorname{tr}(\boldsymbol{U}_{k,l}^{-1}\boldsymbol{J}_{i,o}) d_{i,o} w_{i,k,q_{i,l}} h_{i,l}\Bigg) \\
& + \sum_{i,q,\tau\neq 0} \lambda_{q,i,l}\Bigg(\frac{\operatorname{tr}(\boldsymbol{y}_{k,l}\boldsymbol{y}_{k,l}{}^{\mathsf{H}}\boldsymbol{R}_{i,k,l,\tau,q}\boldsymbol{C}_{i,k,\tau}^{-1}\boldsymbol{R}_{i,k,l,\tau,q})}{w_{i,k,q_{i,l-\tau}} h_{i,l-\tau}} \\
& + \operatorname{tr}(\boldsymbol{U}_{k,l}^{-1}\boldsymbol{C}_{i,k,\tau}) w_{i,k,q_{i,l-\tau}} h_{i,l-\tau}\Bigg) + \log|\boldsymbol{U}_{k,l}| - M\Bigg) \\
& + \sum_{i,l,q} \lambda_{q,i,l}\Bigg((\alpha_{q_{i,l}} - 1)\log h_{i,l} - h_{i,l}/\beta_{q_{i,l}} - \alpha_{q_{i,l}}\log \beta_{q_{i,l}}\Bigg) \\
& + \sum_{q} \lambda_{q,i,l} \log p(\boldsymbol{Z}), \qquad (15)
\end{aligned}
$$

where $\boldsymbol{R}_{i,k,l,\tau,q}$, $\boldsymbol{U}_{k,l}$ and $\lambda_{q,i,l}$ are auxiliary variables. $\boldsymbol{R}_{i,k,l,\tau}$ and $\boldsymbol{U}_{k,l}$ satisfy Hermitian positive definiteness and $\sum_{i,\tau} \boldsymbol{R}_{i,k,l,\tau,q} = \boldsymbol{I}$. $\lambda_{q,i,l}$ satisfies $\sum_q \lambda_{q,i,l} = 1$. We denote the set of the auxiliary variables as Λ. $\operatorname{tr}(\cdot)$ is the trace of a matrix. The equality $L(\boldsymbol{\Theta}) = L^+(\boldsymbol{\Theta}, \Lambda)$ is satisfied when

$$
\begin{aligned}
\boldsymbol{R}_{i,k,l,\tau,q} &= \boldsymbol{C}_{i,k,\tau} w_{i,k,q_{i,l-\tau}} h_{i,l-\tau} \hat{\boldsymbol{X}}_{k,l}^{-1}, \qquad (16) \\
\boldsymbol{U}_{k,l} &= \hat{\boldsymbol{X}}_{k,l}, \qquad (17) \\
\lambda_{q,i,l} &= p(q_{i,l}|\Theta). \qquad (18)
\end{aligned}
$$

TABLE I
THE EXPERIMENTAL CONDITION.

	Q	O	T	α	β
RT=380 ms	5	30	3	1	$\beta_1 = 10^{-1}$, $\beta_{2:10} = 10^{10}$

TABLE II
THE AVERAGE INPUT SDRS/SIRS [dB] AND THESE OBTAINED BY OUR PREVIOUS AND PRESENT METHODS.

RT=380ms	SDR	SIR
Proposed	-4.17	5.90
Method in [7]	-6.49	1.94
Input	-40.33	-4.22

Therefore, we can indirectly maximize L by repeating the following two steps.

1) Maximizing L^+ with respect to $\boldsymbol{R}$, $\boldsymbol{U}$ and λ.
2) Maximizing L^+ with respect to $\boldsymbol{W}$, $\boldsymbol{H}$, $\boldsymbol{C}$ and $\boldsymbol{D}$.

Step 1 consists in updating $\boldsymbol{R}$ and $\boldsymbol{U}$ using Eqs. (16) and (17). With respect to λ, we can apply the Forward-Backward algorithm. The update rule of λ is given by

$$
\lambda_{q,i,l} = F_{q,i,l} B_{q,i,l} / \sum_q F_{q,i,l} B_{q,i,l}, \qquad (19)
$$

$$
F_{q,i,l} = p(\Theta|q_{i,l}) \sum_{q_{i,l-1}} F_{q,i,l-1} \rho_{q_{i,l-1},q_{i,l}}, \qquad (20)
$$

$$
B_{q,i,l} = \sum_{q_{i,l+1}} B_{q,i,l+1} p(\Theta|q_{i,l+1}) \rho_{q_{i,l},q_{i,l+1}}, \qquad (21)
$$

