Bayesian Nonparametrics for Microphone Array Processing

Takuma Otsuka, *Student Member, IEEE*, Katsuhiko Ishiguro, *Member, IEEE*, Hiroshi Sawada, *Senior Member, IEEE*, and Hiroshi G. Okuno, *Fellow, IEEE*

***Abstract*—Sound source localization and separation from a mixture of sounds are essential functions for computational auditory scene analysis. The main challenges are designing a unified framework for joint optimization and estimating the sound sources under auditory uncertainties such as reverberation or unknown number of sounds. Since sound source localization and separation are mutually dependent, their simultaneous estimation is required for better and more robust performance. A unified model is presented for sound source localization and separation based on Bayesian nonparametrics. Experiments using simulated and recorded audio mixtures show that a method based on this model achieves state-of-the-art sound source separation quality and has more robust performance on the source number estimation under reverberant environments.**

***Index Terms*—Audio source separation and enhancement (AUD-SSEN), Bayesian nonparametrics, blind source separation, microphone array processing, sound source localization, spatial and multichannel audio (AUD-SMCA), time-frequency masking.**

I. Introduction

COMPUTATIONAL auditory scene analysis (CASA) aims at a machine listening that can extract and analyze useful information and/or meaningful auditory events such as speech content and sound source type from audio recordings [1], [2]. The decomposition of these constituent sound sources is essential for CASA systems because a mixture of audio signals containing multiple sound sources is common in our daily environment [3].

Many CASA systems use multiple sensors, e.g., a microphone array, to decompose the observed mixture into the individual sound sources [4]. Microphone arrays spatially filter the sound sources to act as a decomposition function. That is, they retrieve audio signals from different directions, which is referred to as sound source separation [3], [5]. If the alignment of the microphone array is available, the direction of arrival of each sound source can be estimated, which is sound source localization [6]. While these two problems of separation and localization are mutually dependent, most existing methods deal with a specific part of these problems, and combined in a cascade manner to handle both problems. The overall quality of this combinational approach is prone to be determined by the worst component. For example, the HARK sound source localization and separation system separates the sound sources using the direction of each source estimated by the preceding localization step [7]. Therefore, a localization failure affects the separation. Thus, a unified method is necessary to optimize the mutually dependent problems.

Designing the unified framework for sound source localization and separation involves two challenges; how to model the unified microphone array processing and how to overcome the auditory uncertainties such as reverberation and an unknown source number. Though the localization and separation have been unified by a Bayesian topic model [8], [9], this method assumes that the source number is available a priori, which is not always the case in practice. On the other hand, the estimation of a source number has also been tackled separately from the separation [10], [11]. The drawbacks of these approaches are the necessity of parameter learning in advance or elaborate configuration depending on the auditory environments. An overall framework that unifies the localization and separation under the uncertainty of the source number will contribute to a more flexible CASA system than that by combinational approaches.

This paper presents a model based on Bayesian nonparametrics for sound source separation and localization with source number estimation using a microphone array. We formulate this as a unified twofold clustering problem in which the sound source separation is formulated as a clustering of time-frequency points in the time-frequency domain of the observed spectrogram and sound source localization is formulated as an assignment of each cluster to a certain direction. The clusters corresponding to the different sound sources are generated using a hierarchical Dirichlet process to cope with the source number uncertainty. To infer the latent variables, we derive a collapsed Gibbs sampler that drastically improves the source number estimation accuracy.

II. Problem and Related Work

Fig. 1 outlines our problem. The inputs are a multichannel mixture audio signal and steering vectors that carry information about the alignment of microphones. The outputs are the respective audio signals comprising the observed mixture, the arrival directions of the sound sources, and the number of sources.

Manuscript received March 11, 2013; revised August 30, 2013; accepted November 18, 2013. Date of publication December 11, 2013; date of current version January 10, 2014. This work was supported in part by JSPS KAKENHI under Grants 24220006 and 236577, and Kyoto University and NTT Research Collaboration Program. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Søren Holdt Jensen.

T. Otsuka and H. G. Okuno are with the Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan (e-mail: ohtsuka@kuis.kyoto-u.ac.jp; okuno@kuis.kyoto-u.ac.jp).

K. Ishiguro is with NTT Communication Science Laboratories, NTT Corporation, Kyoto 619-0237, Japan (e-mail: ishiguro.katsuhiko@lab.ntt.co.jp).

H. Sawada is with NTT Service Evolution Laboratories, NTT Corporation, Kanagawa 239-0847, Japan (e-mail: sawada.hiroshi@lab.ntt.co.jp).

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

Digital Object Identifier 10.1109/TASLP.2013.2294582

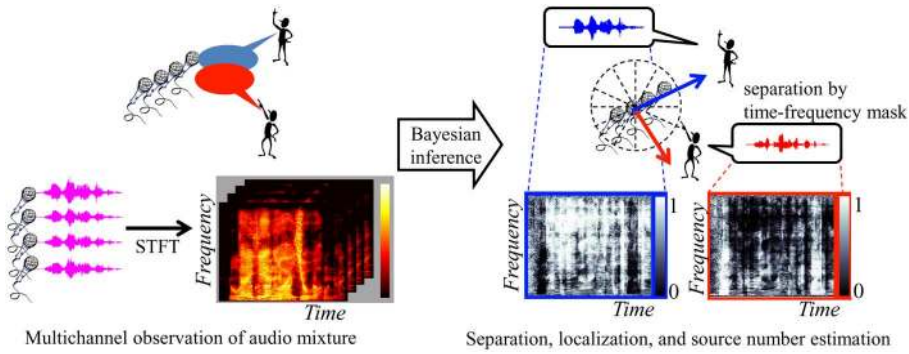


Fig. 1. Illustration of our problem; sound source localization and separation with source number estimation. The process is carried out in the time-frequency domain to generate TF masks. Our Bayesian nonparametrics-based model dispenses with environment-dependent model configurations such as a priori source number information.

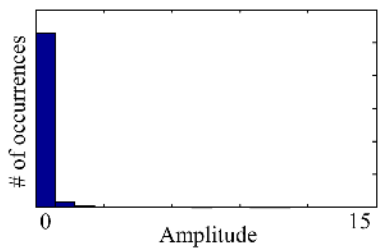


Fig. 2. Histogram of time-frequency amplitudes $|\mathbf{x}_{tf}|$. The power of each time-frequency point is close to zero in most cases. This implies the power sparsity of sound sources in the time-frequency domain. That is, at most one source is assumed dominant at each time-frequency point.

Steering vectors are necessary for sound source localization. A steering vector conveys the time difference of sound arrivals at each microphone given a certain direction and a frequency bin. We use the steering vectors measured in an anechoic chamber so that the vectors can be used independently of the reverberation or environment. We can also synthesize steering vectors when we use a microphone array in a simple shape such as a linear array in Fig. 1.

We make three assumptions on our auditory setup: (1) spatial sparsity of sound sources, (2) power sparsity of the audio in the time-frequency domain, and (3) non-moving sound sources. The first assumption means that all sound sources are located in different directions because a microphone array extracts an audio signal coming from a certain direction. The second assumption is illustrated in Fig. 2. The histogram of the spectrogram amplitudes reveals the energy of most time-frequency points is close to zero. In other words, we are likely to have only one dominant source for each time-frequency point even for a mixture of sound sources. This supports the use of a clustering-based approach for sound source separation [5], [8], [12]. The third assumption means that the sound sources do not change their directions over time and is made for simplicity.

