# UNIFIED APPROACH FOR AUDIO SOURCE SEPARATION WITH MULTICHANNEL FACTORIAL HMM AND DOA MIXTURE MODEL

*Takuya Higuchi*[1)] *and Hirokazu Kameoka*[1),2)]

[1)]Graduate School of Information Science and Technology, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
[2)]NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation
3-1 Morinosatowakamiya, Atsugi, Kanagawa 243-0198, Japan
{higuchi,kameoka}@hil.t.u-tokyo.ac.jp

## ABSTRACT

We deal with the problems of blind source separation, dereverberation, audio event detection and direction-of-arrival (DOA) estimation. We previously proposed a generative model of multichannel signals called the multichannel factorial hidden Markov model, which allows us to simultaneously solve these problems through a joint optimization problem formulation. In this approach, we modeled the spatial correlation matrix of each source as a weighted sum of the spatial correlation matrices corresponding to all possible DOAs. However, it became clear through real environment experiments that the estimate of the spatial correlation matrix tended to deviate from the actual correlation matrix since the plane wave assumption does not hold due to reverberation and noise components. To handle such deviations, we propose introducing a prior distribution over the spatial correlation matrices called the DOA mixture model instead of using the weighted sum model. The experiment showed that the proposed method provided 1.94 [dB] improvement compared with our previous method in terms of the the signal-to-distortion ratios of separated signals.

***Index Terms***— Blind source separation, voice activity detection, dereverberation, DOA estimation

## 1. 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 the BSS problem, 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. 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 us to use the spatial information as an additional clue to source separation [4, 5]. Moreover, we previously proposed the multichannel factorial hidden Markov model [6], where we used the information of audio events for source separation and simultaneously solved the problems of BSS and audio event detection. Furthermore, we extended the model 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. Furthermore, we used direction-of-arrivals (DOAs) of sources as a clue to estimation of spatial correlation matrices [8]. Generally, it is known that the spatial correlation matrix of the direct sound of a point source has a certain structure described by the DOA. In the model proposed in [8], we used a weighted sum of matrices parametrized by all possible DOAs, same as in [9], and estimates of the weight variables corresponded to estimates of spatial correlation matrices at all frequencies. However, in practical situations, a spatial correlation matrix often deviates from a theoretical structure because of reverberation and noises within the frame of the STFT. Therefore, in this paper, we apply the DOA mixture model [10] for modeling the generative process of a spatial correlation matrix, and design a prior distribution over a spatial correlation matrix based on DOAs. In the present model, we estimate a spatial correlation matrix at each frequency bin based on DOAs. 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.

This work was supported by JSPS KAKENHI Grant Number 26730100.

## 2. MULTICHANNEL FACTORIAL HIDDEN MARKOV MODEL

### 2.1. Mixing model with convolutive mixture approximation [7]

First we consider a situation where $I$ source signals are recorded by $M$ microphones. 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.

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

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 signal. $1 \leq k \leq K$ and $1 \leq l \leq L$ are the frequency

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

and time indices, respectively. $\boldsymbol{a}_i(\omega_k, t_\tau)$ denotes the frequency array response for source $i$ at frequency $\omega_k$ and time $t_\tau$. $0 \leq \tau \leq \mathcal{T}$ is the time index of the frequency array response in the time-frequency domain. Note that $\boldsymbol{a}_i(\omega_k, t_{1:\mathcal{T}})$ denote the frequency array responses which correspond to the impulse responses out of the frames of the STFT. For convenience of notation, we hereafter use subscripts $k$, $l$ and $\tau$ to indicate $\omega_k$, $t_l$ and $t_\tau$ respectively.

### 2.2. Generative modeling of source signals based on audio events [6]

We now describe the generative process of the source signal $s_{i,k,l}$ based on its audio event. 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

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

where $w_{i,k}$ and $h_{i,l}$ are non-negative variables. The generative model of $s_{i,k,l}$ is written as

$$\boldsymbol{s}_{i,k,l} | w_{i,k}, h_{i,l} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{s}_{i,k,l}; 0, w_{i,k} h_{i,l}), \tag{3}$$

conditioned on $w_{i,k}$ and $h_{i,l}$. Under the condition of Eq. (3), the generative model of the multichannel observed signal is equivalent to the method proposed in [4, 5].

