UNDERDETERMINED BLIND SEPARATION AND TRACKING OF MOVING SOURCES BASED ON DOA-HMM

Takuya Higuchi[†], *Norihiro Takamune*[†], *Tomohiko Nakamura*[†] *and Hirokazu Kameoka*[†‡]

[†]Graduate School of Information Science and Technology, The University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
[‡]NTT Communication Science Laboratories, NTT Corporation,
3-1 Morinosatowakamiya, Atsugi, Kanagawa 243-0198, Japan

ABSTRACT

This paper deals with the problem of the underdetermined blind separation and tracking of moving sources. In practical situations, sound sources such as human speakers can move freely and so blind separation algorithms must be designed to track the temporal changes of the impulse responses. We propose solving this problem through the posterior inference of the parameters in a generative model of an observed multichannel signal, formulated under the assumption of the sparsity of time-frequency components of speech and the continuity of speakers' movements. Specifically, we describe a generative model of mixture signals by incorporating a generative model of a time-varying frequency array response for each source, described using a path-restricted hidden Markov model (HMM). Each hidden state of the present HMM represents the direction of arrival (DOA) of each source, and so we call it a "DOA-HMM." Through the posterior inference of the overall generative model, we can simultaneously track the DOAs of sources, separate source signals and perform permutation alignment. The experiment showed that the proposed algorithm provided a 6.20 dB improvement compared with the conventional method in terms of the signal-to-interference ratio.

Index Terms— Underdetermined blind separation, moving sources, direction of arrival, hidden Markov model, variational inference

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. The best-known commercial application of BSS techniques is their use in teleconferencing systems. In practical situations, sound sources and microphones are likely to move during sound recording, and the transfer characteristics between the sources and the microphones can change accordingly. Many conventional BSS algorithms have been developed on the assumption that the array response is time-invariant, and thus they do not work satisfactorily when the sources or microphones move. This paper deals with a BSS problem in a situation where the array response has the potential to change over time according to the movements of the sources.

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.

One successful approach for underdetermined BSS is to utilize the fact that the time-frequency components of speech are near zero at most of the time-frequency points [2–9]. This implies that the time-frequency components of speech rarely overlap each other even when multiple speakers are speaking simultaneously. Hence, the main focus of this approach is how to design a time-frequency mask with which we can extract only the components of target speech from the mixture. To exploit the sparse nature of speech, we must convert observed signals to a time-frequency representation. In contrast to a time domain formulation of the BSS problem, a time-frequency domain formulation requires us to solve an additional problem called the permutation alignment problem. That is, we must group together the separated components of different frequency bins that are considered to originate from the same source in order to construct a separated signal. To solve the BSS and permutation alignment problems simultaneously, we have previously proposed constructing a hierarchical generative model consisting of generative models of an observed signal and the array response for each source. This model has allowed us to perform source separation, permutation alignment and direction-of-arrival estimation of the sources simultaneously through the posterior inference [8]. However, this approach has relied on the assumption of fixed source positions.

This paper proposes extending our previous approach to deal with an underdetermined BSS problem allowing for moving sources. Specifically, we formulate a generative model of mixture signals by incorporating a generative model of the time-varying frequency array response for each source, described using a path-restricted hidden Markov model (HMM). Each hidden state of the present HMM represents the direction of arrival (DOA) of each source, and thus we call it a "DOA-HMM." Through the posterior inference of the proposed generative model, we can simultaneously track the DOAs of sources, separate source signals and perform permutation alignment. Sec. 2 reviews our previous generative model designed under a static source position assumption [8] and Sec. 3 extends this model to allow for moving sources by introducing a DOA-HMM.

