IEICE TRANS. FUNDAMENTALS, VOL.E96–A, NO.10 OCTOBER 2013

PAPER Special Section on Sparsity-aware Signal Processing

Bayesian Nonparametric Approach to Blind Separation of Infinitely Many Sparse Sources

Hirokazu KAMEOKA[†,††a)], ***Member***, **Misa SATO**[†††*], **Takuma ONO**[†], **Nobutaka ONO**[††††], ***and*** **Shigeki SAGAYAMA**[†], ***Nonmembers***

SUMMARY This paper deals with the problem of underdetermined blind source separation (BSS) where the number of sources is unknown. We propose a BSS approach that simultaneously estimates the number of sources, separates the sources based on the sparseness of speech, estimates the direction of arrival of each source, and performs permutation alignment. We confirmed experimentally that reasonably good separation was obtained with the present method without specifying the number of sources.

key words: *underdetermined blind signal separation, speech, sparseness, permutation alignment, Bayesian nonparametrics, direction of arrival, Dirichlet process, stick-breaking construction, variational inference*

1. Introduction

Blind Source Separation (BSS) is a technique for separating out individual source signals from microphone inputs when the transfer characteristics between sources and microphones are unknown. BSS is potentially useful for the development of such applications as hands-free teleconference systems and automatic meeting transcription systems. In meeting situations, for instance, it is likely that the number of participants (speakers) will change during the meeting or a loud, unexpected noise such as a door slamming will occur in the room. Thus, it is often difficult to pre-specify the exact number of all possible sources present in real environments. Many conventional BSS algorithms are designed to use the number of sources as the input when performing separation, and most of these algorithms do not work well if the assumed and actual numbers of sources are not the same. This paper presents a novel BSS algorithm that allows the number of sources to be inferred along with the separation.

To estimate the unknown mixing matrix and source signals solely from observed signals, we must 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], which estimates the separation matrix (the inverse of the mixing matrix) such that the independence of the source estimates is maximized. However, to achieve a BSS algorithm that works without assuming the number of sources, we must always consider an underdetermined case where there are fewer observations than sources. In an underdetermined case, there are an infinite number of solutions for source signals even if the mixing process is known. The independence assumption is too weak to allow us to determine a unique solution and so directly applying ICA will not work in this case. In underdetermined situations, we typically need a stronger assumption than independence.

One successful approach for underdetermined BSS involves utilizing the sparse nature of speech [2]–[7], which relies on the fact that the time-frequency components of speech are near zero across most of the time-frequency grid points. Since the time-frequency components of speech rarely overlap even when multiple speakers are speaking simultaneously, 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.

The signals observed at each microphone can be modeled as a convolutive mixture of source signals. To exploit the sparse nature of speech, we must convert it to a time-frequency representation. If we assume the use of a short-time Fourier transform (STFT) to obtain a time-frequency representation with a frame length sufficiently longer than the length of the impulse response from a source to a microphone, an observed signal can be approximated fairly well by an instantaneous mixture in the frequency domain. BSS based on this observation model is called frequency domain BSS. While frequency domain BSS allows for a fast implementation compared with BSS that uses a time domain convolutive mixture model, it 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. Some methods such as [7] are designed to perform frequency-bin-wise source separation followed by permutation alignment. However, permutation aligment and source separation problems should be solved in a cooperative manner since the clues used for permutation alignment can also be helpful

Manuscript received January 22, 2013.
Manuscript revised May 23, 2013.

[†]The authors are with the Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, 113-8656 Japan.
[††]The author is with NTT Communication Science Laboratories, NTT Corporation, Atsugi-shi, 243-0198 Japan.
[†††]The author is with School of Engineering, The University of Tokyo, Tokyo, 113-8656 Japan.
[††††]The author is with National Institute of Informatics, Tokyo, 101-8430 Japan.
[*]Presently, with Graduate School of Engineering, The University of Tokyo.
a) E-mail: kameoka@hil.t.u-tokyo.ac.jp
DOI: 10.1587/transfun.E96.A.1928

for source separation. Thus, we consider it important to develop a method that can simultaneously perform frequency-bin-wise source separation and permutation alignment.

Motivated by the above, this paper proposes a novel BSS approach that simultaneously performs (1) an estimation of the number of sources, (2) source separation based on the sparseness of speech, and (3) permutation alignment.

2. Mixing Model

We first consider a situation where N source signals are captured 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. ω_k is the angular frequency corresponding to the k-th frequency bin and t_l is the time of the l-th frame, respectively. If we assume that the length of the impulse response from a source to a microphone is sufficiently shorter than the frame length of the STFT, the observed signal can be approximated fairly well by an instantaneous mixture in the frequency domain:

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

where $\boldsymbol{y}(\omega_k, t_l) = (y_1(\omega_k, t_l), \ldots, y_M(\omega_k, t_l))^{\mathsf{T}}$. $\boldsymbol{a}_i(\omega_k) = (a_{i,1}(\omega_k), \ldots, a_{i,M}(\omega_k))^{\mathsf{T}}$ is the frequency array response for source i, which is assumed to be time-invariant. $\boldsymbol{n}(\omega_k, t_l)$ is assumed to comprise all kinds of components that cannot be expressed by the instantaneous mixture representation (e.g., background noise and reverberant components).

We now exploit the sparseness of speech and assume that only one source is active at each time-frequency point, as with [2]–[7]. By using $z_{k,l} \in \{1, \ldots, N\}$ to denote the (unknown) active source index at time-frequency point (ω_k, t_l), (1) can be rewritten as

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

Notice that the superscript i is dropped from $s_i(\omega_k, t_l)$ in (2) as it is no longer necessary since we are assuming $s_i(\omega_k, t_l) = 0$ for $i \neq z_{k,l}$. Namely, $s(\omega_k, t_l)$ signifies the component of an active source at time-frequency point (ω_k, t_l). For convenience of notation, we hereafter use subscripts k and l to indicate ω_k and t_l.

3. Generative Model

3.1 Generative Process of Observed Signals

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

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

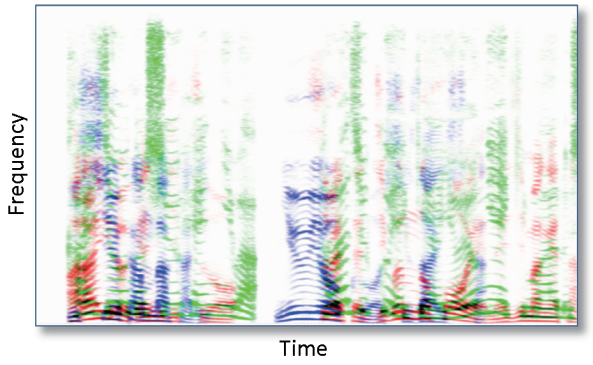

Fig. 1 Spectrograms of speech signals uttered by three speakers, each of which is indicated by a different color. Typically, the time-frequency components of speech rarely overlap even when multiple speakers are speaking simultaneously.

conditioned on $\boldsymbol{a}_{1:N,k} = \{\boldsymbol{a}_{1,k}, \ldots, \boldsymbol{a}_{N,k}\}$, $s_{k,l}$ and $z_{k,l}$, where $\mathcal{N}_{\mathbb{C}}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) \propto \exp(-(\boldsymbol{x} - \boldsymbol{\mu})^{\mathsf{H}} \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}))$. Using this model, Izumi et al. formulated an Expectation-Maximization (EM) algorithm for the maximum likelihood estimation of the unknown parameters by treating $z_{k,l}$ as a latent variable [6], [7]. To flexibly incorporate hierarchical prior distributions over the unknown parameters as well as the number of sources, we employ a Bayesian approach.