In many cases, a sound signal consists of several spectra. Therefore, first we introduce latent variable $z^{(\mathrm{q})}_{i,l} \in \{1, \ldots, Q\}$ to denote a state of $i$-th source's spectrum at time $l$, and the state sequence $z^{(\mathrm{q})}_{i,1}, \ldots, z^{(\mathrm{q})}_{i,L}$ follows a Markov chain:

$$z^{(\mathrm{q})}_{i,l} | z^{(\mathrm{q})}_{i,l-1} \sim \mathrm{Categorical}(z^{(\mathrm{q})}_{i,l}; \boldsymbol{\rho}^{(\mathrm{q})}_{i, z^{(\mathrm{q})}_{i,l-1}}), \tag{4}$$

where $\mathrm{Categorical}(x; \boldsymbol{y}) = y_x$, $\boldsymbol{\rho}^{(\mathrm{q})}_{i,q} = (\rho^{(\mathrm{q})}_{i,q,1}, \ldots, \rho^{(\mathrm{q})}_{i,q,Q})$ denotes the transition probability of state $q$ to each state $1, \ldots, Q$, and $\boldsymbol{\rho}^{(\mathrm{q})}_i = (\rho^{(\mathrm{q})}_{i,q,q'})_{Q \times Q}$ denotes the transition matrix. Then the power spectrum of the $i$-th source at time $l$ is assumed to be determined according to $z^{(\mathrm{q})}_{i,l}$ and $w_{i,k,q}$ denotes a spectrum of $i$-th source at state $q$. Thus, the generative model of $s_{i,k,l}$ is rewritten as

$$s_{i,k,l} | w_{i,k,1:Q}, h_{i,l}, z^{(\mathrm{q})}_{i,l} \sim \mathcal{N}_{\mathbb{C}}(s_{i,k,l}; 0, w_{i,k,z^{(\mathrm{q})}_{i,l}} h_{i,l}). \tag{5}$$

Next, the source's power also tend to differ according to the source's state such as "active" or "inactive." Therefore we first introduce latent variable $z^{(\mathrm{j})}_{i,l} \in 1, \ldots, J$ to denote a state of $i$-th source's power at time $l$, and the state sequence $z^{(\mathrm{j})}_{i,1}, \ldots, z^{(\mathrm{j})}_{i,L}$ follows a Markov chain:

$$z^{(\mathrm{j})}_{i,l} | z^{(\mathrm{j})}_{i,l-1} \sim \mathrm{Categorical}(z^{(\mathrm{j})}_{i,l}; \boldsymbol{\rho}^{(\mathrm{j})}_{i, z^{(\mathrm{j})}_{i,l-1}}). \tag{6}$$

Then we assume $h_{i,l}$ follows a gamma distribution which has different parameters according to $z^{(\mathrm{j})}_{i,l}$,

$$h_{i,l} | z^{(\mathrm{j})}_{i,l} \sim \mathrm{Gamma}(h_{i,l}; \alpha_{z^{(\mathrm{j})}_{i,l}} \beta_{z^{(\mathrm{j})}_{i,l}}), \tag{7}$$

where $\alpha_{1:J}$ and $\beta_{1:J}$ are the shape and scale parameters of a gamma distribution, and $\mathrm{Gamma}(x; \alpha, \beta) = \frac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha) \beta^\alpha}$. As we want $h_{i,l}$ to take a small value when $z^{(\mathrm{j})}_{i,l}$ 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 are like uniform distributions.

### 2.3. Generative process of a spatial correlation matrix based on DOA mixture model

If we assume that a source is far from the microphones, the frequency array response has a certain structure depending on DOA. Therefore, we can express a spatial correlation matrix by using the DOAs. Since a spatial correlation matrix $\boldsymbol{C}_{i,k,\tau}$ is defined as $\boldsymbol{C}_{i,k,\tau} = \boldsymbol{a}_{i,k,\tau} \boldsymbol{a}^{\mathrm{H}}_{i,k,\tau}$ by using $\boldsymbol{a}_{i,k,\tau}$, specifically, with $M = 2$ microphones, the spatial correlation matrix for a source at direction $\theta$ such that $0 \leq \theta \leq \pi$ is defined as a function of $\omega$ depending on $\theta$

$$\boldsymbol{J}(\theta, \omega) = \begin{bmatrix} 1 \\ e^{\mathrm{J} \omega B \cos\theta / \mathcal{C}} \end{bmatrix} \left[ 1 \; e^{\mathrm{J} \omega B \cos\theta / \mathcal{C}} \right]^*, \tag{8}$$