2. GENERATIVE MODEL FOR STATIC SOURCES

2.1. Mixing model

First we consider a situation where I fixed source signals are recorded by M microphones. 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. $1 \leq k \leq K$ and $1 \leq l \leq L$ are the frequency and time indices, respectively. If we assume that the length of the impulse response from a source to microphones is sufficiently shorter than the frame length of the STFT, the observed signal can be approximated

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

fairly well by an instantaneous mixture in the frequency domain:

$$\boldsymbol{y}(\omega_k, t_l) = \sum_{i=1}^{I} \boldsymbol{a}_i(\omega_k) s_i(\omega_k, t_l) + \boldsymbol{n}(\omega_k, t_l), \quad (1)$$

where $\boldsymbol{y}(\omega_k, t_l) = (y_1(\omega_k, t_l), \ldots, y_M(\omega_k, t_l))^{\mathsf{T}} \in \mathbb{C}^M$ and $\boldsymbol{s}(\omega_k, t_l) = (s_1(\omega_k, t_l), \ldots, s_I(\omega_k, t_l))^{\mathsf{T}} \in \mathbb{C}^I$. $\boldsymbol{a}_i(\omega_k)$ denotes the frequency array response for source i at frequency ω_k, which is assumed to be time-invariant throughout this section. $\boldsymbol{n}(\omega_k, t_l)$ is assumed to contain all kinds of components such as background noise and reverberant components, which cannot be represented by the instantaneous mixture representation.

We now utilize the sparseness of speech and assume that only one source is active in each time-frequency bin. By using $z_{k,l} \in \{1, \ldots, I\}$ to denote the (unknown) active source index at time-frequency bin (ω_k, t_l), Eq. (1) can be rewritten as

$$\boldsymbol{y}(\omega_k, t_l) = \boldsymbol{a}_{z_{k,l}}(\omega_k) s(\omega_k, t_l) + \boldsymbol{n}(\omega_k, t_l). \quad (2)$$

Notice that the subscript i is dropped from $s_i(\omega_k, t_l)$ in Eq. (2) as it is no longer necessary since we are assuming $s_i(\omega_k, t_l) = 0$ for $i \neq z_{k,l}$. For convenience of notation, we hereafter use subscripts k and l to indicate ω_k and t_l respectively.

2.2. Generative process of observed signals

Here we describe the generative process of an observed signal based on Eq. (2). We assume that the noise component $\boldsymbol{n}_{k,l}$ follows a complex normal distribution with mean $\mathbf{0}$ and covariance $\Sigma_k^{(n)}$. Then, from Eq. (2), $\boldsymbol{y}_{k,l}$ is also normally distributed such that

$$\boldsymbol{y}_{k,l} | \boldsymbol{a}_{1:I,k}, s_{k,l}, z_{k,l} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{y}_{k,l}; \boldsymbol{a}_{z_{k,l},k} s_{k,l}, \Sigma_k^{(n)}), \quad (3)$$

conditioned on $\boldsymbol{a}_{1:I,k} = \{\boldsymbol{a}_{1,k}, \ldots, \boldsymbol{a}_{I,k}\}$, $s_{k,l}$ and $z_{k,l}$, where $\mathcal{N}_{\mathbb{C}}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) \propto \exp(-(\boldsymbol{x} - \boldsymbol{\mu})^{\mathsf{H}} \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}))$. Moreover, we assume that $z_{k,l}$ derives from a discrete uniform distribution independently among all k and l.

2.3. Generative process of frequency array responses

We now describe the generative process of the frequency array response $\boldsymbol{a}_{i,k}$ by introducing a latent variable indicating the DOA of each source.

Thus far we have treated $\boldsymbol{a}_{i,k}$ as an independent parameter across k. If the index i indicates an identical source across $\omega_1, \ldots, \omega_K$, $\boldsymbol{a}_{i,k}$ will have a certain structure that can be described using the property of acoustic wave propagation. We therefore expect that the incorporation of an appropriate constraint into $\boldsymbol{a}_{i,k}$ would help solve both the permutation alignment problem and the frequency-wise source separation problem simultaneously through parameter inference. If each source is assumed to be located far from the microphones so that the signal can be treated approximately as a plane wave, the time difference between the microphones depends only on the DOA of the source. Since the time delay between two microphones corresponds to the phase difference of the frequency response of the microphone array, the complex array response can be expressed explicitly by using the DOAs. Specifically, with $M = 2$ microphones, the complex array response for a source at direction θ such that $0 \leq \theta \leq \pi$ is defined as a function of ω depending on θ

$$\boldsymbol{h}(\theta, \omega) = \begin{bmatrix} 1 \\ e^{\jmath \omega B \cos\theta / C} \end{bmatrix}, \quad (4)$$

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 frequency array response $\boldsymbol{a}_{i,k}$ should be equal to $\boldsymbol{h}(\theta_i, \omega_k)$. Since the DOAs are unobservable, we regard the DOA of each source as a latent variable and further consider describing its generative process.