3.2 Mixture of Infinitely Many Sparse Sources

We do not usually obtain any information about which source is active at each time-frequency point. Thus, we regard $z_{k,l}$ as a latent variable and consider describing its generative process. If the number of sources is N, it would be natural to assume that the probability of choosing an index $z_{k,l}$ from the set of source indices, $\{1, \ldots, N\}$, can be described as a categorical distribution

$$z_{k,l} | \boldsymbol{\pi} \sim \text{Categorical}(z_{k,l}; \boldsymbol{\pi}), \tag{4}$$

where $\text{Categorical}(x; \boldsymbol{y}) = y_x$ (with $\boldsymbol{y} = (y_1, \ldots, y_N)$ and $\sum_{i=1}^{N} y_i = 1$). The i-th element of $\boldsymbol{\pi}$ defines how likely the source index i is to be chosen. Since we also have no information about $\boldsymbol{\pi}$, we consider describing its generative process using a "symmetric" distribution. For the convenience of the following analysis, we assume that $\boldsymbol{\pi}$ has been generated from a symmetric Dirichlet distribution

$$\boldsymbol{\pi} \sim \text{Dirichlet}(\boldsymbol{\pi}; \alpha_0/N, \ldots, \alpha_0/N), \tag{5}$$

where $\text{Dirichlet}(\boldsymbol{y}; z_1, \ldots, z_N) \propto \prod_{i=1}^{N} y_i^{z_i - 1}$. The shape of the Dirichlet distribution is governed by a concentration hyperparameter α_0.

Thus far, we have considered the case of a finite number N of sources. It can be shown that with a size-biased permutation of (5) followed by taking the limit $N \to \infty$, the above generative processes (3), (4) and (5) approach

$$\boldsymbol{\pi} \sim \text{GEM}(\boldsymbol{\pi}; \alpha_0), \tag{6}$$

$$z_{k,l} | \boldsymbol{\pi} \sim \text{Categorical}(z_{k,l}; \boldsymbol{\pi}), \tag{7}$$

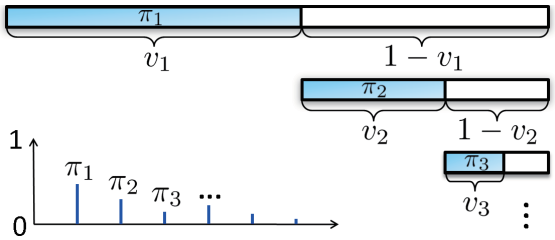

Fig. 2 Illustration of the stick-breaking construction. The construction of $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_\infty)$ can be understood metaphorically as follows. Starting with a stick of length 1, we break it at v_1, assigning π_1 as the length of the stick we just broke off. We now recursively break the other portion to obtain π_2, π_3 and so forth.

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

$\mathrm{GEM}(\boldsymbol{\pi}; \alpha_0)$ is called the stick-breaking process [8], where the letters stand for Griffiths, Engen and McCloskey. For $i = 1, 2, \ldots \infty$, the i-th mixture weight π_i is generated via

$$v_i \sim \mathrm{Beta}(v_i; 1, \alpha_0), \tag{9}$$

$$\pi_i = v_i \prod_{j=1}^{i-1}(1 - v_j), \tag{10}$$

where $\mathrm{Beta}(y; z_1, z_2) \propto y^{z_1-1}(1-y)^{z_2-1}$. This is known as a constructive definition of the Dirichlet process [9], which can be thought of as an infinite dimensional Dirichlet distribution. The construction of $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_\infty)$ can be understood metaphorically as follows. Starting with a stick of length 1, we break it at v_1, assigning π_1 as the length of the piece of stick that we have just broken off. Now recursively break the other portion to obtain π_2, π_3 and so forth (see Fig. 2). $\boldsymbol{\pi} \sim \mathrm{GEM}(\boldsymbol{\pi}; \alpha_0)$ thus produces exponentially decaying weights in expectation. This means that the source with a larger index will be less likely to be active and thus simple models with fewer sources are favored, given observed signals.

3.3 Mixture of Direction-of-Arrivals (DOAs)

Here we describe the generative process of the frequency response $\boldsymbol{a}_{i,k}$ of the mixing system.

So 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 thus 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 interchannel time difference between the microphones depends only on the direction of arrival (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 of the source. Specifically, with $M = 2$ microphones, the complex array response for a source at direction θ such that $0 \le \theta < 2\pi$ is defined as a function of ω depending on θ

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

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 frequency response $\boldsymbol{a}_{i,k}$ should be equal to $\boldsymbol{h}(\theta_i, \omega_k)$. However, due to such factors as the plane wave assumption and the narrowband instantaneous mixture approximation, the actual frequency response $\boldsymbol{a}_{i,k}$ may diverge from the "ideal frequency response" $\boldsymbol{h}(\theta_i, \omega_k)$ to some extent. One way to simplify the process of this kind of deviation is to assume a probability distribution on $\boldsymbol{a}_{i,k}$ with the expected value of $\boldsymbol{h}(\theta_i, \omega_k)$. Here, we assume for convenience that $\boldsymbol{a}_{i,k}$ is generated from a complex normal distribution with mean $\boldsymbol{h}(\theta_i, \omega_k)$. Note that we do not usually obtain any information about the direction from which each source is emanating. Thus, we regard the DOA of each source as a latent variable and further consider describing its generative process. As explained in detail below, the entire generative process of $\boldsymbol{a}_{i,k}$ can then be described as a "mixture of DOAs".

Let us firstly introduce a discrete set of D possible directions, $\vartheta_1, \ldots, \vartheta_D$, which are all assumed to be constants. For instance, consider defining ϑ_d as $\vartheta_d = (d-1)\pi/D$, $(d = 1, \ldots, D)$, which means dividing π into D equal angles. We then assume that each source is emanating 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 $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_D)$

$$c_i|\boldsymbol{\rho} \sim \mathrm{Categorical}(c_i; \boldsymbol{\rho}). \tag{12}$$

By using c_i, θ_i is then given as

$$\theta_i = \vartheta_{c_i}. \tag{13}$$

As with 3.2, we assume that $\boldsymbol{\rho}$ has been generated from a symmetric Dirichlet distribution

$$\boldsymbol{\rho} \sim \mathrm{Dirichlet}(\boldsymbol{\rho}; \beta_0/D, \ldots, \beta_0/D). \tag{14}$$

As mentioned above, the frequency response $\boldsymbol{a}_{i,k}$ is assumed to be generated from a complex normal distribution with mean $\boldsymbol{h}(\vartheta_{c_i}, \omega_k)$, given c_i,

$$\boldsymbol{a}_{i,k}|c_i \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{a}_{i,k}; \boldsymbol{h}(\vartheta_{c_i}, \omega_k), \Sigma_k^{(a)}), \tag{15}$$

where $\Sigma_k^{(a)}$ denotes the covariance of the complex normal distribution, which is assumed to be a constant. An illustration of the generative process of $\boldsymbol{a}_{i,k}$ described above is shown in Fig. 3.

Overall, the entire generative model is described in plate notation in Fig. 4.

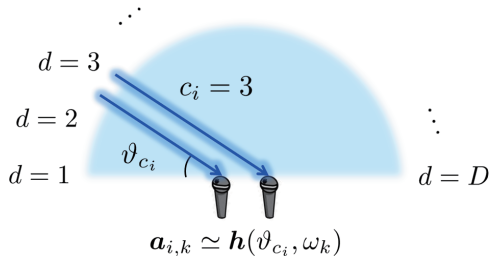

Fig. 3 Illustration of the generative process of the complex array response $\boldsymbol{a}_{i,k}$ based on the DOA mixture model.