where $\mathrm{J}$ is the imaginary unit, $B$ [m] is the distance between the two microphones, and $\mathcal{C}$ [m/s] is the speed of sound. If the DOA $\theta_i$ 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 and noises within the frame of the STFT. One way is modeling a spatial correlation matrix by using a weighted sum of $\boldsymbol{J}(\theta_i, \omega_k)$ of all possible DOAs, same as in [8, 9]. In that model, estimates of the weight variables corresponded to estimates of spatial correlation matrices at all frequencies, and so it would not be enough for modeling the deviation of spatial correlation matrices. Therefore, in this paper, we incorporate the DOA mixture model [10] into the multichannel factorial hidden Markov model, and estimate a spatial correlation matrix at each frequency bin using the information of DOAs. First, we now introduce a discrete set of $O$ possible directions, $\vartheta_1, \ldots, \vartheta_O$, which are all assumed to be constants. We then assume that each source signal propagates from one of these directions. For each source $i$, an index $z^{(\mathrm{o})}_i$ of direction is drawn according to a categorical distribution $\boldsymbol{\rho}^{(\mathrm{o})}_i = (\rho_{i,1}, \ldots, \rho_{i,O})$

$$z^{(\mathrm{o})}_i | \boldsymbol{\rho}^{(\mathrm{o})}_i \sim \mathrm{Categorical}(z^{(\mathrm{o})}_i; \boldsymbol{\rho}^{(\mathrm{o})}_i), \tag{9}$$

$$\theta_i = \vartheta_{z^{(\mathrm{o})}_i}. \tag{10}$$

Then, a spatial correlation matrix follows a whishart distribution:

$$\boldsymbol{C}_{i,k,0} | z^{(\mathrm{o})}_i \sim \mathcal{W}_{\mathbb{C}}(\boldsymbol{C}_{i,k,0}; \gamma, \boldsymbol{J}_{\vartheta_{z^{(\mathrm{o})}_i}, k} + \epsilon \boldsymbol{I}), \tag{11}$$

where $\mathcal{W}_{\mathbb{C}}(\boldsymbol{X}; \gamma, \boldsymbol{Y}) \propto |\boldsymbol{X}|^{(\gamma - M)/2} \exp(-\frac{1}{2} \mathrm{tr}(\boldsymbol{X} \boldsymbol{Y}^{-1}))$ and $\gamma (\geq M + 1)$ is a hyperparameter of the whishart distribution. Here $\boldsymbol{I}$ is an identity matrix and $\epsilon$ is a small value, and they enable inverse operation.

Our overall generative model is given by Eqs. (4), Eqs. (6), Eqs. (7) Eqs. (9), Eqs. (11) and

$$\boldsymbol{y}_{k,l} | \boldsymbol{a}_{1:I,k,0:\mathcal{T}}, w_{1:I,k,1:Q}, h_{1:I,l-\mathcal{T}:l}, z^{(\mathrm{q})}_{1:I,l-\mathcal{T}:l} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{y}_{k,l}; 0, \sum_{i,\tau} \boldsymbol{C}_{i,k,\tau} w_{i,k,z^{(\mathrm{q})}_{i,l-\tau}} h_{i,l-\tau}), \tag{12}$$

conditioned on $\boldsymbol{a}_{1:I,k,0:\mathcal{T}}$, $w_{1:I,k,1:Q}$, $h_{1:I,l-\mathcal{T}:l}$, $z^{(\mathrm{q})}_{1:I,l-\mathcal{T}:l}$. Parameter estimation of the generative model allows us to solve the problems of source separation, source activity detec-

tion, dereverberation and DOA estimation based on a unified maximum likelihood criterion.

## 3. PARAMETER ESTIMATION ALGORITHM BASED ON AUXILIARY FUNCTION METHOD

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 $\boldsymbol{W} = w_{1:I,1:K,1:Q}$, $\boldsymbol{H} = h_{1:I,1:L}$ and $\boldsymbol{C} = \boldsymbol{C}_{1:I,1:K,0:\mathcal{T}}$. We denote the entire set of the above parameters as $\boldsymbol{\Theta}$. $\boldsymbol{Z}^{(\mathrm{q})} = z^{(\mathrm{q})}_{1:I,1:L}$, $\boldsymbol{Z}^{(\mathrm{j})} = z^{(\mathrm{j})}_{1:I,1:L}$ and $\boldsymbol{Z}^{(\mathrm{o})} = z^{(\mathrm{o})}_{1:I}$ are latent variables. In the following, $\boldsymbol{\rho}^{(\mathrm{q})}$, $\boldsymbol{\rho}^{(\mathrm{j})}$, $\boldsymbol{\rho}^{(\mathrm{o})}$, $\alpha$ and $\beta$ are constants that are determined experimentally.