We now introduce a discrete set of D possible directions, $\vartheta_1, \ldots, \vartheta_D$, which are all assumed to be constants. For instance, ϑ_d is defined as $\vartheta_d = (d-1)\pi/D, (d = 1, \ldots, D)$, which means dividing π into D equal angels. We then assume that each source signal propagates from one of these directions. First, we consider the generative process of the DOA θ_i of source i. For each source i, an index c_i of direction is drawn according to a categorical distribution $\boldsymbol{\rho}_i = (\rho_{i,1}, \ldots, \rho_{i,D})$

$$c_i | \boldsymbol{\rho}_i \sim \text{Categorical}(c_i; \boldsymbol{\rho}_i), \quad (5)$$

where $\text{Categorical}(x; \boldsymbol{y}) = y_x$. Given c_i, θ_i is given as ϑ_{c_i}. Since $\boldsymbol{a}_{i,k}$ may deviate from $\boldsymbol{h}(\vartheta_{c_i}, \omega_k)$ due to such factors as the plane wave assumption and the narrow band instantaneous mixture approximation, we assume that the frequency array response $\boldsymbol{a}_{i,k}$ is generated from a complex normal distribution with mean $\boldsymbol{h}(\vartheta_{c_i}, \omega_k)$, given c_i,

$$\boldsymbol{a}_{i,k} | c_i \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{a}_{i,k}; \boldsymbol{h}(\vartheta_{c_i}, \omega_k), \Sigma_k^{(a)}), \quad (6)$$

where $\Sigma_k^{(a)}$ denotes the covariance of the complex normal distribution, which is assumed to be a constant.

3. GENERATIVE MODEL FOR MOVING SOURCES

In the model described in Sec. 2, the position of each source is assumed to be fixed and hence the frequency array response $\boldsymbol{a}_{i,k}$ does not depend on time l. However, in practical situations, sound sources can move during sound recording and so the frequency array response $\boldsymbol{a}_{i,k}$ may vary accordingly. To allow a time-varying frequency array response, we represent the frequency array response as a time sequence $\boldsymbol{a}_{i,k,1}, \ldots, \boldsymbol{a}_{i,k,L}$. Eqs. (2), (3) should thus be rewritten as

$$\boldsymbol{y}_{k,l} = \boldsymbol{a}_{z_{k,l},k,l} s_{k,l} + \boldsymbol{n}_{k,l}, \quad (7)$$

$$\boldsymbol{y}_{k,l} | \boldsymbol{a}_{1:I,k,l}, s_{k,l}, z_{k,l} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{y}_{k,l}; \boldsymbol{a}_{z_{k,l},k,l} s_{k,l}, \Sigma_k^{(n)}). \quad (8)$$

In the same way, we replace the DOA indicator variable c_i with a time sequence $c_{i,1}, \ldots, c_{i,L}$ and assume

$$\boldsymbol{a}_{i,k,l} | c_{i,l} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{a}_{i,k,l}; \boldsymbol{h}(\vartheta_{c_{i,l}}, \omega_k), \Sigma_k^{(a)}). \quad (9)$$

Now, if we simply treat $c_{i,1}, \ldots, c_{i,L}$ as independent free variables we must solve a BSS problem independently for each time l. This implies the need to solve another permutation problem in the time direction. Namely, we must first solve source separations in a frame-by-frame manner and then group together the separated signals over all l's that are considered to originate from the same source. However, it would be better if we could join these two processes, as they are intrinsically interdependent. To solve the permutation alignment in the time direction, frame-by-frame source separation and permutation alignment in the frequency direction simultaneously, we need certain assumptions about the time sequence of the DOAs.