where

$$
\begin{aligned}
& p(\Theta|q_{i,l}) \\
& \propto \exp\Bigg(-\frac{1}{2}\sum_{k,l}\Bigg(\sum_{i,o}\Bigg(\frac{\operatorname{tr}(\boldsymbol{y}_{k,l}\boldsymbol{y}_{k,l}{}^{\mathsf{H}}\boldsymbol{R}_{i,k,l,0,q}\boldsymbol{J}_{i,k,o}^{-1}\boldsymbol{R}_{i,k,l,0,q})}{d_{i,o} w_{i,k,q_{i,l}} h_{i,l}} \\
& + \operatorname{tr}(\boldsymbol{U}_{k,l}^{-1}\boldsymbol{J}_{i,k,o}) d_{i,o} w_{i,k,q_{i,l}} h_{i,l}\Bigg) \\
& + \sum_{i,\tau\neq 0}\Bigg(\frac{\operatorname{tr}(\boldsymbol{y}_{k,l+\tau}\boldsymbol{y}_{k,l+\tau}{}^{\mathsf{H}}\boldsymbol{R}_{i,k,l+\tau,\tau,q}\boldsymbol{C}_{i,k,\tau}^{-1}\boldsymbol{R}_{i,k,l+\tau,\tau,q})}{w_{i,k,q_{i,l}} h_{i,l}} \\
& + \operatorname{tr}(\boldsymbol{U}_{k,l+\tau}^{-1}\boldsymbol{C}_{i,k,\tau}) w_{i,k,q_{i,l}} h_{i,l}\Bigg)\Bigg) \\
& + \sum_{i,l}\Bigg((\alpha_{q_{i,l}} - 1)\log h_{i,l} - h_{i,l}/\beta_{q_{i,l}} - \alpha_{q_{i,l}}\log \beta_{q_{i,l}}\Bigg)\Bigg). \\
& \qquad (22)
\end{aligned}
$$

In step 2, we can obtain update rules of $\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{C}$ and $\boldsymbol{D}$ by setting the partial derivative of L^+ with respect to each of the parameters at zero. Specifically, the partial derivatives of L^+ with respect to $\boldsymbol{D}$ is given by

$$
\begin{aligned}
\frac{\partial L^+}{\partial d_{i,o}} = \sum_{k,l,q} \lambda_{q,i,l}\Bigg(&\frac{\operatorname{tr}(\boldsymbol{y}_{k,l}\boldsymbol{y}_{k,l}{}^{\mathsf{H}}\boldsymbol{R}_{i,k,l,0,q}\boldsymbol{J}_{i,k,o}^{-1}\boldsymbol{R}_{i,k,l,0,q})}{d_{i,o}^2 w_{i,k,q_{i,l}} h_{i,l}} \\
& - \operatorname{tr}(\boldsymbol{U}_{k,l}^{-1}\boldsymbol{J}_{i,k,o}) w_{i,k,q_{i,l}} h_{i,l}\Bigg). \qquad (23)
\end{aligned}
$$

According to the conditions for extremal values, update rules are written as

$$
d_{i,o} \leftarrow \sqrt{\frac{\sum_{k,l,q} \lambda_{q,i,l} \frac{\operatorname{tr}(\boldsymbol{y}_{k,l}\boldsymbol{y}_{k,l}{}^{\mathsf{H}}\boldsymbol{R}_{i,k,l,0}\boldsymbol{J}_{i,k,o}^{-1}\boldsymbol{R}_{i,k,l,0})}{w_{i,k,q_{i,l}} h_{i,l}}}{\sum_{k,l,q} \lambda_{q,i,l} \operatorname{tr}(\boldsymbol{U}_{k,l}^{-1}\boldsymbol{J}_{i,k,o}) w_{i,k,q_{i,l}} h_{i,l}}}. \qquad (24)
$$

Note that to estimate $\boldsymbol{W}$, $\boldsymbol{H}$, $\boldsymbol{C}$, $\boldsymbol{D}$ and $\boldsymbol{\lambda} = \lambda_{1:Q,1:I,1:L}$ means to solve the problems of source separation, source activity detection, dereverberation and DOA estimation simultaneously.

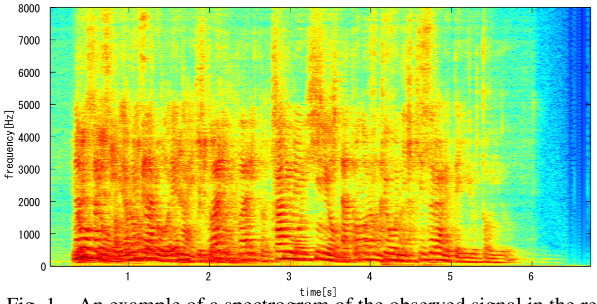

Fig. 1. An example of a spectrogram of the observed signal in the reverberant condition (the reverberation time was 380 ms).