Sound source separation and localization in practical situations have two inherent problems: reverberation and source number uncertainty. When we observe a sound in a room, the observation contains reverberation that can be modeled as a convolutive process [13]. Though methods in the time-frequency domain through a short-time Fourier transform (STFT) are often used to cope with the reverberation, this causes a permutation problem [14]. The permutation problem occurs when the separation is carried out independently of frequency bins in an unsupervised manner, e.g., using independent component analysis (ICA) [3]. To aggregate the spectrogram of a certain source, we must identify the signals of the same sound source from all frequency bins. Independent vector analysis (IVA) [15], [16] avoids the permutation problem by maximizing the independence of the constituent sound sources across all frequency bins simultaneously.

Due to the uncertainty of the number of sources, we have to deal with a model complexity problem and a possibly underdetermined situation. With ICA and IVA, the number of sources N is assumed not to exceed the number of microphones M. However, in practice, N is not always guaranteed to be capped at M, especially when we are unaware of the source number. The case in which $N > M$ is called an underdetermined problem. An approach to this condition is the clustering formulation that generates a time-frequency (TF) mask for each sound source [5], [6], [17], [18], [19].

If the source number is unknown, we need to determine the number of TF masks, which is equal to the number of sources, to estimate. A Bayesian topic model is proposed for the sound source localization and separation [8] in which a TF mask corresponds to a topic regarding the spectrogram as a document. A sufficient number of TF masks are prepared and variational Bayes inference with a sparse prior is carried out to shrink the weight of unnecessary masks. Here, the number of sources is still required when the method extracts the sound sources because the variational Bayes inference is often trapped at a local optimum in terms of the cluster shrinkage. Sources are separated with N masks excluding the other redundant masks that have unnecessary weights. While the posterior inference by Gibbs sampling is presented [9] to avoid local optima, a priori source number information is still necessary due to slow mixing of the Markov chain involving multichannel precision matrices. The local optimum may deteriorate the source number estimation.

The source number uncertainty is closely related to the selection of model complexity. For example, source separation methods using ICA or IVA often reduce the dimensionality of the multichannel observation from the microphone number to the source number by using principal component analysis when the number of sources is available [20]. PCA is employed in order to reduce the number of latent parameters in the separation matrices as a preprocessing of ICA [21], [22]. Similarly, TF

TABLE I
NOTATIONS

Symbol	Meaning
t	Time frame index from 1 to T
f	Frequency bin from 1 to F
k	Class index from 1 to K
d	Direction index from 1 to D
M	Number of microphones
N	Number of sound sources
$\mathbf{x}_{tf}$	Observed M-dimensional complex column vector
z_{tf}	Class indicator at time frame t and frequency bin f
$\boldsymbol{\pi}_t$	Class proportion at time frame t
w_k	Direction indicator for class k
$\boldsymbol{\varphi}$	Direction proportion for all classes
λ_{tf}	Inverse scale parameter for $\mathbf{x}_{tf}$
$\mathbf{H}_{fd}$	Inverse covariance of direction d at frequency bin f
n_{tk}	Number of time-frequency points assigned to class k at time frame t
n_{fk}, n_{fd}	Number of time-frequency points at frequency bin f of class k or direction d, respectively
c_d	Number classes assigned to direction d

masking-based separation methods often use the same number of TF masks as that of sources so that the model complexity should fit the source separation problem [5]. In case of source number uncertainty, where an appropriate model complexity is unknown, a simple solution is to use a sufficiently flexible model. For example, if we can assume the source number is at most four, four TF masks are sufficient. This approach is problematic in two ways: first, a model with a finite number of TF masks fails in the separation when the source number exceeds the number of TF masks. Second, using a too flexible model may affect the performance because redundantly flexible models are apt to overfit the data.

Nonparametric Bayesian models are helpful in such situation since we can bypass a careful selection of the mask number K by assuming an infinite number of TF masks in the model. Furthermore, the prior distribution for the TF masks penalizes unnecessary emergence of TF masks. This property helps the inference to avoid an overfitting that may affect the separation and localization performance. Some Bayesian nonparametric models have been related to microphone array processing techniques. Infinite independent component analysis [23] is a nonparametric counterpart of ICA. Because this model allows only for real-valued variables, the separation is limited to the time domain, which is vulnerable to reverberation. While Nagira *et al.* extend the model into the time-frequency domain [24], they cope with the permutation resolution separately after the separation. This naïve extension into the time-frequency domain is problematic because the inference results in each frequency bin may converge to different number of sources.

The contribution of this paper is twofold. (1) We present a nonparametric Bayesian model that unifies sound source localization, separation, and permutation resolution using a hierarchical Dirichlet process (HDP) [25]. This hierarchical model is advantageous in that the number of sources is globally handled instead of locally for each frequency bin. (2) We derive a collapsed Gibbs sampler (CGS) that promotes the shrinkage of the classes for more accurate sound source estimation. This collapsed inference accelerates the inference by marginalizing out the multichannel precision matrices that a usual Gibbs sampler have to explicitly generate samples [9]. While Kameoka *et al.* develop a similar framework based on Bayesian nonparametrics without an HDP [26], the use of this hierarchical structure in our model is expedient to encourage the temporal synchronization of source dominance over frequency bins so as to generate the time-frequency masks. This mechanism gains a robustness against the reverberation because reverberation is apt to obscure the temporal synchronization in the time-frequency domain.

III. HDP-based Sound Source Localization and Separation

As mentioned, the problem of sound source separation and localization is tackled as a clustering problem. The observed multichannel mixture signal is converted into the TF domain by using STFT. The separation is the clustering of multichannel vectors at each TF point while the localization is the matching of each cluster with steering vectors. A separation with permutation resolution has been developed based on latent Dirichlet allocation (LDA) [27] in which the time frames are regarded as documents and the TF points are treated as words [8]. In this model, a few sound sources (corresponding to topics in the context of LDA) are preferred in each time frame to help the permutation resolution by synchronizing the appearance of the same source across frequency bins. Because LDA is limited to a finite number of sources in spite of the source number uncertainty, we introduce an unbounded model in terms of the number of sources by using HDP [25].

Our model is designed to achieve a balance between the capability to deal with an unbounded number of sound sources and tractable inference of the latent parameters. In order to satisfy these properties, we employ a likelihood distribution suitable to model multichannel observation of directional sound sources as well as conjugate prior distributions. The conjugate priors are helpful to develop an efficient inference procedure in that the parameter estimation is accelerated and stabilized with analytic derivation of the posterior distribution and marginalization of some of the latent parameters.

The notations used in this section are listed in Table I. Fig. 3 shows the graphical representation of our model. The double-circled $\mathbf{x}_{tf}$ is the observation, the circled symbols are latent probability variables, and the plain symbols are fixed values. Section III-A explains how the multichannel input signal is observed and associated with steering vectors. Section III-B de-

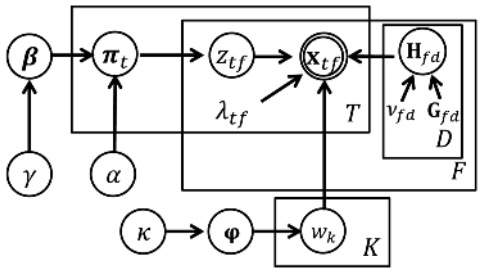


Fig. 3. Graphical model depicting the generative process. Observed variables are double-circled. Latent random variables are denoted with a single circle. Fixed values are denoted by plain symbols. Variables inside a box with an upper-case symbol are independent and identically distributed with respect to the corresponding lower-case index.