Here we assume that each source moves continuously and thus the DOA of each source varies continuously. That is, the DOA of a source at time l is assumed to be close to that at time $l-1$. To incorporate this assumption into our generative model, we propose modeling the time sequence of the frequency array responses $\boldsymbol{a}_{i,k,1}, \ldots, \boldsymbol{a}_{i,k,L}$ using an HMM in which the direction indices $d = 1, \ldots, D$ are regarded as the hidden states (Fig. 1). Thus, Eq. (9) can be seen as a state emission density. The state sequence $c_{i,1}, \ldots, c_{i,L}$ follows a Markov chain:

$$c_{i,l} | c_{i,l-1} \sim \text{Categorical}(c_{i,l}; \boldsymbol{\rho}_{c_{i,l-1}}), \quad (10)$$

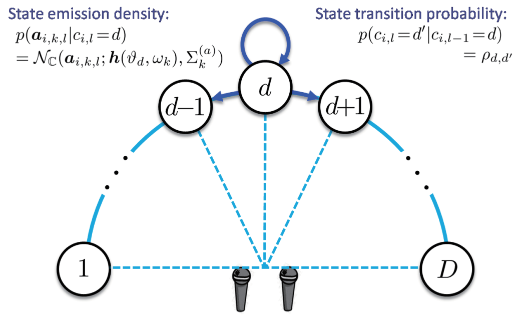
State emission density:
$p(\boldsymbol{a}_{i,k,l}|c_{i,l}=d)$
$=\mathcal{N}_{\mathbb{C}}(\boldsymbol{a}_{i,k,l};\boldsymbol{h}(\vartheta_d,\omega_k),\Sigma_k^{(a)})$
State transition probability:
$p(c_{i,l}=d'|c_{i,l-1}=d)$
$=\rho_{d,d'}$
d
d–1
d+1
1
D

Fig. 1. The time-varying frequency array response for each source is modeled by an HMM. The DOA indices $d = 1, \dots, D$ correspond to the hidden states and the time sequence of frequency array responses $\boldsymbol{a}_{i,k,1}, \dots, \boldsymbol{a}_{i,k,L}$ corresponds to the output sequence of the HMM, respectively.

where $\boldsymbol{\rho}_d = (\rho_{d,1}, \dots, \rho_{d,D})$ denotes the transition probability of state d to each state $1, \dots, D$, and $\boldsymbol{\rho} = (\rho_{d,d'})_{D\times D}$ denotes the transition matrix. Note that we can straightforwardly constrain the DOAs to be continuously time-varying by setting the transition probabilities from state d to states d, $d-1$ and $d+1$, respectively, at reasonably large values. Overall, our new generative model is given by Eqs. (8), (9) and (10).

4. APPROXIMATE POSTERIOR INFERENCE

4.1. Variational Bayesian Approach

In this section, we describe an approximate posterior inference algorithm for our generative model based on variational inference. The random variables of interest in our model are $\boldsymbol{A} = \boldsymbol{a}_{1:I,1:K,1:L}$, $\boldsymbol{S} = s_{1:K,1:L}$, $\boldsymbol{Z} = Z_{1:K,1:L}$ and $\boldsymbol{C} = c_{1:I,1:L}$. We denote the entire set of the above parameters as Θ. In the following, $\boldsymbol{\rho}$, $\Sigma_{1:K}^{(n)}$ and $\Sigma_{1:K}^{(a)}$ are constants that are determined experimentally. Our goal is to compute the posterior

$$p(\Theta|\boldsymbol{Y}) = \frac{p(\boldsymbol{Y},\Theta)}{p(\boldsymbol{Y})}, \tag{11}$$

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 and 3, we can write the joint distribution $p(\boldsymbol{Y},\Theta)$ as

$$p(\boldsymbol{Y},\boldsymbol{A},\boldsymbol{S},\boldsymbol{Z},\boldsymbol{C}) = p(\boldsymbol{Y}|\boldsymbol{A},\boldsymbol{S},\boldsymbol{Z})p(\boldsymbol{Z})p(\boldsymbol{A}|\boldsymbol{C})p(\boldsymbol{C}), \tag{12}$$

but to obtain the exact posterior $p(\Theta|\boldsymbol{Y})$, we must compute $p(\boldsymbol{Y})$, which involves many intractable integrals.