The objective function $L(\boldsymbol{\Theta})$ is defined as $L(\boldsymbol{\Theta}) = \log p(\boldsymbol{\Theta}|\boldsymbol{Y})$, where $\boldsymbol{Y} = \boldsymbol{y}_{1:K,1:L}$ is a set consisting of the time-frequency components of observed multichannel signals. By using the conditional distributions defined in Sec. 2, we can write down $\log p(\boldsymbol{\Theta}|\boldsymbol{Y})$. However, $L(\boldsymbol{\Theta})$ is difficult to maximize analytically, and so we derive an efficient parameter estimation algorithm based on an auxiliary function method.

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

$$
\begin{aligned}
L(\boldsymbol{\Theta}) \geq & L^+(\boldsymbol{\Theta}, \boldsymbol{\Lambda}) \\
= & -\frac{1}{2}\sum_{k,l}\Bigg[\sum_{i,q,\tau}\lambda^{(\mathrm{q})}_{q,i,l-\tau} \\
& \Bigg(\frac{\mathrm{tr}(\boldsymbol{y}_{k,l}\boldsymbol{y}_{k,l}{}^{\mathsf{H}}\boldsymbol{R}_{i,k,l,q,\tau}\boldsymbol{C}^{-1}_{i,k}\boldsymbol{R}_{i,k,l,q,\tau})}{w_{i,k,q}h_{i,l-\tau}} \\
& + \mathrm{tr}(\boldsymbol{U}^{-1}_{k,l}\boldsymbol{C}_{i,k})w_{i,k,q}h_{i,l-\tau}\Bigg) + \log|\boldsymbol{U}_{k,l}|\Bigg] \\
& + \sum_{j,i,l}\lambda^{(\mathrm{j})}_{j,i,l}\Bigg((\alpha_j - 1)\log h_{i,l} - h_{i,l}/\beta_j \\
& - \alpha_j \log\beta_j - \log\Gamma(\alpha_j)\Bigg) \\
& + \sum_{i,k,o} d_{i,o}\Bigg[-\frac{1}{2}\mathrm{tr}(\boldsymbol{C}_{i,k,0}(\boldsymbol{J}_{k,o} + \epsilon\boldsymbol{I})^{-1}) \\
& + \frac{\gamma - M}{2}\Bigg(\mathrm{tr}(\tilde{\boldsymbol{U}}^{-1}_{i,k}\boldsymbol{C}_{i,k,0}) \\
& + \log|\tilde{\boldsymbol{U}}_{i,k}|\Bigg)\Bigg] + \sum_{q,i,l}\lambda^{(\mathrm{q})}_{q,i,l}\log p(\boldsymbol{Z}^{(\mathrm{q})}) \\
& + \sum_{j,i,l}\lambda^{(\mathrm{j})}_{j,i,l}\log p(\boldsymbol{Z}^{(\mathrm{j})}) + C_{L^+}, \qquad (13)
\end{aligned}
$$

where $\boldsymbol{R}_{i,k,l,\tau,q}$, $\boldsymbol{U}_{k,l}$, $\tilde{\boldsymbol{U}}_{i,k}$, $\lambda^{(\mathrm{q})}_{q,i,l}$, $\lambda^{(\mathrm{j})}_{j,i,l}$ and $d_{i,o}$ are auxiliary variables. $\boldsymbol{R}_{i,k,l,\tau,q}$, $\boldsymbol{U}_{k,l}$ and $\tilde{\boldsymbol{U}}_{i,k}$ satisfy Hermitian positive definiteness and $\sum_{i,\tau}\boldsymbol{R}_{i,k,l,\tau,q} = \boldsymbol{I}$. $\lambda^{(\mathrm{q})}_{q,i,l}$, $\lambda^{(\mathrm{j})}_{j,i,l}$ and $d_{i,o}$ satisfy $\sum_q \lambda^{(\mathrm{q})}_{q,i,l} = 1$, $\sum_j \lambda^{(\mathrm{j})}_{j,i,l} = 1$ and $\sum_o d_{i,o} = 1$. We denote the set of the auxiliary variables as $\boldsymbol{\Lambda}$. $C_{L^+}$ is the sum of constant terms. $\mathrm{tr}(\cdot)$ is the trace of a matrix. The equality $L(\boldsymbol{\Theta}) = L^+(\boldsymbol{\Theta}, \Lambda)$ is satisfied when

$$\boldsymbol{R}_{i,k,l,\tau,q} = \boldsymbol{C}_{i,k,\tau}w_{i,k,q}h_{i,l-\tau}\hat{\boldsymbol{X}}^{-1}_{k,l}, \qquad (14)$$