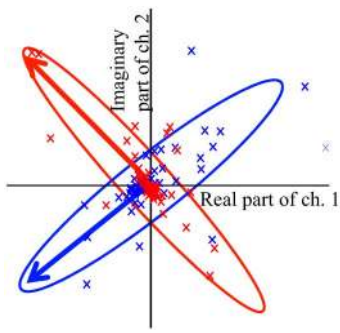


Fig. 4. Scatter plot of multichannel observation at 3200 Hz. Each plot corresponds to each time-frequency point in the multichannel space. Two sources located in different directions form two subspaces (in blue and red dots). The subspace indicated by red and blue arrows corresponds to the direction of arrival of the source.

scribes the inference by using CGS. Section III-C shows how the sound sources are retrieved or localized and how the number of sources is estimated using the sampled latent variables. Finally, Section III-D shows the initialization procedures. A set of variables is denoted with a tilde without subscripts, e.g., $\tilde{\mathbf{x}} = \{\mathbf{x}_{tf} | 1 \leq t \leq T, 1 \leq f \leq F\}$. As revealed in the subsequent sections, the inference of $\tilde{\mathbf{z}}$ corresponds to the estimation of TF masks for separation, and the inference of $\tilde{\mathbf{w}}$ corresponds to the localization.

A. Multichannel Observation and Generative Model

This section explains the generative process described in Fig. 3. We use a covariance model [28] for the likelihood function of the multichannel observation in the TF domain: each sample follows a complex normal distribution with zero mean and time-varying covariance. Fig. 4 shows a scatter plot of the two-channel observations for two sources drawn in blue and red, respectively. We assume that these samples are generated as follows. Let a source located in direction d be dominant in time frame t and frequency bin f. The multichannel signal at t and f is then observed in parallel to a steering vector as $\mathbf{x}_{tf} = s_{tf}\mathbf{q}_{fd}$, where s_{tf} corresponds to the source signal existing at t and f and $\mathbf{q}_{fd}$ is the steering vector for direction d. Vector $\mathbf{x}_{tf}$ is an M-dimensional vector, and each element in $\mathbf{x}_{tf}$ corresponds to a microphone observation. The covariance is $\mathbb{E}[\mathbf{x}_{tf}\mathbf{x}_{tf}^H] = \mathbb{E}[|s_{tf}|^2\mathbf{q}_{fd}\mathbf{q}_{fd}^H]$, where $\cdot^H$ means Hermitian transposition.

The covariance matrix of each source, shown as ellipses in Fig. 4, has an eigenvector with a salient eigenvalue. This vector corresponds to the steering vector associated with the direction of the source. That is, the class estimation of each sample corresponds to the separation, and the investigation of the eigenvectors of the clustered covariances corresponds to the localization of sources.

The covariance above is factorized into a time-varying scale term $|s_{tf}|^2$ and a fixed direction term $\mathbf{q}_{fd}\mathbf{q}_{fd}^H$, and the sound sources are assumed not to move over time. We rewrite these terms as $\lambda_{tf}\mathbf{H}_{fd}$, where λ_{tf} is the scale parameter corresponding to the inverse of $|s_{tf}|^2$, $\mathbf{H}_{fd} \approx (\mathbf{q}_{fd}\mathbf{q}_{fd}^H + \varepsilon\mathbf{I}_M)^{-1}$, and $\mathbf{I}_M$ is the $M \times M$ identity matrix. We use this inverse notation for the convenience of placing a conjugate prior over $\mathbf{H}_{fd}$. While λ_{tf} has been treated as a probability variable in [8], we fix this parameter as $\lambda_{tf} = \frac{1}{\mathbf{x}^H\mathbf{x}}$ so that an efficient collapsed inference is analytically derived by marginalizing out $\mathbf{H}_{fd}$. The likelihood function is a complex normal distribution:

$$\mathbf{x}_{tf}|z_{tf}, \tilde{\mathbf{w}}, \lambda_{tf}, \tilde{\mathbf{H}} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{x}_{tf}|\mathbf{0}, (\lambda_{tf}\mathbf{H}_{fw_{z_{tf}}})^{-1}), \quad (1)$$

where z_{tf} and w_k indicate the class of $\mathbf{x}_{tf}$ and the direction of class k, respectively. Thus, $w_{z_{tf}}$ denotes the direction in which $\mathbf{x}_{tf}$ is located. The probability density function (pdf) of a complex normal distribution $\mathcal{N}_{\mathbb{C}}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$ is defined as $\frac{|\boldsymbol{\Lambda}|}{\pi^M}\exp(-\mathbf{x}^H\boldsymbol{\Lambda}\mathbf{x})$ [29] with mean $\boldsymbol{\mu}$ and precision $\boldsymbol{\Lambda}$. $|\boldsymbol{\Lambda}|$ is the determinant of matrix $\boldsymbol{\Lambda}$.

The direction matrix $\mathbf{H}_{fd}$ follows the conjugate prior, i.e., complex Wishart distribution [30].

$$\mathbf{H}_{fd} \sim \mathcal{W}_{\mathbb{C}}(\mathbf{H}_{fd}|\nu_{fd}, \mathbf{G}_{fd}), \quad (2)$$

where the pdf of complex Wishart distribution $\mathcal{W}_{\mathbb{C}}(\mathbf{H}|\nu, \mathbf{G})$ is $\frac{|\mathbf{H}|^{\nu-M}\exp\{-\mathrm{tr}(\mathbf{H}\mathbf{G}^{-1})\}}{|\mathbf{G}|^{\nu}\pi^{M(M-1)/2}\prod_{i=0}^{M-1}\Gamma(\nu-i)}$; $\mathrm{tr}(\mathbf{A})$ is the trace of $\mathbf{A}$ and $\Gamma(x)$ is the gamma function. The hyperparameters of the complex Wishart distribution are set as $\nu_{fd} = M$ and $\mathbf{G}_{fd} = (\mathbf{q}_{fd}\mathbf{q}_{fd}^H + \varepsilon\mathbf{I}_M)^{-1}$. $\mathbf{G}_{fd}$ is generated from the given steering vectors $\mathbf{q}_{fd}$, where $\mathbf{q}_{fd}$ is normalized s.t. $\mathbf{q}_{fd}^H\mathbf{q}_{fd} = 1$ and ε is set to 0.01 to enable inverse operation.

An HDP [25] is used as the generative process of z_{tf}, which is an infinite extension of an LDA. We introduce this hierarchical generative process to resolve the permutation ambiguity [8]. First, global class proportion $\boldsymbol{\beta}$ is generated, where the dimensionality of $\boldsymbol{\beta}$ is infinitely large. Each element represent the average weights of infinitely-many classes throughout the spectrogram. Then, the time-wise class proportion $\boldsymbol{\pi}_t$ is sampled in accordance with $\boldsymbol{\beta}$. Again, $\boldsymbol{\pi}_t$ is an infinite-dimensional vector where the elements represent the weights of infinite classes at the specific time frame t. Finally, each z_{tf} is sampled in accordance with the time-wise class proportion $\boldsymbol{\pi}_t$. As Fig. 5 shows the dominance of each source is synchronized across frequency bins. Therefore, we achieve the permutation resolution by introducing $\boldsymbol{\pi}_t$. The stick-breaking construction for an HDP [25] is given by:

$$\boldsymbol{\beta}|\gamma \sim \mathrm{GEM}(\gamma), \quad \boldsymbol{\pi}_t|\alpha, \boldsymbol{\beta} \sim \mathrm{DP}(\alpha, \boldsymbol{\beta}), \quad z_{tf}|\boldsymbol{\pi}_t \sim \boldsymbol{\pi}_t, \quad (3)$$

where $\mathrm{GEM}(\gamma)$ is the Griffiths-Engen-McCloskey distribution with concentration γ; $\mathrm{DP}(\alpha, \boldsymbol{\beta})$ denotes the Dirichlet process with concentration α and base measure $\boldsymbol{\beta}$. Here, the expectation