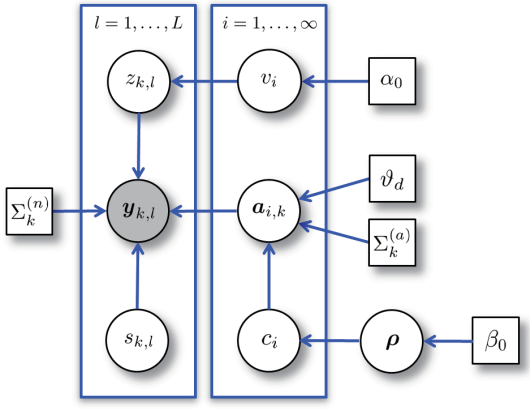

Fig. 4 Plate notation of present overall generative model.

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 [10]. The random variables of interest in our model are

$$\begin{aligned}
A &= \{\boldsymbol{a}_{i,k}\}_{i,k} &&: \text{complex array response for source } i,\\
S &= \{s_{k,l}\}_{k,l} &&: \text{component of active source at } (\omega_k, t_l),\\
Z &= \{z_{k,l}\}_{k,l} &&: \text{index of active source at } (\omega_k, t_l),\\
V &= \{v_i\}_i &&: \text{stick breaking proportion},\\
C &= \{c_i\}_i &&: \text{index of direction for source } i,\\
\boldsymbol{\rho} &= (\rho_1, \ldots, \rho_D) &&: \text{mixture weight for each DOA.}
\end{aligned}$$

We denote the entire set of the above parameters as Θ. Our goal is to compute the posterior $p(\Theta|Y)$ where $Y = \{\boldsymbol{y}_{k,l}\}_{k,l}$ is a set consisting of the time-frequency components of observed multichannel signals. By using the conditional distributions defined in 3.2 and 3.3, we can write the joint distribution $p(Y, \Theta)$ as

$$\begin{aligned}
&p(Y, A, S, Z, V, C, \boldsymbol{\rho})\\
&\quad = p(Y|A, S, Z)p(Z|V)p(V)p(A|C)p(C|\boldsymbol{\rho})p(\boldsymbol{\rho}), \quad (16)
\end{aligned}$$

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

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

$$\operatorname*{argmin}_{q \in \mathcal{Q}} \mathrm{KL}(q(\Theta)\|p(\Theta|Y)), \quad (17)$$

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

$$\mathrm{KL}(q(\Theta)\|p(\Theta|Y)) = \int q(\Theta) \log \frac{q(\Theta)}{p(\Theta|Y)} \mathrm{d}\Theta. \quad (18)$$

Indeed, if we let $\mathcal{Q}$ be the family of all distributions over Θ, the solution to the optimization problem is the exact posterior $p(\Theta|Y)$, 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 of simpler form than $p(\Theta|Y)$ allows us to obtain principled approximate solutions.

For our model, we define the set of approximate distributions $\mathcal{Q}$ as those that factor as follows:

$$\mathcal{Q} = \{q : q(A)q(S)q(Z)q(V)q(C)q(\boldsymbol{\rho})\}. \quad (19)$$

This approximation is often called a naive mean-field approximation. To define $q(A)$, $q(V)$ and $q(C)$, we need to construct distributions on the infinite sets $\{v_1, v_2, \ldots\}$, $\{\boldsymbol{a}_{1,k}, \boldsymbol{a}_{2,k}, \ldots\}$ and $\{c_1, c_2, \ldots\}$. For this approach to be tractable, we truncate the variational distribution at some value N^* by setting $q(v_{N^*} = 1) = 1$. The mixture proportions π_i for $i > N^*$ will thus be zero, and we can ignore $\boldsymbol{a}_{i,k}$ and c_i for $i > N^*$. In practice, we set N^* at a sufficiently large integer. It is important to emphasize that truncating the variational distribution does not mean that the true posterior itself is truncated. As the truncation level N^* becomes larger, the approximations to the true posterior become more accurate.

4.2 Coordinate Ascent

We now present an algorithm for solving the optimization problem described in (17) and (19). Unfortunately, the optimization problem is non-convex, and it is intractable to find the global optimum. However, we can use a simple coordinate ascent algorithm to find a local optimum. Notice that (18) can be written as

$$\begin{aligned}
&\mathrm{KL}(q(\Theta)\|p(\Theta|Y))\\
&\qquad = \int q(\Theta) \log \frac{q(\Theta)}{p(Y, \Theta)} \mathrm{d}\Theta + \log p(Y). \quad (20)
\end{aligned}$$

As the log evidence $\log p(Y)$ is fixed with respect to $q(\Theta)$, minimizing the first term, which is known as the (negative) variational free energy, minimizes the KL divergence of $p(\Theta|Y)$ from $q(\Theta)$. The algorithm optimizes one factor in the mean-field approximation of the posterior 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:

$$\hat{q}(X) \propto \exp \mathbb{E}_{\Theta \setminus X}[\log p(Y, \Theta)], \quad (21)$$

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

$$\hat{q}(A) \leftarrow \prod_{i,k} \mathcal{N}_{\mathbb{C}}(\boldsymbol{a}_{i,k}; \boldsymbol{m}_{i,k}, \Gamma_{i,k}), \tag{22}$$

$$\hat{q}(S) \leftarrow \prod_{k,l} \mathcal{N}_{\mathbb{C}}(s_{k,l}; \mu_{k,l}, \sigma^2_{k,l}), \tag{23}$$

$$\hat{q}(Z) \leftarrow \prod_{k,l} \mathrm{Categorical}(z_{k,l}; \boldsymbol{\phi}_{k,l}), \tag{24}$$

$$\hat{q}(V) \leftarrow \prod_{i} \mathrm{Beta}(v_i; \gamma_{i,0}, \gamma_{i,1}), \tag{25}$$

$$\hat{q}(C) \leftarrow \prod_{i} \mathrm{Categorical}(c_i; \boldsymbol{\psi}_i), \tag{26}$$

$$\hat{q}(\boldsymbol{\rho}) \leftarrow \mathrm{Dirichlet}(\boldsymbol{\rho}; \zeta_1, \ldots, \zeta_D), \tag{27}$$

where

$$\Gamma^{-1}_{i,k} = \Big(\sum_l [\boldsymbol{\phi}_{k,l}]_i (|\mu_{k,l}|^2 + \sigma^2_{k,l}) \Big) \Sigma^{(n)-1}_k + \Sigma^{(a)-1}_k, \tag{28}$$

$$\boldsymbol{m}_{i,k} = \Gamma_{i,k} \Big(\Sigma^{(n)-1}_k \sum_l [\boldsymbol{\phi}_{k,l}]_i \mu^*_{k,l} \boldsymbol{y}_{k,l} + \Sigma^{(a)-1}_k \sum_d [\boldsymbol{\psi}_i]_d \boldsymbol{h}(\vartheta_d, \omega_k) \Big), \tag{29}$$

$$\frac{1}{\sigma^2_{k,l}} = \sum_i [\boldsymbol{\phi}_{k,l}]_i \mathrm{tr}\left[(\boldsymbol{m}_{i,k} \boldsymbol{m}^{\mathsf{H}}_{i,k} + \Gamma_{i,k}) \Sigma^{(n)-1}_k \right], \tag{30}$$