$$\boldsymbol{U}_{k,l} = \hat{\boldsymbol{X}}_{k,l}, \qquad (15)$$

$$\tilde{\boldsymbol{U}}_{i,k} = \boldsymbol{C}_{i,k,0}, \qquad (16)$$

with respect to $\boldsymbol{R}_{i,k,l,\tau,q}$, $\boldsymbol{U}_{k,l}$ and $\tilde{\boldsymbol{U}}_{i,k}$. With respect to $\lambda^{(\mathrm{q})}_{q,i,l}$,

$$\lambda^{(\mathrm{q})}_{q,i,l} = F^{(\mathrm{q})}_{q,i,l}B^{(\mathrm{q})}_{q,i,l} / \sum_q F^{(\mathrm{q})}_{q,i,l}B^{(\mathrm{q})}_{q,i,l}, \qquad (17)$$

$$F^{(\mathrm{q})}_{q,i,l} = p(\boldsymbol{\Theta}|z^{(\mathrm{q})}_{i,l} = q)\sum_{q'} F^{(\mathrm{q})}_{q,i,l-1}\rho^{(\mathrm{q})}_{i,q',q}, \qquad (18)$$

$$B^{(\mathrm{q})}_{q,i,l} = \sum_{q'} B^{(\mathrm{q})}_{q',i,l+1}p(\boldsymbol{\Theta}|z^{(\mathrm{q})}_{i,l+1} = q')\rho^{(\mathrm{q})}_{i,q,q'}, \qquad (19)$$

where

$$
\begin{aligned}
& p(\boldsymbol{\Theta}|z^{(\mathrm{q})}_{i,l} = q) \\
\propto\ & \exp\Bigg[-\frac{1}{2}\sum_{i,k,\tau}\Big(\mathrm{tr}(\boldsymbol{y}_{k,l+\tau}\boldsymbol{y}_{k,l+\tau}{}^{\mathsf{H}}\boldsymbol{R}_{i,k,l+\tau,\tau,q} \\
& \boldsymbol{C}^{-1}_{i,k,\tau}\boldsymbol{R}_{i,k,l+\tau,\tau,q})/w_{i,k,q}h_{i,l} \\
& + \mathrm{tr}(\boldsymbol{U}^{-1}_{k,l+\tau}\boldsymbol{C}_{i,k,\tau})w_{i,k,q}h_{i,l}\Big)\Bigg]. \qquad (20)
\end{aligned}
$$

With respect to $\lambda^{(\mathrm{j})}_{j,i,l}$,

$$\lambda^{(\mathrm{j})}_{j,i,l} = F^{(\mathrm{j})}_{j,i,l}B^{(\mathrm{j})}_{j,i,l} / \sum_j F^{(\mathrm{j})}_{j,i,l}B^{(\mathrm{j})}_{j,i,l}, \qquad (21)$$

$$F^{(\mathrm{j})}_{j,i,l} = p(\boldsymbol{\Theta}|z^{(\mathrm{j})}_{i,l} = j)\sum_{j'} F^{(\mathrm{j})}_{j,i,l-1}\rho^{(\mathrm{j})}_{i,j',j}, \qquad (22)$$

$$B^{(\mathrm{j})}_{j,i,l} = \sum_{j'} B^{(\mathrm{j})}_{j',i,l+1}p(\boldsymbol{\Theta}|z^{(\mathrm{j})}_{i,l+1} = j')\rho^{(\mathrm{j})}_{i,j,j'}, \qquad (23)$$

where

$$
\begin{aligned}
& p(\boldsymbol{\Theta}|z^{(\mathrm{j})}_{i,l} = j) \\
\propto\ & \exp\Big((\alpha_j - 1)\log h_{i,l} - h_{i,l}/\beta_j \\
& - \alpha_j\log\beta_j - \log\Gamma(\alpha_j)\Big). \qquad (24)
\end{aligned}
$$

With respect to $d_{i,o}$,

$$
\begin{aligned}
d_{i,o} &= p(\boldsymbol{\Theta}|z^{(\mathrm{o})}_i = o) \\
&\propto \exp(-\frac{1}{2}\mathrm{tr}(\boldsymbol{C}_{i,k,0}(\boldsymbol{J}_{\vartheta_o,k} + \epsilon\boldsymbol{I})^{-1})), \qquad (25)
\end{aligned}
$$

**Table 1**. The experimental condition of parameters

| Q | O | $\alpha$ | $\beta$ | $\gamma$ | $\epsilon$ | $\epsilon'$ |
|---|---|---|---|---|---|---|
| 15 | 30 | 1 | $\beta_1 = 10^{-3}, \beta_2 = 10^{10}$ | 3 | $10^{-3}$ | $10^{-3}$ |

and $\sum_o d_{i,o} = 1$.