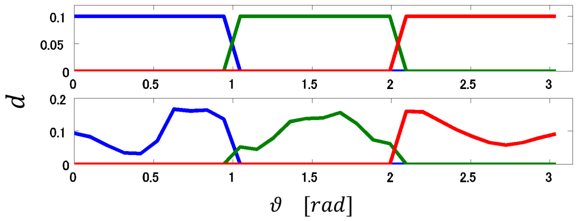

Fig. 2. Initial condition of d (top) and an example of the results of d with the echoic signals (bottom).

V. Related works

Our proposed model is equal to the model proposed in [7] except Eqs. (12). The expression of Eqs. (12) was proposed in [10] for the spatial matrix of multichannel NMF with the mean-square-error criterion. We incorporate this model of the spatial correlation matrix into our previous generative model , which means that we model reverberation in the frame of the STFT by using the DOAs. The estimation of $\boldsymbol{D}$ allows us to estimate the DOAs, and we derive update rules of $\boldsymbol{D}$ for optimizing our objective function. Moreover, we marginalize $\boldsymbol{Z}$ in Sec. IV, while the algorithm proposed in [7] allows us only to obtain the optimal sequence of $\boldsymbol{Z}$ by applying the Viterbi algorithm.

VI. Experimental evaluation

We evaluated the performance of the present method in terms of the abilities of source separation. We used 10 mixed stereo signals (therefore the number of the microphones M is 2) as the experimental data, each of which we obtained by mixing three speech signals (speech of a male and two females) from the ATR database [14] and was convolved with the measured room impulse response from the RWCP database [15] (in which the distance between the microphones was 11.48 cm and the reverberation time was 380 ms). Thus, the three signals were artificially located 30°, 90° and 130° from the microphones respectively. Fig. 1 shows a spectrogram of the observed signal (the reverberation time was 380 ms). The sampling rate was 16 kHz. To compute the STFT components of the observed signal, the STFT frame length was set at 64 ms and a Hamming window was used with an overlap length of 48 ms. We set the parameters as Table I. We expected that $q = 1$ is an inactive state and $q = 2 : 5$ are active states. $\boldsymbol{D}$ was initially set as Fig. 2 (top). For $\tau = 1, \dots, T$, The diagonal elements of $\boldsymbol{C}_{i,k,1:T}$ were set to $10^{-1}/\sqrt{M}$, and the off-diagonal elements were also set to zero initially. $\boldsymbol{W}$ was initially randomized. $\boldsymbol{H}$ was set as 1 initially. $\rho_{q,q'}$ was set as $1/Q$ The parameter estimation algorithm was run for 30 iterations. In order to avoid an undesirable local optima, we first set T as 0 and gradually increased T up to 3 with the iteration. We chose the method proposed in [7] as a comparison. The separated signal $\hat{\boldsymbol{y}}_{i,k,l}$ was obtained by Wiener filtering

$$\hat{\boldsymbol{y}}_{i,k,l} = \sum_q \lambda_{q,i,l} w_{i,k,q_{i,l}} h_{i,l} \boldsymbol{C}_{i,k,0} \hat{\boldsymbol{X}}_{k,l}^{-1} \boldsymbol{y}_{k,l}. \quad (25)$$

As evaluation measures, we used the signal-to-distortion ratio (SDR) and the signal-to-interference ratio (SIR) [16]. The SDR and SIR are expressed in decibels (dB), and a higher SDR(/SIR) indicates superior quality.