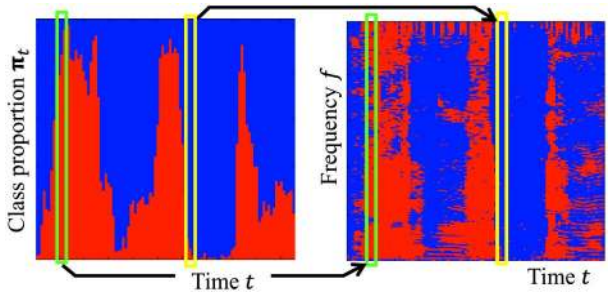


Fig. 5. Left: class proportion $\boldsymbol{\pi}_t$ for each time frame. Right: TF mask of two sources denoted by z_{tf}. Each TF point is assigned to a class in accordance with the class proportion of the time frame.

of $\boldsymbol{\pi}_t$ satisfies $\mathbb{E}[\boldsymbol{\pi}_t] = \boldsymbol{\beta}$. We place gamma distribution priors for concentrations $\gamma \sim \mathcal{G}(\gamma|a_\gamma, b_\gamma)$ and $\alpha \sim \mathcal{G}(\alpha|a_\alpha, b_\alpha)$. The hyperparameters are set as $a_\gamma = 0.05$, $b_\gamma = 5$, $a_\alpha = 0.01$, and $b_\alpha = 1$.

Direction indicator w_k contributes to the sound source localization as well as to the permutation resolution because classes from the same direction are associated with each other across all frequency bins. This variable is drawn from proportion $\boldsymbol{\varphi}$ generated from a flat Dirichlet distribution.

$$\boldsymbol{\varphi}|\kappa \sim \mathcal{D}\left(\boldsymbol{\varphi}|\frac{\kappa}{D}\mathbf{1}_D\right), \quad w_k|\boldsymbol{\varphi} \sim \boldsymbol{\varphi}, \tag{4}$$

where $\mathbf{1}_D$ is a D-dimensional vector in which all elements are 1 and $\mathcal{D}(\cdot|\boldsymbol{\alpha})$ denotes the Dirichlet distribution with parameter $\boldsymbol{\alpha}$. Our model is a finite mixture with regard to direction due to the limitation of the spatial resolution of microphone arrays. We also place a gamma prior over κ as $\mathcal{G}(\kappa|a_\kappa, b_\kappa)$, where $a_\kappa = 1$ and $b_\kappa = 1$.

B. Inference by Collapsed Gibbs Sampler

For sound source separation and localization, the inference of $\tilde{\mathbf{z}}$ and $\tilde{\mathbf{w}}$ is important. These variables are inferred by using a CGS with $\boldsymbol{\pi}, \boldsymbol{\varphi}$, and $\tilde{\mathbf{H}}$ marginalized out. The joint distribution of $\tilde{\mathbf{x}}, \tilde{\mathbf{z}}$, and $\tilde{\mathbf{w}}$ becomes

$$\begin{aligned}
&p(\tilde{\mathbf{x}}, \tilde{\mathbf{z}}, \tilde{\mathbf{w}}|\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}}, \tilde{\mathbf{G}}, \alpha, \boldsymbol{\beta}, \gamma, \kappa) \\
&= \int p(\tilde{\mathbf{x}}, \tilde{\mathbf{H}}|\tilde{\mathbf{z}}, \tilde{\mathbf{w}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}}, \tilde{\mathbf{G}}) p(\tilde{\mathbf{z}}, \tilde{\boldsymbol{\pi}}|\alpha, \boldsymbol{\beta}, \gamma) \\
&\qquad p(\tilde{\mathbf{w}}, \boldsymbol{\varphi}|\kappa) d\tilde{\mathbf{H}} d\boldsymbol{\pi} d\boldsymbol{\varphi} \\
&= \prod_{tf}\left(\frac{\lambda_{tf}}{\pi}\right)^M \prod_{fd} \frac{\prod_{i=0}^{M-1}\Gamma(\hat{\nu}_{fd}-i)|\hat{\mathbf{G}}_{fd}|^{\hat{\nu}_{fd}}}{\prod_{i=0}^{M-1}\Gamma(\nu_{fd}-i)|\mathbf{G}_{fd}|^{\nu_{fd}}} \\
&\quad \prod_t \left\{\frac{\Gamma(\alpha)}{\Gamma(\alpha+n_{t\cdot})}\prod_k \frac{\Gamma(\alpha\beta_k + n_{tk})}{\Gamma(\alpha\beta_k)}\right\}\frac{\Gamma(\kappa)}{\Gamma(\kappa+c_\cdot)} \\
&\quad \prod_d \frac{\Gamma\left(\frac{\kappa}{D}+c_d\right)}{\Gamma\left(\frac{\kappa}{D}\right)},
\end{aligned} \tag{5}$$

where a dot in the subscripts denote summation over the index, e.g., $n_{t\cdot} = \sum_k n_{tk}$ and $c_\cdot = \sum_d c_d$. Note that, as explained in [25], a finite k can be handled during the inference so that the product over k is valid. The dimensionality of $\boldsymbol{\beta}$ dynamically changes over sampling iterations in accordance with the number of classes actually drawn. The posterior parameters of the complex Wishart distribution, $\hat{\nu}_{fd}$ and $\hat{\mathbf{G}}_{fd}$, are updated using sufficient statistics:

$$\begin{aligned}
\hat{\nu}_{fd} &= \nu_{fd} + \sum_{t:w_{z_{tf}}=d} 1 = \nu_{fd} + n_{fd}, \\
\hat{\mathbf{G}}_{fd}^{-1} &= \mathbf{G}_{fd}^{-1} + \sum_{t:w_{z_{tf}}=d} \lambda_{tf}\mathbf{x}_{tf}\mathbf{x}_{tf}^H,
\end{aligned} \tag{6}$$

where $\sum_{t:w_{z_{tf}}=d}\cdot$ means a summation over the samples assigned to direction d in frequency bin f.

From Eq. (5), z_{tf} and w_k are stochastically updated:

$$\begin{aligned}
&p\left(z_{tf}=k|\tilde{\mathbf{x}}, \boldsymbol{\vartheta}\backslash z_{tf}\right) \propto \left(\alpha\beta_k + n_{tk}^{\backslash tf}\right)\frac{\Gamma\left(\hat{\nu}_{fw_k}^{\backslash tf}+1\right)}{\Gamma\left(\hat{\nu}_{fw_k}^{\backslash tf}-M+1\right)} \\
&\qquad \frac{\left|\mathrm{inv}\left(\hat{\mathbf{G}}_{fw_k}^{\backslash tf}\right)\right|^{\hat{\nu}_{fw_k}^{\backslash tf}}}{\left|\mathrm{inv}\left(\hat{\mathbf{G}}_{fw_k}^{\backslash tf}\right)+\lambda_{tf}\mathbf{x}_{tf}\mathbf{x}_{tf}^H\right|^{\hat{\nu}_{fw_k}^{\backslash tf}+1}},
\end{aligned} \tag{7}$$