By setting the partial differential of $L^+$ at zero, we can obtain the update rule of each parameter in $\mathbf{\Theta}$. Specifically, the partial differential of $L^+$ with respect to $\boldsymbol{C}_{i,k,0}$ is given by

$$\begin{aligned}&\frac{\partial L^+}{\partial \boldsymbol{C}_{i,k,0}}\\ =&\frac{1}{2}\sum_{l,q}\lambda_{q,i,l}\Bigg(\frac{\boldsymbol{C}_{i,k,0}^{-1}\boldsymbol{R}_{i,k,l,0,q}\boldsymbol{y}_{k,l}\boldsymbol{y}_{k,l}{}^{\mathsf{H}}\boldsymbol{R}_{i,k,l,0,q}\boldsymbol{C}_{i,k,0}^{-1}}{w_{i,k,q}h_{i,l}}\\ &-\boldsymbol{U}_{k,l}^{-1}w_{i,k,q}h_{i,l}\Bigg)+\sum_o d_{i,o}\Bigg(-\frac{1}{2}(\boldsymbol{J}_{\vartheta_o,k}+\epsilon\boldsymbol{I})^{-1})\Bigg)\\ &+\frac{\gamma-M}{2}\tilde{\boldsymbol{U}}_{i,k}^{-1}.\end{aligned} \tag{26}$$

By setting this at zero, we obtain an algebraic Riccati equation:

$$\boldsymbol{C}_{i,k,0}\boldsymbol{A}_{i,k,0}\boldsymbol{C}_{i,k,0} = \boldsymbol{B}_{i,k,0}, \tag{27}$$

where

$$\begin{aligned}\boldsymbol{A}_{i,k,0} =&\sum_{l,q}\lambda_{q,i,l}w_{i,k,q}h_{i,l}\hat{\boldsymbol{X}}_{k,l}^{-1}\\ &+\sum_o d_{i,o}\Bigg(-(\boldsymbol{J}_{k,o}+\epsilon\boldsymbol{I})^{-1})\Bigg)+(\gamma-M)\tilde{\boldsymbol{U}}_{i,k}^{-1},\end{aligned} \tag{28}$$

$$\boldsymbol{B}_{i,k,0} =\boldsymbol{C}_{i,k,0}(\sum_{l,q}\lambda_{q,i,l}w_{i,k,q}h_{i,l}\hat{\boldsymbol{X}}_{k,l}^{-1}\boldsymbol{y}_{k,l}\boldsymbol{y}_{k,l}{}^{\mathsf{H}}\hat{\boldsymbol{X}}_{k,l}^{-1})\boldsymbol{C}_{i,k,0}. \tag{29}$$

We can solve this equation with the same procedure proposed in [5] and obtain the update rule of $\boldsymbol{C}_{i,k,0}$. For updating $\boldsymbol{C}_{i,k,0}$, we must calculate the inverse of $\tilde{\boldsymbol{U}}_{i,k}$. Even if $\boldsymbol{C}_{i,k,0}$ equals $\boldsymbol{J}_{\theta_{i,k}}+\epsilon\boldsymbol{I}$, we can calculate the inverse of $\tilde{\boldsymbol{U}}_{i,k}$ which is updated by Eq. (16). However, experimentally, $\tilde{\boldsymbol{U}}_{i,k}$ would not always be a full-rank matrix in the iterations. Therefore, we update $\tilde{\boldsymbol{U}}_{i,k}$ approximately as:

$$\tilde{\boldsymbol{U}}_{i,k} \approx \boldsymbol{C}_{i,k,0} + \epsilon'\boldsymbol{I}, \tag{30}$$

where $\epsilon'$ is a small value. On account of the space, we omit the update rules of the other parameters in $\mathbf{\Theta}$, however, they can be also obtained analytically by setting the partial differentials of $L^+$ at zeros.