$$\mu_{k,l} = \sigma^2_{k,l} \Big(\sum_i [\boldsymbol{\phi}_{k,l}]_i \boldsymbol{m}^{\mathsf{H}}_{i,k} \Big) \Sigma^{(n)-1}_k \boldsymbol{y}_{k,l}, \tag{31}$$

$$\begin{aligned} \varphi_{i,k,l} = \exp\Big[& 2\mathrm{Re}(\mu_{k,l} \boldsymbol{y}^{\mathsf{H}}_{k,l} \Sigma^{(n)-1}_k \boldsymbol{m}_{i,k}) \\ & - (|\mu_{k,l}|^2 + \sigma^2_{k,l}) \mathrm{tr}\Big[(\boldsymbol{m}_{i,k} \boldsymbol{m}^{\mathsf{H}}_{i,k} + \Gamma_{i,k}) \Sigma^{(n)-1}_k \Big] \\ & + \Psi(\gamma_{i,0}) - \Psi(\gamma_{i,0} + \gamma_{i,1}) \\ & + \sum_{j=1}^{i-1} (\Psi(\gamma_{j,1}) - \Psi(\gamma_{j,0} + \gamma_{j,1})) \Big], \end{aligned}$$

$$[\boldsymbol{\phi}_{k,l}]_i = \frac{\varphi_{i,k,l}}{\sum_{i'=1}^{N^*} \varphi_{i',k,l}}, \tag{32}$$

$$\gamma_{i,0} = 1 + \sum_{k,l} [\boldsymbol{\phi}_{k,l}]_i, \tag{33}$$

$$\gamma_{i,1} = \alpha_0 + \sum_{j=i+1}^{N^*} \sum_{k,l} [\boldsymbol{\phi}_{k,l}]_i, \tag{34}$$

$$\begin{aligned} \varrho_{i,d} = \exp\Big[& \sum_k \Big\{ 2\mathrm{Re}(\boldsymbol{h}(\vartheta_d, \omega_k)^{\mathsf{H}} \Sigma^{(a)-1}_k \boldsymbol{m}_{i,k}) \\ & - \mathrm{tr}[\boldsymbol{h}(\vartheta_d, \omega_k) \boldsymbol{h}(\vartheta_d, \omega_k)^{\mathsf{H}} \Sigma^{(a)-1}_k] \Big\} + \Psi(\zeta_d) \Big], \end{aligned}$$

$$[\boldsymbol{\psi}_i]_d = \frac{\varrho_{i,d}}{\sum_{d'=1}^{D} \varrho_{i,d'}}, \tag{35}$$

$$\zeta_d = \frac{\beta_0}{D} + \sum_{i=1}^{N^*} \psi_{i,d}. \tag{36}$$

$[\cdot]_i$ denotes the i-th element of a vector, $\mathrm{tr}[\cdot]$ denotes the trace of a matrix, and $\Psi(\cdot)$ denotes the digamma function.

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

5. Related Work

It has been brought to our attention that several papers related to this work have been presented independently by different authors [11], [12], after the publication of our conference paper [17].

Taghia et al. proposed in [11] a variational Bayesian approach for the approximate inference of the parameters of Sawada's model [7]. This method requires a model selection procedure to estimate the number of sources. The variational lower bound (an approximation of the log marginal likelihood) is used as the criterion for the model selection. By contrast, our method avoids selecting the number of sources explicitly and instead pushes this task down into the inference algorithm, incorporating automated Bayesian model selection in the inference procedure.

Otsuka et al. proposed in [12] a combined source separation and source localization method based on a Bayesian approach. Similar to the idea described in 3.3, they described the prior distribution of the spatial covariance matrix of each source using a mixture model of Wishart-distributed spatial covariance matrices, each of which is associated with a latent DOA variable. By contrast, the present model describes the prior distribution of the frequency array response of each source using a mixture model of Gaussian-distributed frequency array responses, each of which is associated with a latent DOA variable. Although these models appear slightly different, the concepts are very similar. It would be interesting to compare the different ways in which the use of these models affect the source separation ability. This should be investigated in the future.

6. Experiment

We evaluated the performance of the proposed method in terms of source separation ability.

We used stereo speech signals with a sampling rate of 16 kHz as test signals, which we obtained by mixing three speech signals [13] (one male and two female speakers) using a measured room impulse response [14] (in which the distance between the microphones was 5 cm and the reverberation time was 0 ms). The three sources were spaced 30 degrees apart. The geometry setting for the test mixtures

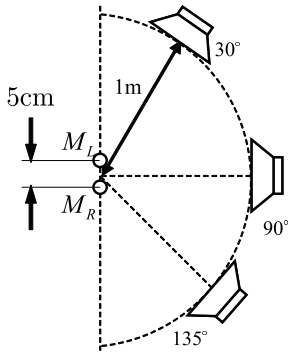

Fig. 5 Geometry setting for text mixtures.

is shown in Fig. 5. 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.

Our preliminary experiments revealed that ζ_d 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 up to the Nyquist frequency during the iteration. $\Sigma_k^{(n)}$ and $\Sigma_k^{(a)}$ were set respectively at $\boldsymbol{I}$ and $10^{-1.5} \times \boldsymbol{I}$. D was set at 180. All the variational parameters were initialized randomly. After convergence, each separated signal was obtained by multiplying $\mu_{k,l}$ by $\phi_{i,k,l}$. Figure 6 shows some examples of the spectrograms of separated source signals obtained with the present method. Owing to the property of the stick-breaking process, sources that are dominant at many time-frequency points are likely to be assigned small indices. Thus, we considered the separated signals of indices $i = 1, 2, 3$ to be the source estimates, which we used for the evaluation. We chose Sawada's method described in [7] as a comparison. The source code was kindly provided to us by its author. In this method, the number of sources must be specified manually. The following results report the performance in terms of the Signal-to-Distortion Ratio (SDR) [15]. The SDR is expressed in decibels (dB), and a higher SDR indicates superior quality.

The present method was tested with various settings of the truncation level N^*. As for Sawada's method, it was tested with various settings of the assumed number N of sources. Figure 7 shows the average SDRs obtained with Sawada's and the present methods with various N and N^* settings. As expected, the performance of the present method improves with increasing N^*, while that of Sawada's method deteriorates significantly when the assumed number of sources departs from the actual number. It is worth emphasizing that Sawada's method requires prior knowledge of the exact number of sources to achieve good separation, while the present approach does not. With Sawada's method, each source signal tends to be further separated into multiple components when the specified number of sources is greater than the actual number. Unlike with Sawada's

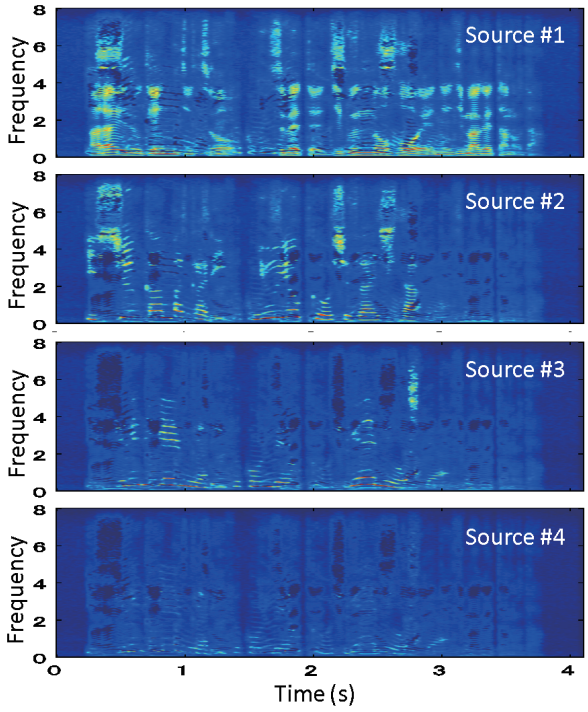

Fig. 6 Separated source signals obtained with present method.