$$\begin{aligned}
&p\left(w_k=d|\tilde{\mathbf{x}}, \boldsymbol{\vartheta}\backslash w_k\right) \propto \left(\frac{\kappa}{D}+c_d^{\backslash k}\right) \\
&\quad \prod_f \left\{\prod_{i=0}^{M-1}\frac{\Gamma\left(\hat{\nu}_{fd}^{\backslash k}+n_{fk}-i\right)}{\Gamma\left(\hat{\nu}_{fd}^{\backslash k}-i\right)}\right. \\
&\qquad \left.\frac{\left|\mathrm{inv}\left(\hat{\mathbf{G}}_{fd}^{\backslash k}\right)\right|^{\hat{\nu}_{fd}^{\backslash k}}}{\left|\mathrm{inv}\left(\hat{\mathbf{G}}_{fd}^{\backslash k}\right)+\sum_{t:z_{tf}=k}\lambda_{tf}\mathbf{x}_{tf}\mathbf{x}_{tf}^H\right|^{\hat{\nu}_{fd}^{\backslash k}+n_{fk}}}\right\},
\end{aligned} \tag{8}$$

where $\boldsymbol{\vartheta}\backslash z$ denotes all latent variables except z, superscripts $\backslash tf$ and $\backslash k$ mean the statistics without the sample at t and f or samples of class k, respectively, and $\mathrm{inv}(\mathbf{G})$ is the inverse matrix of $\mathbf{G}$.

Let K be the number of sampled classes. To allow for the probability of z_{tf} taking an unassigned class $K+1$ in Eq. (7), $\boldsymbol{\beta}$ has $K+1$ elements, as explained in [25]. To calculate the probability of $z_{tf} = K+1$, $w_{K+1} = d$ is temporarily drawn with probability $\frac{\kappa/D+c_d}{\kappa+c_\cdot}$. If z_{tf} is chosen to be $K+1$, K is updated as $K \leftarrow K+1$, and the dimensionality of $\boldsymbol{\beta}$ increases by one with $\beta_K \leftarrow b\beta_K$ and $\beta_{K+1} \leftarrow (1-b)\beta_K$, where b is drawn from a beta distribution: $b \sim B(1, \gamma)$.

The updates of the other parameters, $\alpha, \boldsymbol{\beta}, \gamma$, and κ, follow a procedure described in [25], [31]. These parameters are updated using auxiliary variables.

C. Localization, Separation, and Source Number Estimation

The collapsed Gibbs sampler described in Eqs. (7, 8) produces the samples of latent variables indexed by i: $\{\tilde{\mathbf{z}}^{(i)}, \tilde{\mathbf{w}}^{(i)}\}_{i=1}^I$. Sound sources are retrieved by applying a TF mask corresponding to a certain direction. The multichannel

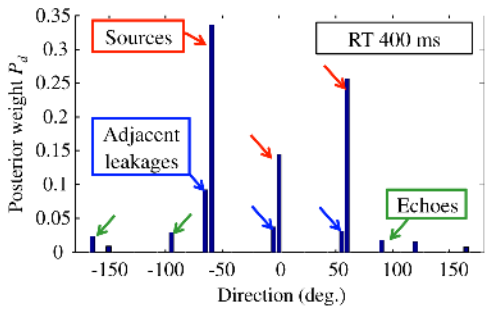


Fig. 6. Posterior weights for three-source mixture with 400-ms reverberation time. Three salient peaks are found at $-60°$, $0°$, and $60°$ indicated by red arrows. Some echo component (e.g., reflection on the wall) is observed as TF masks with small weights.

spectrogram of a sound source in direction d, denoted by $\hat{\mathbf{x}}^d_{tf}$, is retrieved using

$$\hat{\mathbf{x}}^d_{tf} = \frac{1}{I}\sum_{i=1}^{I} \delta\left(w^{(i)}_{z^{(i)}_{tf}}, d\right)\mathbf{x}_{tf}, \tag{9}$$

where $\delta(m,n)$ is the Kronecker delta, i.e., $\delta(m,n) = 1$ if $m = n$, and 0 otherwise. The factor $\frac{1}{I}\sum_{i=1}^{I}\delta(w^{(i)}_{z^{(i)}_{tf}}, d)$ is the estimated TF mask for direction d at time t and frequency f. We can distinguish in which direction sound sources exist by defining the posterior weight for each direction as

$$P_d = \frac{1}{I}\sum_{i=1}^{I}\sum_{tf}\delta\left(w^{(i)}_{z^{(i)}_{tf}}, d\right). \tag{10}$$

If we want N sources from the mixture, we choose N directions in descending order of P_d. The sound sources are thereby localized and separated.

The number of sound sources is estimated using the posterior weights defined in Eq. (10). Fig. 6 shows the posterior weights of a three-source mixture with a reverberation time of 400 (ms). We can set three salient peaks (indicated by red arrows) with smaller peaks in the adjacent directions (blue arrows). Reverberation causes additional peaks corresponding to echoes (green arrows). The number of sources is estimated using a three-step process. (1) Ignore the weights adjacent to larger peaks: $P_d \leftarrow 0$, if $P_d < P_{d+1}$ or $P_d < P_{d-1}$. (2) Sort the weights in descending order: $P'_1 > P'_2 > \ldots > P'_D$. (3) Find the number $\hat{N}$ where the weight drops most sharply: $\hat{N} = \arg\max_N P'_N/P'_{N+1}$ while $P'_{N+1} > 0$. If P'_N/P'_{N+1} monotonically increases until $P'_{N+1} = 0$, $\hat{N} = N$.

D. Initialization of the Inference

The inference is initialized in a similar way as previously reported [8]. The inference starts with a certain number of classes K. First, w_k is initialized with a uniform distribution whose support has no overlap with the other classes. Then, each z_{tf} is drawn using the sampled w_k and the hyperparameter of Wishart distribution, $\mathbf{G}_{fd}$, generated from the steering vectors:

$$\begin{aligned} p(w_k = d) &= \mathcal{U}\left(\left\{d \middle| \frac{k-1}{K}D \le d < \frac{k}{K}D\right\}\right), \\ p(z_{tf} = k) &\propto \exp\left(-\mathbf{x}^H_{tf}\mathbf{G}_{fw_k}\mathbf{x}_{tf}\right), \end{aligned} \tag{11}$$

where $\mathcal{U}(A)$ is a pdf of uniform distribution on set A.

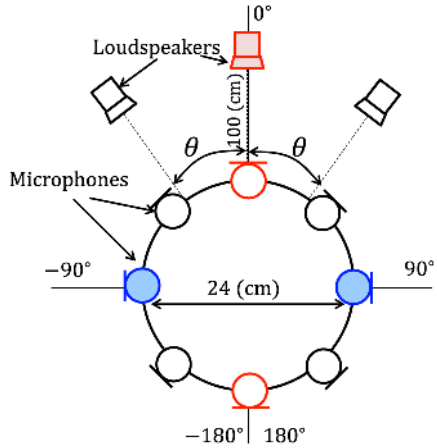


Fig. 7. Microphone array configuration and positions of sound sources. The number of microphones M is 2, 4, or 8 whereas the number of sources N is set as 2 or 3. When $M = 2$, blue microphones are used. When $M = 4$, blue and red microphones are used. All microphones are used when $M = 8$. When $N = 2$, the center source illustrated in red is omitted.

IV. Evaluation

We evaluate the sound source separation, localization, and source number estimation performances of our HDP-CGS method using simulated and recorded mixtures. We compare the source separation performance with those of state-of-the-art sound source separation methods: LDA-VB [8] and IVA [16] for $M \ge N$ and TF masking with permutation resolution (TF-perm.) [5] for $M < N$. The localization and source number estimation performance are compared between HDP-CGS and LDA-VB.