## 4. EXPERIMENTAL EVALUATION

We evaluated the performance of the present method in terms of the abilities of source separation. We used 5 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 [11] and was convolved with the measured room impulse response from the RWCP database [12] (in which the distance between the microphones was 11.48 cm and the reverberation time was 0 ms/380 ms). Thus, the three signals were artificially located $30°$, $90°$ and $130°$ from the microphones respectively. Fig. 1 shows an example of 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 hyperparameters as Table 1. We expected that $j = 1$ is an inactive state and $j = 2$ is an active state. We set $\mathcal{T}$ as 0 for the anechoic mixed signals, as 3 for echoic ones. $d$ was initially set as shown in Fig. 2 (a). $\boldsymbol{C}_{i,k,0}$ was set to $1/\sqrt{M} \times \boldsymbol{I}$, $\boldsymbol{C}_{i,k,1:\mathcal{T}}$ were set to $10^{-2}/\sqrt{M} \times \boldsymbol{I}$. $\boldsymbol{W}$ was initially randomized. $\boldsymbol{H}$ was set as 1 initially. $\boldsymbol{\rho}_i^{(\mathrm{j})}$ was set as $\boldsymbol{\rho}_{i,1}^{(\mathrm{j})} = (0.9, 0.1)$ and $\boldsymbol{\rho}_{i,2}^{(\mathrm{j})} = (0.1, 0.9)$, all components of $\boldsymbol{\rho}_i^{(\mathrm{q})}$ were set as $1/Q$. We set $\boldsymbol{\rho}^{(\mathrm{o})}$ individually for each source so that directions of 3 sources did not overlap each other. The parameter estimation algorithm was run for 50 iterations. In order to avoid an undesirable local optima, we iterated the proposed algorithm 25 times with setting $\mathcal{T}$ as 0, then gradually increased $\mathcal{T}$ up to 3 according to the iteration. Moreover, in order to avoid huge numerical errors, we fixed $\boldsymbol{W}$ when $\mathcal{T} > 0$. The numerical errors occurred probably because we fitted the product of $\boldsymbol{W}_{i,k}$ and $\boldsymbol{C}_{i,k,1:\mathcal{T}}$ to observed signals, and so each of $\boldsymbol{W}_{i,k}$ and $\boldsymbol{C}_{i,k,1:\mathcal{T}}$ was arbitrary scaled. We chose the methods proposed in [7] and [8] as comparisons. The method of [7] does not use the information of DOAs, and that of [8] models a spatial correlation matrix by using a weighted sum of $\boldsymbol{J}$. The separated signal was obtained by Wiener filtering. As evaluation measures, we used the signal-to-distortion ratio (SDR) and the signal-to-interference ratio (SIR) [13]. The SDR and SIR are expressed in decibels (dB), and a higher SDR(/SIR) indicates superior quality. The SDRs and SIRs were calculated with magnitude spectrograms. The averages of SDRs and SIRs of the mixed signals were -1.98 and -1.04[dB] when RT=0 ms, -4.77, -0.97[dB] when RT=380 ms respectively.

**Table 2**. SDR and SIR improvements obtained by the proposed method, the method in [7] and [8]

| | RT=0 ms | | RT=380 ms | |
|---|---|---|---|---|
| | SDR | SIR | SDR | SIR |
| method in [7] | -0.53 | 4.65 | -1.55 | 3.73 |
| method in [8] | 1.90 | 7.88 | -1.57 | 4.99 |
| proposed | 3.78 | 14.12 | 0.37 | 8.65 |

Table 2 shows the average SDR and SIR improvements obtained by our previous and present methods. The average SDRs and SIRs obtained with the present method were superior to these obtained with our previous methods. Fig. 2 (b) and (c) show examples of $d$ obtained by the proposed method when RT=0 ms and RT=380 ms respectively. The arrows show the actual direction of the sources. We can see that the DOA of the sources were estimated roughly by the proposed method. Fig. 3 shows (a)a spectrogram of an anechoic source signal, (b)that of a separated and dereverbed signal with the proposed method and (c)the result of $\lambda^{(\mathrm{j})}$ which corresponds to the estimated activity of the source obtained with the mixed signal showed in Fig. 1. We expected $q = 1$ was an inactive state by setting the hyperparameters of the gamma distribution properly (see Table 1), and the result shows the voice activity was roughly detected.