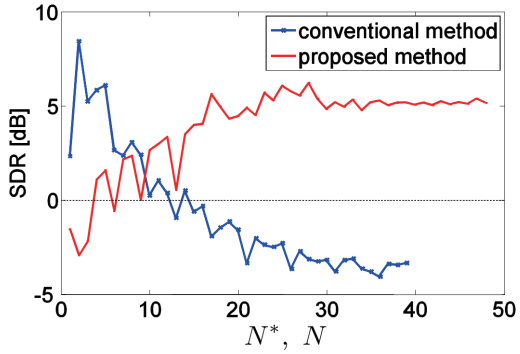

Fig. 7 SDRs obtained with conventional and present methods with different N and N^* settings.

method, such a phenomenon did not occur with the present method. This is because the present method searches for an explanation for an observed signal using a model with as low a complexity as possible.

Figure 8 shows an example of the estimate of the inter-channel phase difference of each source. The slope of $\arg([\boldsymbol{m}_{i,k}]_2/[\boldsymbol{m}_{i,k}]_1)$ along the frequency direction k corresponds to the DOA estimate of source i. Hence, if the DOAs of the sources were estimated correctly and the permutation problem was successfully solved, $\arg([\boldsymbol{m}_{i,k}]_2/[\boldsymbol{m}_{i,k}]_1)$ with the same index i should lie on a straight line along k. Figure 8 thus reveals that the DOA estimation and permutation alignment both worked successfully with the present method.

It is important to note that the DOA mixture model described in 3.3 assumes a completely anechoic environment. Moreover, the assumption of the sparseness of speech

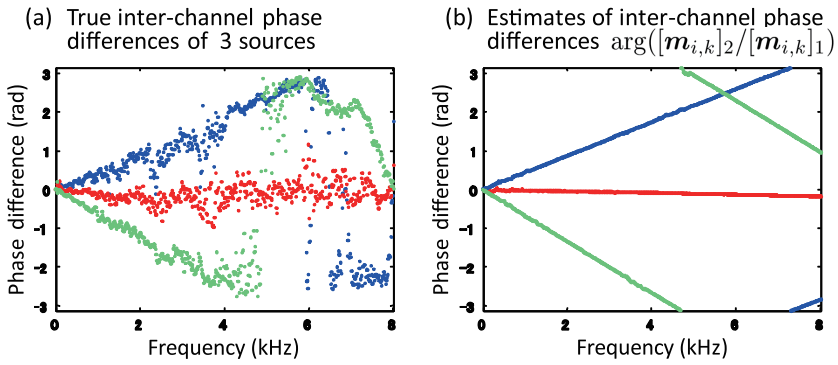

Fig. 8 An example of the estimate of the inter-channel phase difference of each source plotted in a different color. The slope of $\arg([\boldsymbol{m}_{i,k}]_2/[\boldsymbol{m}_{i,k}]_1)$ along the frequency direction k corresponds to the DOA estimate of source i. For each i, $\arg([\boldsymbol{m}_{i,k}]_2/[\boldsymbol{m}_{i,k}]_1)$ lies on a straight line along k. This means that both the DOA estimation and permulation alignment worked successfully with the present method.

described in Sect. 2 does not hold in reverberant environments. Indeed, the performance of the present method was poor when using real data recorded in reverberant environments. This implies the need for a combined generative model of the present model and a reverberation process, e.g., by adopting the idea described in [16].

7. Conclusion

This paper aimed at developing a BSS algorithm that works well even when the number of sources is unknown and proposed a novel BSS approach that simultaneously performs an estimation of the number of sources, source separation based on the sparseness of speech, an estimation of the direction of arrival of each source, and permutation alignment. We confirmed experimentally that reasonably good separations were obtained with the present method without specifying the number of sources.

Blind separation of moving sound sources is an important task to be solved for real applications. One possible extension to the present method involves representing the DOA indicator variable c_i as a time sequence, i.e., $\{c_{i,l}\}_{1\le l\le L}$, and modeling its generative process using a hidden Markov model. This is currently under development.

It should be noted that this paper is an extended journal version of our conference papers [17], [18].

Acknowledgments

We thank Dr. Hiroshi Sawada (NTT) for providing us with the source code of his method. We also thank Mr. Yosuke Izumi (previously with The University of Tokyo) for providing us with Fig. 1.

References

[1] A. Hyvärinen, J. Karhunen, and E. Oja, Independent Component Analysis, John Wiley & Sons, 2001.
[2] Ö. Yılmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," IEEE Trans. Signal Process., vol.52, no.7, pp.1830–1847, 2004.
[3] M.I. Mandel, D.P.W. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," Adv. Neural Info. Process. Syst., pp.953–960, 2006.
[4] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," Signal Process., vol.87, no.8, pp.1833–1847, 2007.
[5] 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," Proc. 9th International Workshop on Acoustic Echo and Noise Control (IWAENC 2005), pp.229–232, 2005.
[6] Y. Izumi, N. Ono, and S. Sagayama, "Sparseness-based 2ch BSS using the EM algorithm in reverberant environment," Proc. 2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (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 Trans. Audio Speech Language Process., vol.19, no.3, pp.516–527, 2010.
[8] J. Sethuraman, "A constructive definition of Dirichlet priors," Statistica Sinica, vol.4, pp.639–650, 1994.
[9] T.S. Ferguson, "A Bayesian analysis of some nonparametric problems," Annals of Statistics, vol.1, no.2, pp.209–230, 1973.
[10] D.M. Blei and M.I. Jordan, "Variational inference for Dirichlet process mixtures," Journal of Bayesian Analysis, vol.1, no.1, pp.121–144, 2006.
[11] J. Taghia, N. Mohammadiha, and A. Leijon, "A variational Bayes approach to the underdetermined blind source separation with automatic determination of the number of sources," Proc. 2012 IEEE International Conferences on Acoustics, Speech and Signal Processing (ICASSP2012), pp.253–256, March 2012.
[12] T. Otsuka, K. Ishiguro, H. Sawada, and H.G. Okuno, "Bayesian unification of sound source localization and separation with permutation resolution," Proc. Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI-12), pp.2038–2045, July 2012.
[13] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," Transactions on Speech Communication, pp.357–363, 1990.
[14] 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," Proc. 2nd International Conference on Language Resources & Evaluation (LREC 2000), pp.965–968, 2000.
[15] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," IEEE Trans. Audio Speech Language Process., pp.1462–1469, 2006.

[16] T. Yoshioka, T. Nakatani, M. Miyoshi, and H.G. Okuno, "Blind Separation and Dereverberation of Speech Mixtures by Joint Optimization," IEEE Trans. Audio Speech Language Process., vol.19, no.1, pp.69–84, 2011.
[17] H. Kameoka, M. Sato, T. Ono, N. Ono, and S. Sagayama, "Bayesian nonparametric approach to underdetermined sparse BSS," Proc. ASJ meeting, 1-1-19, pp.713–716, March 2012 (in Japanese).
[18] H. Kameoka, M. Sato, T. Ono, N. Ono, and S. Sagayama, "Blind separation of infinitely many sparse sources," Proc. 13th International Workshop on Acoustic Signal Enhancement (IWAENC 2012), Sept. 2012.