A. Experimental Setup

Fig. 7 illustrates the experimental setup. We used two, four, or eight microphones ($M = 2, 4, 8$) to observe two or three sound source mixtures ($N = 2, 3$) with the interval $\theta = 30$, 60, and 90°. The microphones depicted in shaded blue were used when $M = 2$, those depicted in blue and red were used when $M = 4$, and all microphones in Fig. 7 were used when $M = 8$. The center speaker (in red) was omitted when $N = 2$. The steering vectors, $\tilde{\mathbf{q}}$, of the microphone array were measured in an anechoic room such that $D = 72$ with 5° resolution. When $M = 2$, we used the steering vectors ranging from $-90°$ to 90° so as to avoid the front-back ambiguity. The steering vectors were generated from a Fourier transform of the first 1024 points of the anechoic impulse responses.

The experiments used both simulated and recorded mixtures in three rooms with reverberation times (RT) of 150, 400, and 600 (ms). The simulated mixtures were generated by convoluting the impulse responses measured in each room. The spectrograms of the impulse responses are shown in Fig. 20 with an explanation in the appendix. For each condition, 20 mixtures were tested using JNAS phonetically-balanced Japanese utterances. The average length of these mixtures is around 5 (sec). The audio signals were sampled at 16,000 (Hz), and STFT was carried out with a 1024 (pt) hanning window and a 256 (pt) shift size.

We use the signal-to-distortion ratio (SDR) as the metric for separation quality [32]. Since this ratio is calculated from the

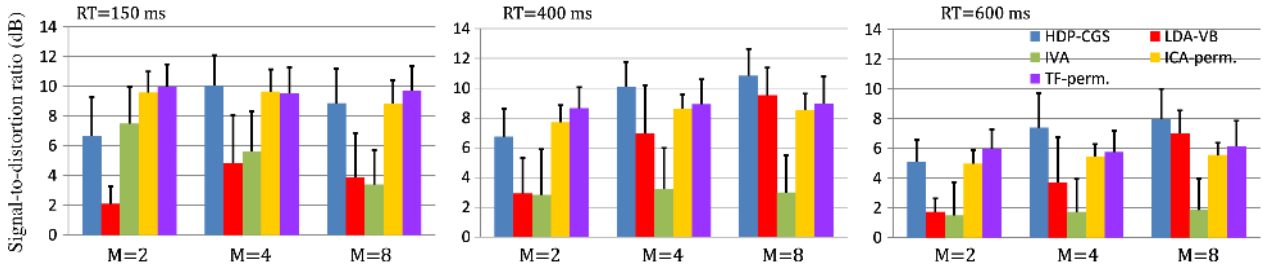


Fig. 8. Separation results for simulated mixtures with two sources. Larger value means better separation. Bars are the means, and the segments are the standard deviations. Color represents each method. Left: RT $= 150$ (ms); middle: RT $= 400$ (ms); right: RT $= 600$ (ms).

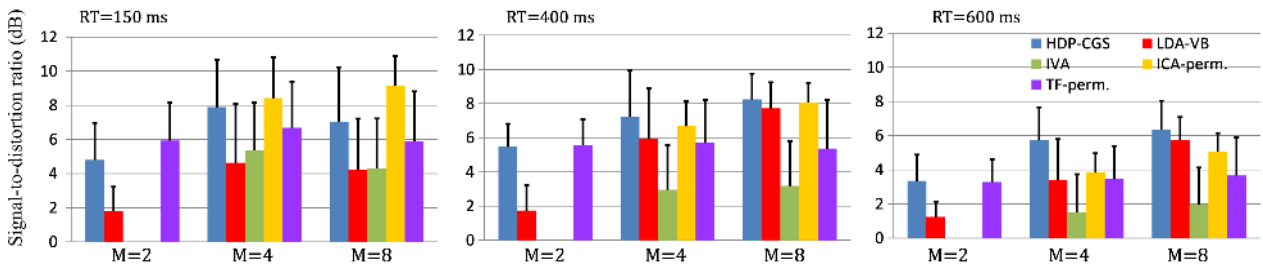


Fig. 9. Separation results for simulated mixtures with three sources.

N original signals and the identical number of separated signals, we extracted N sound sources regardless of the source number estimation results. We compare five methods in this experiments: HDP-CGS, LDA-VB, IVA, ICA with permutation resolution [33] (ICA-perm.), and TF masking with permutation resolution (TF-perm.). HDP-CGS and LDA-VB separate N sources in descending order of the posterior weight, as explained in Section III-C, whereas IVA and ICA-perm. take N sources in descending order of the power of the separated audio signals, and TF-perm. carries out TF mask clustering assuming N sources. Note that TF-perm. uses the fact of N sources for the inference while the inferences of HDP-CGS and LDA-VB are independent of N. The number of classes K used by LDA-VB was 12, and HDP-CGS was initialized with $K = 12$.

In Section IV-D, the source number estimation results are compared between HDP-CGS, LDA-VB, and source separation and source counting method developed by Araki *et al.* [34]. Since this method is developed for stereo observation ($M = 2$), we refer to this method as Stereo hereafter. The idea of the source counting of Stereo is similar to our method in that Stereo generates TF masks for each source and then estimates the source number by counting the TF masks the weight of which is above a certain threshold. The TF masks are estimated through the EM algorithm where the observation is based on the phase difference of two microphone, that is, the phase of non-diagonal elements of $\mathbf{x}_{tf}\mathbf{x}_{tf}^H$. In contrast, our method uses both the phase and level difference by considering $\lambda_{tf}\mathbf{x}_{tf}\mathbf{x}_{tf}^H$ in Eq. (6), and extends the model to any number of microphones.

The inference (parameter estimation) procedures are configured as follows. The collapsed Gibbs sampler for HDP-CGS was iterated 50 times with the first 20 cycles discarded as a burn-in period. The other methods are iterated until the evaluation function converges. LDA-VB typically converged in about 15 iterations. The iteration of IVA was 50 cycles. ICA-perm. carried out 50 iterations for the separation for each frequency bin and 30 iterations for the permutation resolution. TF-perm. required 50 iterations for the separation and 30 iterations for the permutation resolution, respectively. Computational complexity of each method is compared in Table II. The number of iterations is the necessary cycles for the convergence. Here, one iteration involves the whole spectrogram; for example, TF masking-based methods updates the weight of TF masks at all TF points in one iteration whereas linear separation methods updates the separation matrices of all frequency bins in each iteration. The class number K for HDP-CGS is the number of instantiated classes during the inference. HDP-CGS requires iterative M^3 operations due to the calculation of matrix determinants in Eqs. (7) and (8). In practice, we can accelerate the computation by skipping the evaluation of the probability for almost empty classes and directions.