We can express this posterior variationally as the solution to an optimization problem:

$$\underset{q\in Q}{\operatorname{argmin}}\, \mathrm{KL}(q(\Theta)||p(\Theta|\boldsymbol{Y})), \tag{13}$$

where $\mathrm{KL}(\cdot||\cdot)$ denotes the Kullback-Leibler (KL) divergence between its two arguments, i.e.,

$$\mathrm{KL}(q(\Theta)||p(\Theta|\boldsymbol{Y})) = \int q(\Theta)\log\frac{q(\Theta)}{p(\Theta|\boldsymbol{Y})}d\Theta. \tag{14}$$

Indeed, if we let Q be the family of all distributions over Θ, the solution to the optimization problem is the exact posterior $p(\Theta|\boldsymbol{Y})$, since KL divergence is minimized when its two arguments are exactly equal. Of course, solving this optimization problem is just as intractable as directly computing the posterior. Although it may appear that no progress has been made, restricting $q(\Theta)$ to belong to a family of distributions with a simpler form than $p(\Theta|\boldsymbol{Y})$ allows us to obtain principled approximate solutions.

For our model, we define the set of approximate distributions Q as those that factor as follows:

$$Q = \{q : q(\boldsymbol{A})q(\boldsymbol{S})q(\boldsymbol{Z})q(\boldsymbol{C})\}. \tag{15}$$

This approximation is often called a naive mean-field approximation.

4.2. Coordinate Ascent

We now present an algorithm for solving the optimization problem described in (12) and (14). Unfortunately, the optimization problem is non-convex, and it is difficult to find the global optimum. However, we can use a simple coordinate ascent algorithm to find a local optimum. Note that (13) can be written as

$$\mathrm{KL}(q(\Theta)||p(\Theta|\boldsymbol{Y})) = \int q(\Theta)\log\frac{q(\Theta)}{p(\boldsymbol{Y},\Theta)}d\Theta + \log p(\boldsymbol{Y}). \tag{16}$$

As the log evidence $\log p(\boldsymbol{Y})$ is fixed with respect to $q(\Theta)$, minimizing the first term, which is known as the (negative) variational free energy, amounts to minimizing the KL divergence of $p(\Theta|\boldsymbol{Y})$ from $q(\Theta)$. In the mean-field approximation of the posterior, the algorithm can optimize one factor at a time while fixing all other factors. It can be shown using the calculus of variations that the "optimal" distribution for each of the factors can be expressed as:

$$\hat{q}(\boldsymbol{X}) \propto \exp \boldsymbol{E}_{\Theta\backslash\boldsymbol{X}}[\log p(\boldsymbol{Y},\Theta)], \tag{17}$$

where $\boldsymbol{X}$ indicates one of the factors and $\boldsymbol{E}_{\Theta\backslash\boldsymbol{X}}[\log p(\boldsymbol{Y},\Theta)]$ is the expectation of the joint probability of the data and latent variables, taken over all variables except $\boldsymbol{X}$. The update equations for the variational distributions are given in the following form:

$$\hat{q}(\boldsymbol{A}) = \prod_{i,k,l}\mathcal{N}_{\mathbb{C}}(\boldsymbol{a}_{i,k,l};\boldsymbol{m}_{i,k,l},\Gamma_{i,k,l}), \tag{18}$$

$$\hat{q}(\boldsymbol{S}) = \prod_{k,l}\mathcal{N}_{\mathbb{C}}(s_{k,l};\mu_{k,l},\sigma_{k,l}^2), \tag{19}$$

$$\hat{q}(\boldsymbol{Z}) = \prod_{k,l}\hat{q}(z_{k,l}), \hat{q}(z_{k,l}=i) = \phi_{i,k,l}, \tag{20}$$

where

$$\Gamma_{i,k,l}^{-1} =(\phi_{i,k,l}(|\mu_{k,l}|^2+\sigma_{k,l}^2))\Sigma_k^{(n)-1}+\Sigma_k^{(a)-1}, \tag{21}$$