Appendix A: Probability Density Functions

The circular complex normal distribution with mean $\boldsymbol{\mu}$ and covariance P is defined as

$$\mathcal{N}_{\mathbb{C}}(\boldsymbol{x};\boldsymbol{\mu},P) = \frac{1}{|\pi P|}\exp\Big((\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{H}}P^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\Big) \quad (\mathrm{A}\cdot 1)$$

$$= \exp\Big[-\boldsymbol{x}^{\mathsf{H}}P^{-1}\boldsymbol{x} + 2\mathrm{Re}(\boldsymbol{\mu}^{\mathsf{H}}P^{-1}\boldsymbol{x}) - \boldsymbol{\mu}^{\mathsf{H}}P^{-1}\boldsymbol{\mu} - \log|\pi P|\Big], \quad (\mathrm{A}\cdot 2)$$

and so

$$\mathbb{E}[\boldsymbol{x}] = \boldsymbol{\mu}, \quad (\mathrm{A}\cdot 3)$$

$$\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathsf{H}}] = P + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{H}}, \quad (\mathrm{A}\cdot 4)$$

where $\mathbb{E}[\cdot]$ denotes expectation.

The Dirichlet distribution of order I with parameters $\alpha_1,\ldots,\alpha_I > 0$ is defined as

$$\mathrm{Dirichlet}(\boldsymbol{x};\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{I}\Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{I}\alpha_i)}\prod_{i=1}^{I}x_i^{\alpha_i-1}, \quad (\mathrm{A}\cdot 5)$$

for all $x_1,\ldots,x_I > 0$ satisfying $\sum_{i=1}^{I}x_i = 1$. The density is zero outside this open $(I-1)$-dimensional simplex. $\Gamma(\cdot)$ is the gamma function. The mean and entropy of the Dirichlet-distributed variable are given by

$$\mathbb{E}[x_i] = \frac{\alpha_i}{\sum_{i'=1}^{I}\alpha_{i'}}, \quad (\mathrm{A}\cdot 6)$$

$$\mathbb{E}[\log x_i] = \Psi(\alpha_i) - \Psi\Big(\sum_i \alpha_i\Big). \quad (\mathrm{A}\cdot 7)$$

Note that when $I = 2$ the Dirichlet distribution reduces to a Beta distribution.

Appendix B: Derivation of Update Equations

Here we show how we obtained the variational update equations. Recall that each variational update equation can be obtained via the formula given by (21).

B.1 Update Equation for $q(A)$

The term in $\log p(Y,\Theta)$ depending on A is given as

$$-\sum_{k,l}(\boldsymbol{y}_{k,l}-\boldsymbol{a}_{z_{k,l},k}s_{k,l})^{\mathsf{H}}\Sigma_k^{(n)-1}(\boldsymbol{y}_{k,l}-\boldsymbol{a}_{z_{k,l},k}s_{k,l})$$
$$-\sum_{i,k}(\boldsymbol{a}_{i,k}-\boldsymbol{h}(\vartheta_{c_i},\omega_k))^{\mathsf{H}}\Sigma_k^{(a)-1}(\boldsymbol{a}_{i,k}-\boldsymbol{h}(\vartheta_{c_i},\omega_k))$$
$$\overset{A}{=}\sum_{i,k}\Big[\sum_l \mathbf{1}[z_{k,l}=i](2\mathrm{Re}[s_{k,l}^*\boldsymbol{a}_{i,k}^{\mathsf{H}}\Sigma_k^{(n)-1}\boldsymbol{y}_{k,l}] - |s_{k,l}|^2\boldsymbol{a}_{i,k}^{\mathsf{H}}\Sigma_k^{(n)-1}\boldsymbol{a}_{i,k}) + \sum_d \mathbf{1}[c_i=d](2\mathrm{Re}[\boldsymbol{a}_{i,k}^{\mathsf{H}}\Sigma_k^{(a)-1}\boldsymbol{h}(\vartheta_d,\omega_k)] - \boldsymbol{a}_{i,k}^{\mathsf{H}}\Sigma_k^{(a)-1}\boldsymbol{a}_{i,k})\Big], \quad (\mathrm{A}\cdot 8)$$

where $\overset{x}{=}$ denotes equality up to a term independent of x and $\mathbf{1}[\cdot]$ denotes the indicator function that takes the value 1 if its argument is true and 0 otherwise. The expectation of (A·8) taken over all the variables except A is given as

$$\sum_{i,k}\Big[\sum_l \mathbb{E}[z_{k,l}=i](2\mathrm{Re}[\mathbb{E}[s_{k,l}^*]\boldsymbol{a}_{i,k}^{\mathsf{H}}\Sigma_k^{(n)-1}\boldsymbol{y}_{k,l}] - \mathbb{E}[|s_{k,l}|^2]\boldsymbol{a}_{i,k}^{\mathsf{H}}\Sigma_k^{(n)-1}\boldsymbol{a}_{i,k}) + \sum_d \mathbb{E}[c_i=d](2\mathrm{Re}[\boldsymbol{a}_{i,k}^{\mathsf{H}}\Sigma_k^{(a)-1}\boldsymbol{h}(\vartheta_d,\omega_k)] - \boldsymbol{a}_{i,k}^{\mathsf{H}}\Sigma_k^{(a)-1}\boldsymbol{a}_{i,k})\Big], \quad (\mathrm{A}\cdot 9)$$

where

$$\mathbb{E}[z_{k,l}=i] = q(z_{k,l}=i) = [\boldsymbol{\phi}_{k,l}]_i \quad (\mathrm{A}\cdot 10)$$

$$\mathbb{E}[c_i=d] = q(c_i=d) = [\boldsymbol{\psi}_i]_d \quad (\mathrm{A}\cdot 11)$$

$$\mathbb{E}[s_{k,l}] = \mu_{k,l} \quad (\mathrm{A}\cdot 12)$$

$$\mathbb{E}[|s_{k,l}|^2] = |\mu_{k,l}|^2 + \sigma_{k,l}^2. \quad (\mathrm{A}\cdot 13)$$

We can thus confirm that (A·9) is equal up to a constant term to a quadratic form $-\sum_{i,k}(\boldsymbol{a}_{i,k}-\boldsymbol{m}_{i,k})^{\mathsf{H}}\Gamma_{i,k}^{-1}(\boldsymbol{a}_{i,k}-\boldsymbol{m}_{i,k})$ where $\boldsymbol{m}_{i,k}$ and $\Gamma_{i,k}$ are given by (29) and (28), respectively. Hence, $q(A)$ is shown to be the product of multivariate complex normal distributions with mean $\boldsymbol{m}_{i,k}$ and covariance $\Gamma_{i,k}$

$$q(A) = \prod_{i,k}\mathcal{N}_{\mathbb{C}}(\boldsymbol{a}_{i,k};\boldsymbol{m}_{i,k},\Gamma_{i,k}). \quad (\mathrm{A}\cdot 14)$$

B.2 Update Equation for $q(S)$

The term in $\log p(Y,\Theta)$ depending on S is given as

$$-\sum_{k,l}(\boldsymbol{y}_{k,l}-\boldsymbol{a}_{z_{k,l},k}s_{k,l})^{\mathsf{H}}\Sigma_k^{(n)-1}(\boldsymbol{y}_{k,l}-\boldsymbol{a}_{z_{k,l},k}s_{k,l})$$
$$\overset{S}{=}-\sum_{k,l}\sum_i \mathbf{1}[z_{k,l}=i](|s_{k,l}|^2\mathrm{tr}[\boldsymbol{a}_{i,k}\boldsymbol{a}_{i,k}^{\mathsf{H}}\Sigma_k^{(n)-1}] - 2\mathrm{Re}[s_{k,l}^*\boldsymbol{a}_{i,k}^{\mathsf{H}}\Sigma_k^{(n)-1}\boldsymbol{y}_{k,l}]). \quad (\mathrm{A}\cdot 15)$$