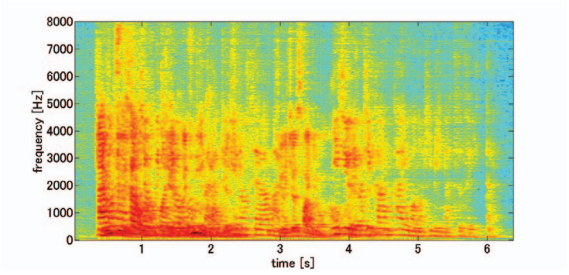


**Fig. 1**. An example of the spectrogram of the mixed signal (RT=380 ms).

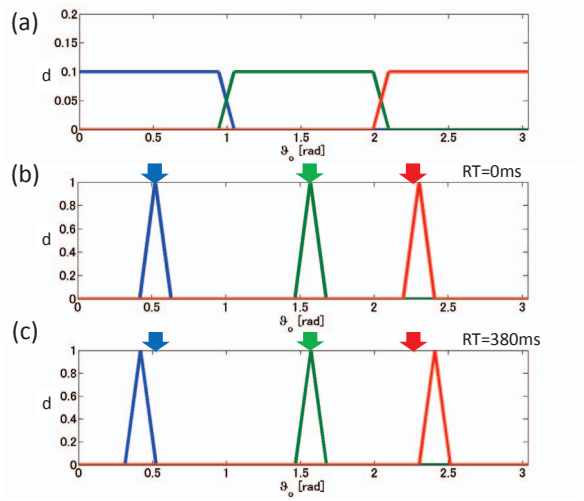


**Fig. 2**. (a)First condition of $d$, (b)an example of the estimation results of $d$ obtained with the mixed signal (RT=0 ms) and (c)that obtained with the mixed signal (RT=380 ms). $d$ of each source is colored differently. $d$ corresponds to how likely the source is at the direction.

## 5. CONCLUSION

In this paper, we incorporate the DOA mixture model into the multichannel factorial hidden Markov model. Parameter estimation allows us to solve the problems of BSS, VAD, dereverberation and DOA estimation simultaneously. The average of SDRs obtained by the proposed method was 1.94 dB higher than that obtained by our previous method.

## REFERENCES

[1] A. Hyvàrinen *et al.*, *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," *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. ASLP*, vol. 18, no. 3, pp. 550–563, 2010.

[5] H. Sawada *et al.*, "Efficient algorithms for multichannel extensions of Itakura-Saito nonnegative matrix factorization," *ICASSP 2012*, pp. 261–264, 2012.

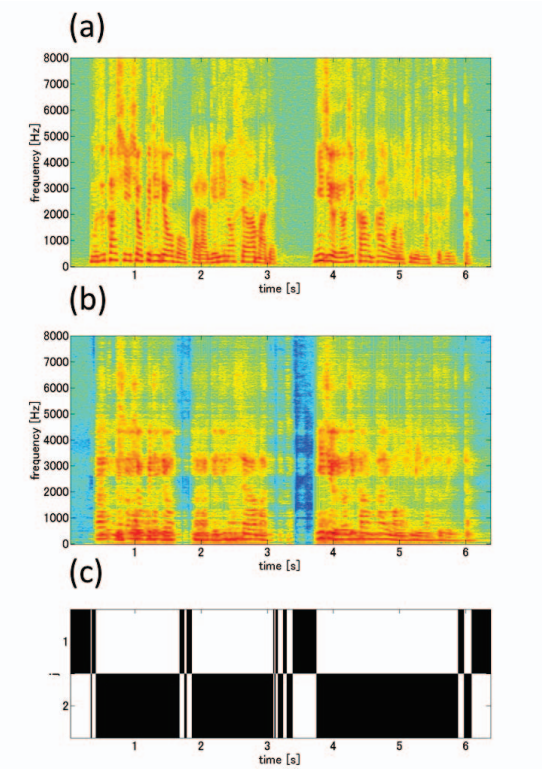


**Fig. 3**. (a)An example of a spectrogram of a source signal, (b)that of the separated and dereverbed signal obtained by the proposed method and (c)the estimation result of $\lambda^{(j)}$ which corresponds to source's activity obtained with the mixed signal showed in Fig. 1. Black indicates higher value and the state is more likely to be assigned at that time in (c).

[6] T. Higuchi, *et al.*, "A unified approach for underdetermined blind signal separation and source activity detection by multichannel factorial hidden Markov models," *Interspeech 2014*, pp. 850–854, 2014.

[7] T. Higuchi and H. Kameoka, "Joint audio source separation and dereverberation based on multichannel factorial hidden Markov model," *MLSP 2014*.

[8] T. Higuchi and H. Kameoka, "Unified approach for underdetermined BSS, VAD, dereverberation and DOA estimation with multichannel factorial HMM," *GlobalSIP 2014*, pp. 725-729, 2014.

[9] J. Nikunen and T. Virtanen, "Multichannel audio separation by direction of arrival based spatial covariance model and non-negative matrix factorization," *ICASSP 2014*, pp. 6727–6731, 2014.

[10] H. Kameoka *et al.*, "Blind separation of infinitely many sparse sources," *IWAENC 2012*, pp. 1–4, 2012.

[11] A. Kurematsu *et al.*, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, pp. 357–363, 1990.

[12] S. Nakamura *et al.*, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," *LREC 2000*, pp. 965–968, 2000.

[13] E. Vincent *et al.*, "Performance measurement in blind audio source separation," *IEEE Trans. ASLP*, pp. 1462–1469, 2006.