$$\begin{aligned}\boldsymbol{m}_{i,k,l} =&\Gamma_{i,k,l}(\Sigma_k^{(n)-1}\phi_{i,k,l}\mu_{k,l}^*\boldsymbol{y}_{k,l}\\ &+\Sigma_k^{(a)-1}\sum_d\hat{q}(c_{i,l}=d)\boldsymbol{h}(\vartheta_d,\omega_k)),\end{aligned} \tag{22}$$

$$\frac{1}{\sigma_{k,l}^2} =\sum_i\phi_{i,k,l}\mathrm{tr}[(\boldsymbol{m}_{i,k,l}\boldsymbol{m}_{i,k,l}^{\mathsf{H}}+\Gamma_{i,k,l})\Sigma_k^{(n)-1}], \tag{23}$$

$$\mu_{k,l} =\sigma_{k,l}^2(\sum_i\phi_{i,k,l}\boldsymbol{m}_{i,k,l}^{\mathsf{H}})\Sigma_k^{(n)-1}\boldsymbol{y}_{k,l}, \tag{24}$$

$$\begin{aligned}\varphi_{i,k,l} =&\exp(2\mathrm{Re}[\mu_{k,l}\boldsymbol{y}_{k,l}^{\mathsf{H}}\Sigma_k^{(n)-1}\boldsymbol{m}_{i,k,l}]\\ &-(|\mu_{k,l}|^2+\sigma_{k,l}^2)\mathrm{tr}[(\boldsymbol{m}_{i,k,l}\boldsymbol{m}_{i,k,l}^{\mathsf{H}}+\Gamma_{i,k,l})\Sigma_k^{(n)-1}]),\end{aligned} \tag{25}$$

$$\phi_{i,k,l} =\frac{\varphi_{i,k,l}}{\sum_i\varphi_{i,k,l}}. \tag{26}$$

For updating $\hat{q}(\boldsymbol{C})$, we can use the forward-backward algorithm. The update equation of $\hat{q}(\boldsymbol{C})$ is given as:

$$\hat{q}(\boldsymbol{C}) = \prod_{i,l} \frac{\alpha(\vartheta_{c_{i,l}})\beta(\vartheta_{c_{i,l}})}{\sum_{\vartheta_{c_{i,l}}} \alpha(\vartheta_{c_{i,l}})\beta(\vartheta_{c_{i,l}})}, \tag{27}$$

where α and β denote the forward and backward variables, that can be computed using the emission probabilities $\hat{q}(\boldsymbol{a}_{i,k,l}|\vartheta_d)$:

$$\alpha(\vartheta_{c_{i,l}}) = \hat{q}(\boldsymbol{a}_{i,k,l}|\vartheta_{c_{i,l}}) \sum_{\vartheta_{c_{i,l-1}}} \alpha(\vartheta_{c_{i,l-1}})\rho_{c_{i,l-1},c_{i,l}}, \tag{28}$$

$$\beta(\vartheta_{c_{i,l}}) = \sum_{\vartheta_{c_{i,l+1}}} \beta(\vartheta_{c_{i,l+1}})\hat{q}(\boldsymbol{a}_{i,k,l+1}|\vartheta_{c_{i,l+1}})\rho_{c_{i,l},c_{i,l+1}}. \tag{29}$$

In the sense of variational inference, we can obtain "optimal" $\hat{q}(\boldsymbol{a}_{i,k,l}|\vartheta_d)$ as:

$$\begin{aligned}\hat{q}(\boldsymbol{a}_{i,k,l}|\vartheta_d) \propto& \exp \boldsymbol{E}_{\boldsymbol{a}_{i,k,l}}[\log p(\boldsymbol{a}_{i,k,l}|\vartheta_d)] \\ =& \exp(-\mathrm{tr}[(\boldsymbol{m}_{i,k,l}^{\mathsf{H}}\boldsymbol{m}_{i,k,l} + \Gamma_{i,k,l})\Sigma_k^{(a)-1}] \\ &+ 2\mathrm{Re}[\boldsymbol{h}(\vartheta_d,\omega_k)^{\mathsf{H}}\Sigma_k^{(a)-1}\boldsymbol{m}_{i,k,l}] \\ &- \boldsymbol{h}(\vartheta_d,\omega_k)^{\mathsf{H}}\Sigma_k^{(a)-1}\boldsymbol{h}(\vartheta_d,\omega_k)).\end{aligned} \tag{30}$$

Note that $\hat{q}(\boldsymbol{a}_{i,k,l})$ and $p(\boldsymbol{a}_{i,k,l}|\vartheta_d)$ are both expressed as complex normal distributions.