The expectation of (A·15) taken over all the variables except S is given as

$$-\sum_{k,l}\sum_i \mathbb{E}[z_{k,l}=i](|s_{k,l}|^2\mathrm{tr}[\mathbb{E}[\boldsymbol{a}_{i,k}\boldsymbol{a}_{i,k}^{\mathsf{H}}]\Sigma_k^{(n)-1}] - 2\mathrm{Re}[s_{k,l}^*\mathbb{E}[\boldsymbol{a}_{i,k}^{\mathsf{H}}]\Sigma_k^{(n)-1}\boldsymbol{y}_{k,l}]), \quad (\mathrm{A}\cdot 16)$$

where

$$\mathbb{E}[z_{k,l} = i] = [\boldsymbol{\phi}_{k,l}]_i, \tag{A·17}$$

$$\mathbb{E}[\boldsymbol{a}_{i,k}^{\mathsf{H}}] = \boldsymbol{m}_{i,k}^{\mathsf{H}}, \tag{A·18}$$

$$\mathbb{E}[\boldsymbol{a}_{i,k}\boldsymbol{a}_{i,k}^{\mathsf{H}}] = \boldsymbol{m}_{i,k}\boldsymbol{m}_{i,k}^{\mathsf{H}} + \Gamma_{i,k}. \tag{A·19}$$

We can thus confirm that (A·16) is equal up to a constant term to a quadratic function $-\sum_{k,l}|s_{k,l} - \mu_{k,l}|^2/\sigma_{k,l}^2$ where $\mu_{k,l}$ and $\sigma_{k,l}^2$ are given by (31) and (30), respectively. Hence, $q(S)$ is shown to be the product of complex normal distributions with mean $\mu_{k,l}$ and variance $\sigma_{k,l}^2$

$$q(S) = \prod_{k,l} \mathcal{N}_{\mathbb{C}}(s_{k,l}; \mu_{k,l}, \sigma_{k,l}^2). \tag{A·20}$$

B.3 Update Equation for $q(Z)$

The term in $\log p(Y, \Theta)$ depending on Z is given as

$$\begin{aligned}
&\sum_{k,l}\Big\{ -(\boldsymbol{y}_{k,l} - \boldsymbol{a}_{z_{k,l},k}s_{k,l})^{\mathsf{H}}\Sigma_k^{(n)-1}(\boldsymbol{y}_{k,l} - \boldsymbol{a}_{z_{k,l},k}s_{k,l}) \\
&\qquad + \log v_{z_{k,l}}\prod_{j=1}^{z_{k,l}-1}(1-v_j)\Big\} \\
&\overset{Z}{=} \sum_{k,l}\sum_i \mathbf{1}[z_{k,l} = i]\Big\{2\mathrm{Re}[s_{k,l}\boldsymbol{y}_{k,l}^{\mathsf{H}}\Sigma_k^{(n)-1}\boldsymbol{a}_{i,k}] \qquad (A·21) \\
&- |s_{k,l}|^2\mathrm{tr}[\boldsymbol{a}_{i,k}\boldsymbol{a}_{i,k}^{\mathsf{H}}\Sigma_k^{(n)-1}] + \log v_i + \sum_{j=1}^{i-1}\log(1-v_j)\Big\}.
\end{aligned}$$

The expectation of (A·21) taken over all the variables except Z is given as

$$\begin{aligned}
&\sum_{k,l}\sum_i \mathbf{1}[z_{k,l} = i]\Big\{2\mathrm{Re}[\mathbb{E}[s_{k,l}]\boldsymbol{y}_{k,l}^{\mathsf{H}}\Sigma_k^{(n)-1}\mathbb{E}[\boldsymbol{a}_{i,k}] \\
&\qquad - \mathbb{E}[|s_{k,l}|^2]\mathrm{tr}[\mathbb{E}[\boldsymbol{a}_{i,k}\boldsymbol{a}_{i,k}^{\mathsf{H}}]\Sigma_k^{(n)-1}] \\
&\qquad + \mathbb{E}[\log v_i] + \sum_{j=1}^{i-1}\mathbb{E}[\log(1-v_j)]\Big\}, \qquad (A·22)
\end{aligned}$$

where

$$\mathbb{E}[\log v_i] = \Psi(\gamma_{i,0}) - \Psi(\gamma_{i,0} + \gamma_{i,1}), \tag{A·23}$$

$$\mathbb{E}[\log(1-v_j)] = \Psi(\gamma_{j,1}) - \Psi(\gamma_{j,0} + \gamma_{j,1}). \tag{A·24}$$

Hence,

$$\begin{aligned}
q(Z) &= \prod_{k,l} q(z_{k,l}) \\
q(z_{k,l}) &\propto \exp\Big\{2\mathrm{Re}[\mu_{k,l}\boldsymbol{y}_{k,l}^{\mathsf{H}}\Sigma_k^{(n)-1}\boldsymbol{m}_{z_{k,l},k} \\
&\quad - (|\mu_{k,l}|^2 + \sigma_{k,l}^2)\mathrm{tr}[(\boldsymbol{m}_{z_{k,l},k}\boldsymbol{m}_{z_{k,l},k}^{\mathsf{H}} + \Gamma_{z_{k,l},k})\Sigma_k^{(n)-1}] \\
&\quad + \Psi(\gamma_{z_{k,l},0}) - \Psi(\gamma_{z_{k,l},0} + \gamma_{z_{k,l},1}) \\
&\quad + \sum_{j=1}^{z_{k,l}-1}(\Psi(\gamma_{j,1}) - \Psi(\gamma_{j,0} + \gamma_{j,1}))\Big\}. \qquad (A·25)
\end{aligned}$$

B.4 Update Equation for $q(V)$

The term in $\log p(Y, \Theta)$ depending on V is given as

$$\begin{aligned}
&\sum_{k,l}\log v_{z_{k,l}}\prod_{j=1}^{z_{k,l}-1}(1-v_j) + \sum_{i=1}^{\infty}(\alpha_0 - 1)\log(1-v_i) \\
&= \sum_i\Big(\sum_{k,l}\mathbf{1}[z_{k,l} = i]\Big)\log v_i \qquad (A·26) \\
&\quad + \sum_i\Big(\sum_{j=i+1}^{\infty}\sum_{k,l}\mathbf{1}[z_{k,l} = j] + \alpha_0 - 1\Big)\log(1-v_i).
\end{aligned}$$

The expectation of (A·26) taken over all the variables except V is given as

$$\begin{aligned}
&\sum_i\Big\{\Big(\sum_{k,l}\mathbb{E}[z_{k,l} = i]\Big)\log v_i \qquad (A·27) \\
&\quad + \Big(\sum_{j=i+1}^{N^*}\sum_{k,l}\mathbb{E}[z_{k,l} = j] + \alpha_0 - 1\Big)\log(1-v_i)\Big\}.
\end{aligned}$$

We can thus confirm that the exponential of (A·27) is equal up to a constant factor to

$$q(V) = \prod_i \mathrm{Beta}(v_i; \gamma_{i,0}, \gamma_{i,1}), \tag{A·28}$$

where $\gamma_{i,0}$ and $\gamma_{i,1}$ are given by (33) and (34), respectively.