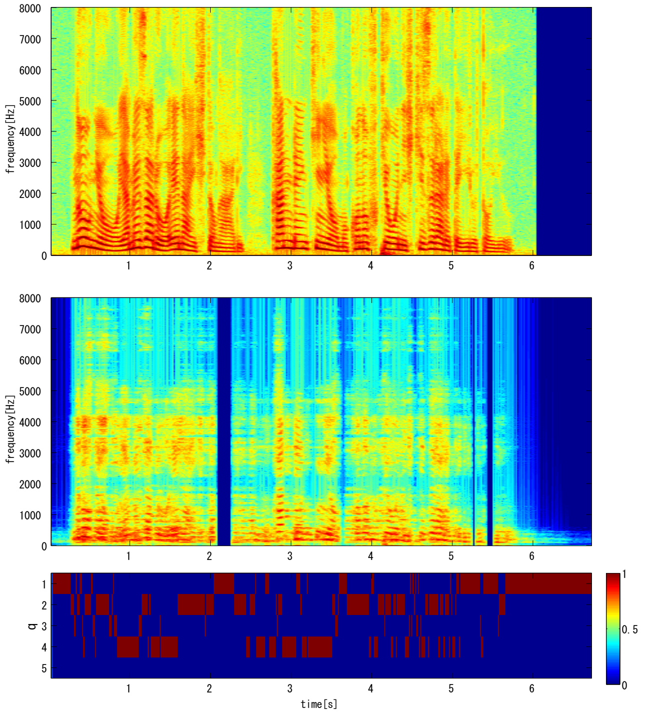

Fig. 3. A spectrogram of an anechoic source signal (top), that of a separated and dereverbed signal with the proposed method (middle) and the result of λ which corresponds to the estimated activity of the source (bottom). $q = 1$ corresponds to the inactive state.

Table II shows the average SDRs and SIRs obtained by our previous and present methods. The average SDR and SIR obtained with the present method were superior to these obtained with conventional method. Fig. 2 (bottom) shows an example of d obtained by the proposed method. We can see that the DOA of the sources were estimated roughly by the proposed method. Fig. 3 shows a spectrogram of an anechoic source signal (top), that of a separated and dereverbed signal with the proposed method (middle) and the result of λ which corresponds to the estimated activity of the source (bottom). We expected $q = 1$ is an inactive state by setting the hyperparameters of the gamma distribution properly (Table I), and the result shows the voice activity was roughly detected.

VII. Conclusion

In this paper we extend the MFHMM and propose a unified approach for source separation, source activity detection, dereverberation and DOA estimation. Specifically, we describe a spatial correlation matrix by using weight variable and kernels and incorporate into the MFHMM. Through the estimation of the parameters of the overall generative model, we can simultaneously performed source separation, source activity detection, dereverberation and DOA estimation. The experiment showed that the proposed algorithm were superior to our previous method in terms of the signal-to-distortion ratio with echoic mixed signals.

VIII. Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 26730100, 26280060.

References

[1] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[2] D. D. Lee, and H. S. Seung, "Learning the parts of objects with nonnegative matrix factorization," *Nature*, vol. 401, pp.788–791, 1999.

[3] P. Smaragdis, and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) 2003*, pp. 177–180, Oct. 2003.

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

[5] Hiroshi Sawada, Hirokazu Kameoka, Shoko Araki and Naonori Ueda, "Efficient algorithms for multichannel extensions of Itakura-Saito nonnegative matrix factorization," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2012*, pp. 261–264, 2012.

[6] Takuya Higuchi, Hirofumi Takeda, Tomohiko Nakamura and Hirokazu Kameoka, "A unified approach for underdetermined blind signal separation and source activity detection by multichannel factorial hidden Markov models," *The 15th Annual Conference of the International Speech Communication Association (Interspeech 2014)*, pp. 850–854, 2014.

[7] Takuya Higuchi and Hirokazu Kameoka, "Joint audio source separation and dereverberation based on multichannel factorial hidden Markov model," *The 24th IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2014)*.

[8] Hirokazu Kameoka, Misa Sato, Takuma Ono, Nobutaka Ono, Shigeki Sagayama, "Blind separation of infinitely many sparse sources," *The 13th International Workshop on Acoustic Signal Enhancement (IWAENC 2012)*, H-09, Sep. 2012.

[9] Takuya Higuchi, Norihiro Takamune, Tomohiko Nakamura, Hirokazu Kameoka, "Underdetermined blind separation and tracking of moving sources based on DOA-HMM," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2014*, pp. 3215-3219, May 2014.

[10] Joonas Nikunen and Tuomas Virtanen, "Multichannel audio separation by direction of arrival based spatial covariance model and non-negative matrix factorization," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2014*, pp. 6727–6731, May 2014.

[11] Tomohiro Nakatani, Takuya Yoshioka, Keisuke Kinoshita, Masato Miyoshi and Biing-Hwang Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2008*, pp. 85–88, 2008.

[12] C. Févotte, N. Bertin, and J. -L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence. with application to music analysis," *Neural Computation*, vol. 21, no. 3, 2009.

[13] Masahiro Nakano, Le Roux Jonathan, Hirokazu Kameoka, Yu Kitano, Nobutaka Ono, and Shigeki Sagayama, "Nonnegative matrix factorization with Markov-chained bases for modeling time-varying patterns in music spectrograms," *Latent Variable Analysis and Signal Separation*, vol. 6365, pp. 149-156, 2010.

[14] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano,, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, pp. 357–363, 1990.

[15] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada,"Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *2nd International Conference on Language Resources & Evaluation (LREC 2000)*, pp. 965–968, 2000.

[16] E. Vincent, R. Gribonval, and C. Févotte,"Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 1462–1469, 2006.