Finally, the STFT components of the i-th separated signal can be obtained by multiplying $\mu_{k,l}$ by $\phi_{i,k,l}$. Since $q(\Theta)$ is an approximation of the true posterior $p(\Theta|\boldsymbol{Y})$, $\phi_{i,k,l}\mu_{k,l}$ corresponds to an approximation of the minimum mean square error estimator of the i-th source signal, i.e., $\hat{s}_{i,k,l} = \mathbb{E}\big[\mathbf{1}[z_{k,l} = i]s_{k,l}|\boldsymbol{Y}\big] \simeq \mathbb{E}[z_{k,l} = i|\boldsymbol{Y}]\mathbb{E}[s_{k,l}|\boldsymbol{Y}] = \phi_{i,k,l}\mu_{k,l}$, where $\mathbf{1}[\cdot]$ denotes the indicator function that takes the value 1 if its argument is true and 0 otherwise.

5. EXPERIMENTAL EVALUATION

We evaluated the performance of the present method in terms of the ability to separate moving sources and track their DOAs. We used ten mixed stereo signals as the experimental data, each of which we obtained by mixing the speech signals of one static female speaker and two moving male speakers. The static source was obtained from the ATR Japanese speech database [10] and was convolved with the measured room impulse response from the RWCP database [11] (in which the distance between the microphones was 2.83 cm and the reverberation time was 0 ms). We selected the two moving sources from ten different moving sources obtained from the RWCP database [11]. 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 16 ms. $\Sigma_k^{(n)}$ and $\Sigma_k^{(a)}$ were set at $\boldsymbol{I}$ and $10^{1.5} \times \boldsymbol{I}$, respectively. D was set at 180. In the experiments, we fixed source signal $s(\omega_k, t_l)$ as $y_1(\omega_k, t_l)$. This may be reasonable since the noise was relatively lower than the speech, and could help prevent $s(\omega_k, t_l)$ from being trapped in local optima. Moreover, our preliminary experiments revealed that $\hat{q}(\boldsymbol{C})$ was likely to be trapped in local optima due to the spatial aliasing that occurs at high frequencies. To avoid this, we adopted the following procedure: we first ran the variational inference algorithm using only the low-frequency region of the observed signals, after which we gradually increased the frequency range to the Nyquist frequency during the iterations. The variational inference algorithm was run for 100 iterations. $\hat{q}(z_{1:K,1:L} = i)$ was initially set equally at $1/3$ for $i = 1, 2, 3$. As for $\hat{q}(c_{1:3,1:L} = d)$, $\hat{q}(c_{1,1:L} = 46)$, $\hat{q}(c_{2,1:L} = 91)$ and $\hat{q}(c_{3,1:L} = 136)$ were set at relatively large values. The estimated DOAs were obtained from $\hat{q}(\boldsymbol{C})$ as the most possible direction at each time. We chose the method proposed in [8] as a comparison. This method assumed that the source directions were fixed. As an evaluation measure, we used the signal-to-interference ratio (SIR) [12]. The SIR is expressed in decibels (dB), and a higher SIR indicates superior quality. The average input SIRs(±SD) dB of moving source A, moving source B and the fixed source were -4.81(±1.15), -5.22(±1.03) and 0.29(±0.78) dB, respectively. The SIRs were calculated until the time at which the shortest signal of the three source signals was finished.

Table 1. The average output SIRs and standard deviations of the three sources by the conventional and proposed methods.