TABLE II
COMPUTATIONAL COMPLEXITY OF EACH METHOD

Method	Complexity per iteration	# of iterations until convergence
HDP-CGS	$O(TFKM^3 + FKDM^3)$	50
LDA-VB	$O(TFKDM^2 + FDM^3)$	15
IVA	$O(TFM^3)$	50
ICA-perm.	$O(TFM^2)$	50
TF-perm.	$O(TFNM + TFN^2)$	50

B. Separation Results

Figs. 8–11 show the separation results for simulated and recorded mixtures with two or three sources. The bars are grouped by the microphone number M for each method. The

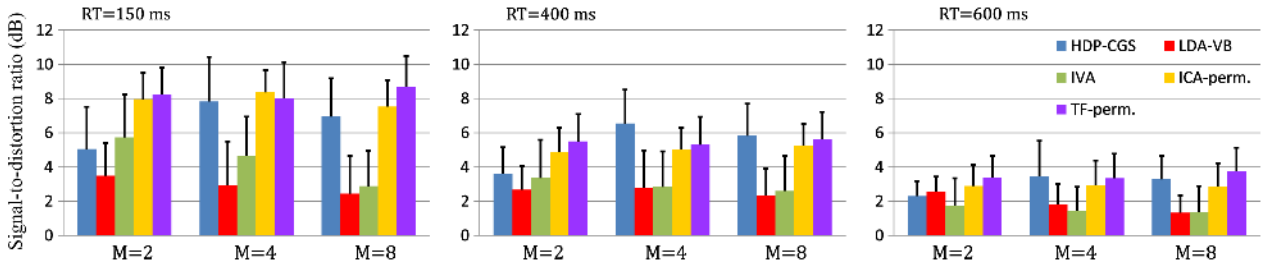


Fig. 10. Separation results for recorded mixtures with two sources.

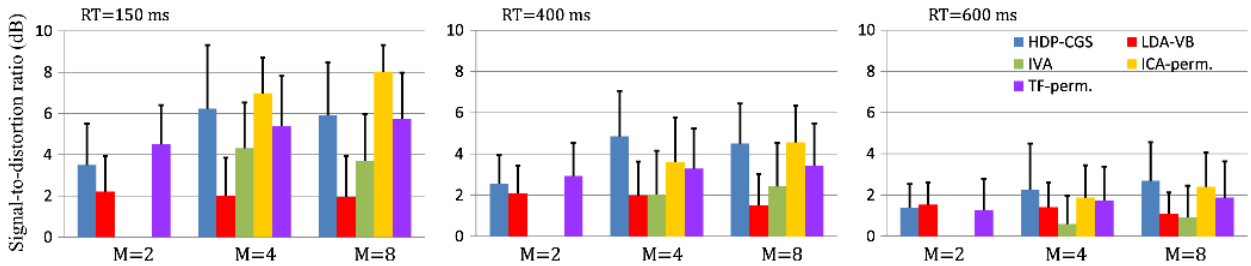


Fig. 11. Separation results for recorded mixtures with three sources.

SDR scores are averaged over the source interval θ because the interval θ makes little difference in the separation quality. The color represents each method according to the legend in the rightmost figures. In general, a longer reverberation time degrades the SDR of all methods. A comparison of Figs. 8 with 9, and Figs. 10 with 11 shows that a larger number of sources in the observed mixture degrades the separation quality of the respective sources.

Our method is superior to or competitive with the other methods when $M = 4$ and 8. In particular, HDP-CGS tends to produce better SDR than LDA-VB. This is as expected because LDA-VB has more than N masks with non-negligible weights due to local optima, which results in the limited SDR scores. In contrast, the performance of our method is limited especially when $M = 2$. This is explained as follows. Even though the microphone number is small $M = 2$, the proposed approach separates the sources considering a variety of possible numbers of sources with the limited dimensionality of the observation. This source number uncertainty limits the performance of HDP-CGS. On the other hand, linear models including ICA and IVA can assume that the possible source number is two when $M = 2$. The determined problem of two-source and two-microphone is also suitable for the linear models in terms of the model complexity. Thus, the $M = 2$ setup is advantageous for the linear models. Similarly, TF-perm. uses the same number of TF masks as that of the sources. This assumption improves the separation quality of TF-perm. method. Another remaining issue is the reverberation. We can note that the long reverberation (600 ms) in the recorded mixtures deteriorates the separation quality of any methods. To cope with these situations, an explicit model for the reverberation is enumerated as a future work.

The performance with the recorded mixtures in Figs. 10 and 11 is worse than that with the simulated mixtures in Figs. 8 and 9. This is because the recorded audio contains more reverberation in the lower frequency region than simulated mixtures. As shown in the appendix, the energy of the reverberation in the impulse responses used to generate the simulated mixtures is attenuated in the lower frequency range. In contrast, the recorded mixtures preserve the lower frequency reverberation of the environments. The intensity of the reverberation in the low frequency region severely affects the separation and localization performance because the subspace structure shown in Fig. 4 is originally vague and is further disturbed by the reverberation. Furthermore, the SDR score is likely to be influenced from the disturbance of the separation quality in the lower frequency region because speech signals concentrate their power on the lower part in the frequency domain.

C. Localization Results

Figs. 12–15 show the localization results of HDP-CGS and LDA-VB in terms of the absolute errors of the localization results. Similarly to the separation results, the larger number of microphones improves the localization performance while the reverberation tends to affect the localization due to the reflection of the sounds. The errors in LDA-VB is more prominent than those in HDP-CGS because the posterior probability of $w_k = d$ can fall into a local optimum with the variational Bayesian inference of LDA-VB.

For some applications, the localization resolution specified by the steering vectors (5° in our experiment) may be insufficient. We can apply the following post-processing to the separated sound image $\hat{\mathbf{x}}_{tf}^d$ to enhance the localization resolution. Let $\mathbf{R}_f^d \equiv \sum_t \hat{\mathbf{x}}_{tf}^d$ be the autocorrelation of the sound image and $\hat{\mathbf{q}}_{fd}$ be the eigenvector associated with the largest eigenvalue of $\mathbf{R}_f^d$. The vector $\hat{\mathbf{q}}_{fd}$ is a clue to investigate the direction of the sound source since this vector is parallel to one of the subspaces illustrated in Fig. 4. The direction that matches $\hat{\mathbf{q}}_{fd}$ is investigated by interpolating the given steering vectors of adjacent directions, $\mathbf{q}_{fd}$ and $\mathbf{q}_{fd\pm1}$ [35], [36].

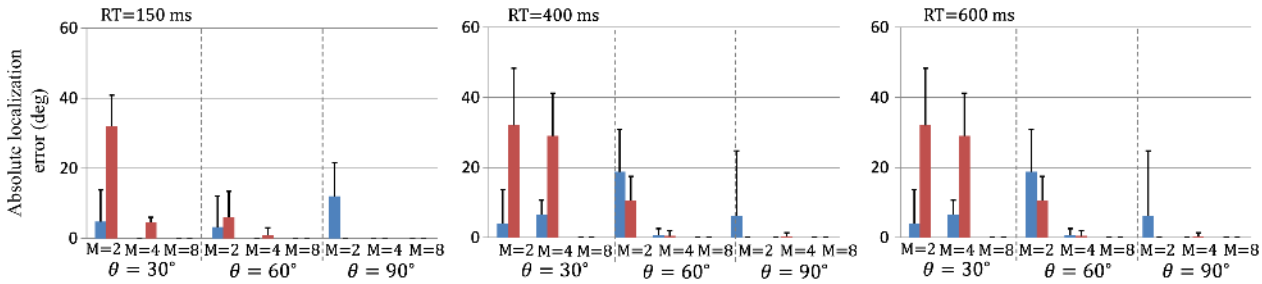


Fig. 12. Localization results for simulated mixtures with two sources in terms of absolute errors. Smaller value means better localization. Bars are the means, and the segments are the standard deviations. Color represents each method: blue bars indicate our HDP-CGS while red bars denote LDA-VB. Left: $\mathrm{RT} = 150$ (ms); middle: $\mathrm{RT} = 400$ (ms); right: $\mathrm{RT} = 600$ (ms).