B.5 Update Equation for $q(C)$

The term in $\log p(Y, \Theta)$ depending on C is given as

$$\begin{aligned}
&-\sum_{i,k}(\boldsymbol{a}_{i,k} - \boldsymbol{h}(\vartheta_{c_i}, \omega_k))^{\mathsf{H}}\Sigma_k^{(a)-1}(\boldsymbol{a}_{i,k} - \boldsymbol{h}(\vartheta_{c_i}, \omega_k)) + \sum_i \log\rho_{c_i} \\
&\overset{C}{=} \sum_i\Big\{\sum_d \mathbf{1}[c_i = d]\Big(\sum_k(2\mathrm{Re}[\boldsymbol{a}_{i,k}^{\mathsf{H}}\Sigma_k^{(a)-1}\boldsymbol{h}(\vartheta_d, \omega_k)] \\
&\quad - \boldsymbol{h}(\vartheta_d, \omega_k)^{\mathsf{H}}\Sigma_k^{(a)-1}\boldsymbol{h}(\vartheta_d, \omega_k)) + \log\rho_d\Big)\Big\}. \qquad (A·29)
\end{aligned}$$

The expectation of (A·29) taken over all the variables except C is given as

$$\begin{aligned}
&\sum_i\Big\{\sum_d \mathbf{1}[c_i = d]\Big(\sum_k(2\mathrm{Re}[\boldsymbol{m}_{i,k}^{\mathsf{H}}\Sigma_k^{(a)-1}\boldsymbol{h}(\vartheta_d, \omega_k)] \\
&\quad - \boldsymbol{h}(\vartheta_d, \omega_k)^{\mathsf{H}}\Sigma_k^{(a)-1}\boldsymbol{h}(\vartheta_d, \omega_k)) + \mathbb{E}[\log\rho_d]\Big)\Big\}, \qquad (A·30)
\end{aligned}$$

where

$$\mathbb{E}[\log\rho_d] = \Psi(\zeta_d) - \Psi\Big(\sum_{d'}\zeta_{d'}\Big). \tag{A·31}$$

We can thus confirm that the exponential of (A·30) is equal up to a constant factor to

$$q(C) = \prod_i \mathrm{Categorical}(c_i; \boldsymbol{\psi}_i), \tag{A·32}$$

where $[\boldsymbol{\psi}_i]_d$ is given by (35).

B.6 Update Equation for $q(\boldsymbol{\rho})$

The term in $\log p(Y, \Theta)$ depending on $\boldsymbol{\rho}$ is given as

$$\sum_i \log \rho_{c_i} + (\beta_0/D - 1) \sum_d \log \rho_d \qquad (A\cdot 33)$$
$$= \sum_i \sum_d \mathbf{1}[c_i = d] \log \rho_d + (\beta_0/D - 1) \sum_d \log \rho_d$$

The expectation of (A· 33) taken over all the variables except $\boldsymbol{\rho}$ is given as

$$\sum_d \Big(\sum_i [\boldsymbol{\psi}_i]_d + \beta_0/D - 1 \Big) \log \rho_d. \qquad (A\cdot 34)$$

We can thus confirm that the exponential of (A· 34) is equal up to a constant factor to

$$q(\boldsymbol{\rho}) = \text{Dirichlet}(\boldsymbol{\rho}; \zeta_1, \ldots, \zeta_D), \qquad (A\cdot 35)$$

where ζ_d is given by (36).

Hirokazu Kameoka received B.E., M.S. and Ph.D. degrees all from the University of Tokyo, Japan, in 2002, 2004 and 2007, respectively. He is currently a research scientist at NTT Communication Science Laboratories and a Visiting Associate Professor at the University of Tokyo. His research interests include computational auditory scene analysis, statistical signal processing, speech andmusic processing, and machine learning. He is a member of IEEE, the Information Processing Society of Japan (IPSJ) and the Acoustical Society of Japan (ASJ). He received 13 awards over the past 10 years, including the Yamashita Memorial Research Award in 2005 from IPSJ, the Itakura Prize Innovative Young Researcher Award in 2007 and the Awaya Prize Young Researcher Award in 2008 from ASJ, IEEE Signal Processing Society 2008 SPS Young Author Best Paper Award in 2009, and IEICE ISS Young Researcher's Award in Speech Field in 2011.

Misa Sato received her B.E. degree from the University of Tokyo, Japan, in 2012. She is currently an MSc student at the Graduate School of Engineering, the University of Tokyo, Japan. Her research interests include natural language processing and machine learning.

Takuma Ono received the B.E. and M.S. degrees from the University of Tokyo, Japan, in 2010 and 2012, respectively. Since then, he has been with Zenkyoren (National Mutual Insurance Federation of Agricultural Cooperatives).

Nobutaka Ono received the B.E., M.S., and Ph.D. degrees in Mathematical Engineering and Information Physics from the University of Tokyo, Japan, in 1996, 1998, 2001, respectively. He joined the Graduate School of Information Science and Technology, the University of Tokyo, Japan, in 2001 as a Research Associate and became a Lecturer in 2005. He moved to the National Institute of Informatics, Japan, as an Associate Professor in 2011. His research interests include acoustic signal processing, specifically, microphone array processing, source localization and separation, music signal processing, audio coding and watermarking, and optimization algorithms for them. He is the author or co-author of more than 90 articles in international journals and conference proceedings. He was a Tutorial speaker at ISMIR2010 and will organize a special session in EUSIPCO2013. Dr. Ono has been an Associate Editor of the IEEE Transactions on Audio, Speech and Language Processing since 2012 and an Associate Editor of Acoustic Science and Technology since 2012. He is a chair of SiSEC (Signal Separation Evaluation Campaign) committee in 2013. He is a senior member of the IEEE Signal Processing Society and a member of the Acoustical Society of Japan (ASJ), the Information Processing Society of Japan (IPSJ), the Institute of Electrical Engineers of Japan (IEEJ), and the Society of Instrument and Control Engineers (SICE). He received the Sato Paper Award and the Awaya Award from ASJ in 2000 and 2007, respectively, and received the Igarashi Award at the Sensor Symposium on Sensors, Micromachines, and Applied Systems from IEEJ in 2004.

Shigeki Sagayama received the B.E., M.S., and Ph.D. degrees from the University of Tokyo, Tokyo, Japan, in 1972, 1974, and 1998, respectively, all in mathematical engineering and information physics. He joined Nippon Telegraph and Telephone Public Corporation (currently, NTT) in 1974 and started his career in speech analysis, synthesis, and recognition at NTT Labs in Musashino, Japan. From 1990, he was Head of the Speech Processing Department, ATR Interpreting Telephony Laboratories, Kyoto, Japan where he was in charge of an automatic speech translation project. In 1993, he was responsible for speech recognition, synthesis, and dialog systems at NTT Human Interface Laboratories, Yokosuka, Japan. In 1998, he became a Professor of the Graduate School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), Ishikawa. In 2000, he was appointed Professor at the Graduate School of Information Science and Technology (formerly, Graduate School of Engineering), the University of Tokyo. On his retirement from the University of Tokyo in 2013, he became a Project Professor at the National Institute of Informatics (NII). His major research interests include the processing and recognition of speech, music, acoustic signals, handwriting, and images. He was the leader of anthropomorphic spoken dialog agent project (Galatea Project) from 2000 to 2003. Prof. Sagayama received the National Invention Award from the Institute of Invention of Japan in 1991, the Chief Official's Award for Research Achievement from the Science and Technology Agency of Japan in 1996, and other academic awards including Paper Awards from the Institute of Electronics, Information and Communications Engineers, Japan (IEICEJ) in 1996 and from the Information Processing Society of Japan (IPSJ) in 1995. He is a member of the Acoustical Society of Japan, IEICEJ, and IPSJ.