SIR(±SD) [dB]	moving source A	moving source B	fixed source
Proposed	4.82(±3.94)	6.07(±3.24)	8.16(±1.50)
Conventional	-1.10(±1.37)	-2.00(±1.70)	3.55(±0.78)

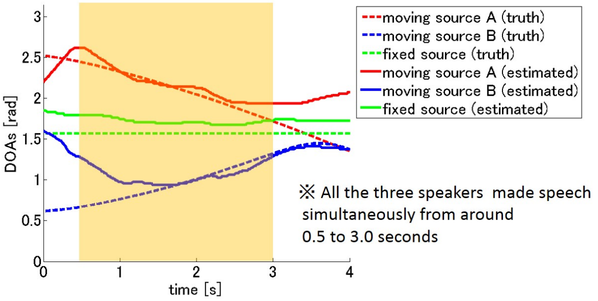

Fig. 2. Example of the ground truth DOA trajectories (dashed lines) and the estimated DOA trajectories (solid lines).

Table 1 shows the average SIRs and standard deviations for the ten mixed signals obtained by the conventional and proposed methods. The average SIRs of the proposed method were superior to those of the conventional method for each signal, especially for moving signals. The total average of the SIRs obtained with the proposed method was 6.20 dB more than that obtained with the conventional approach. These results show the effectiveness of the proposed method for BSS of moving sources. Examples of separated signals are available at `http://www.hil.t.u-tokyo.ac.jp/~higuchi/demo/Examples.htm`. In this experiment, the DOAs of different sources hardly overlapped each other. Another experiment revealed that if the DOAs of different sources overlapped each other, the proposed method did not work well because it separates each signal based on its DOA. Fig. 2 shows a DOA estimation result. The DOAs were almost all estimated correctly from 0.5 to 3.0 seconds, when all of the speakers were speaking.

6. CONCLUSION

This paper proposed a novel BSS approach that simultaneously estimates the directions of moving sources, separates the sources based on sparseness of speech and performs permutation alignment. Focusing on the fact that the DOAs of a source tend to change gradually in practical situations, we modeled a time-varying frequency array response for each source as a path-restricted HMM. Each hidden state of the HMM represented the DOA of each source and we integrated the assumption of the smoothness of the DOA into the transition probabilities of the HMM. The experiment showed that the proposed algorithm provided a 6.20 dB improvement compared with the conventional algorithm as regards the signal-to-interference ratio and estimated the DOAs of sources when two of the three sources were moving.

7. REFERENCES

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

[2] Ö. Yılmaz & S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, 52(7), pp. 1830–1847, 2004.

[3] Y. Mori, H. Saruwatari, T. Takatani, S. Ukai, K. Shikano, T. Hiekata, and T. Morita, "Real-time implementation of two-stage blind source separation combining SIMO-ICA and binary masking," in *Proc. 9th International Workshop on Acoustic Echo and Noise Control (IWAENC 2005)*, pp. 229–232, 2005.

[4] M. I. Mandel, D. P. W. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," in *Advances in Neural Information Processing Systems*, 2006, pp. 953–960.

[5] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, 87(8), pp. 1833–1847, 2007.

[6] Y. Izumi, N. Ono and S. Sagayama, "Sparseness-based 2ch BSS using the EM algorithm in reverberant environment," in *Proc. 2007 IEEE Workshop on Applications of Signal Processing (WASPAA 2007)*, pp. 147–150, 2007.

[7] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516–527, 2010.

[8] H. Kameoka, M. Sato, T. Ono, N. Ono and S. Sagayama, "Bayesian nonparametric approach to blind separation of infinitely many sparse sources," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer sciences*, vol. E96-A, no. 10, pp. 1928–1937, 2013.

[9] T. Otsuka, K. Ishiguro, H. Sawada, and H. G. Okuno,, K. Ishiguro, H. Sawada, and H. G. Okuno, "Bayesian unification of sound source localization and separation with permutation resolution," in *Proc. of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI-12)*, pp. 2038–2045, 2012.

[10] 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.

[11] 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 *Proc. 2nd International Conference on Language Resources & Evaluation (LREC 2000)*, pp. 965–968, 2000.

[12] 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.