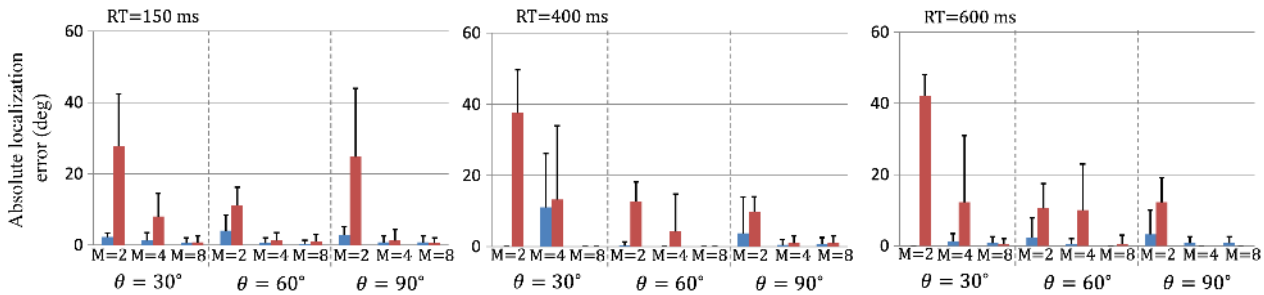


Fig. 13. Localization results for simulated mixtures with three sources. Blue: HDP-CGS; red: LDA-VB.

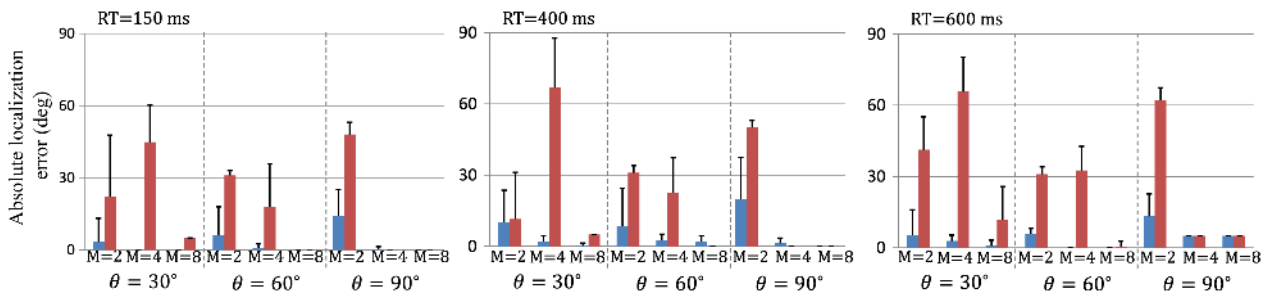


Fig. 14. Localization results for recorded mixtures with two sources. Blue: HDP-CGS; red: LDA-VB.

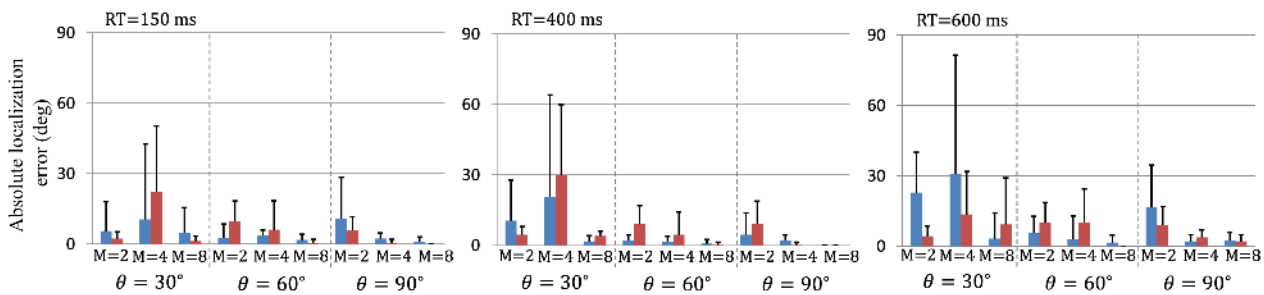


Fig. 15. Localization results for recorded mixtures with three sources. Blue: HDP-CGS; red: LDA-VB.

D. Source Number Estimation Results

Figs. 16–19 show the source number estimation results with HDP-CGS, LDA-VB, and Stereo. Each figure show the histogram of source number estimates for each microphone number and reverberation. Note that the results of Stereo is presented for only $M = 2$ case. The results are merged in terms of θ for this evaluation because this parameter made little difference to the source number estimation performance. An ideal result of the estimation is that the bar is concentrated at the ground truth source number N.

A comparison of HDP-CGS and LDA-VB reveals that HDP-CGS clearly outperforms LDA-VB because the blue bars are mostly located at the true source number where as the red bars are distracted to larger source numbers. These results demonstrate that the CGS works well for source number estimation because it avoids local optima of the latent space, unlike variational Bayes inference. The results of Stereo tends to have a larger variance than HDP-CGS with $M = 2$ case. This is considered because the observation model of Stereo uses only the phase difference between the two microphones and thus the TF mask generation is sometimes unstable. This makes it diffi-

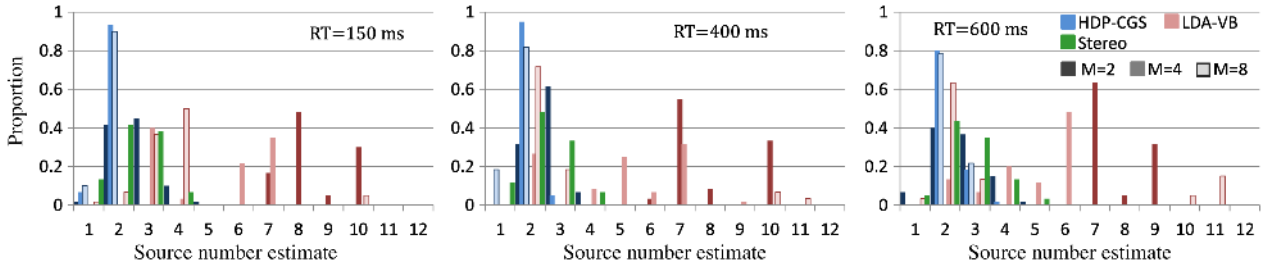


Fig. 16. Source number estimation results for simulated mixtures with two sources ($N = 2$). Each bar represents the proportion of source number estimates. Color represents each method, and shade represents the number of microphones. Stereo method is only for $M = 2$. Left: $\mathrm{RT} = 150$ (ms); middle: $\mathrm{RT} = 400$ (ms); right: $\mathrm{RT} = 600$ (ms).

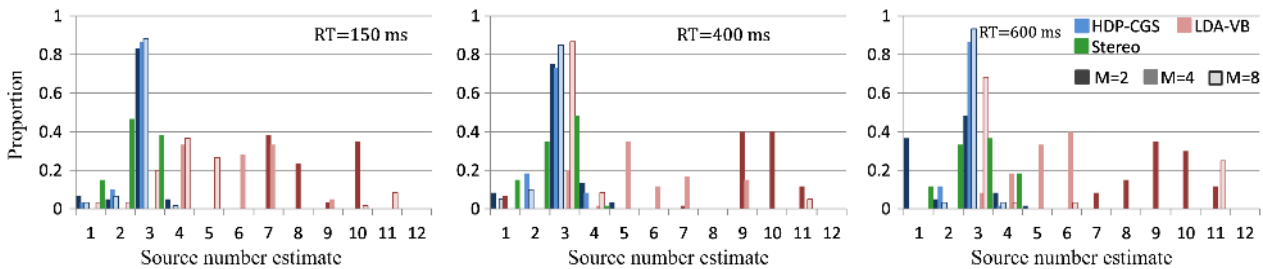


Fig. 17. Source number estimation results for simulated mixtures with three sources ($N = 3$).

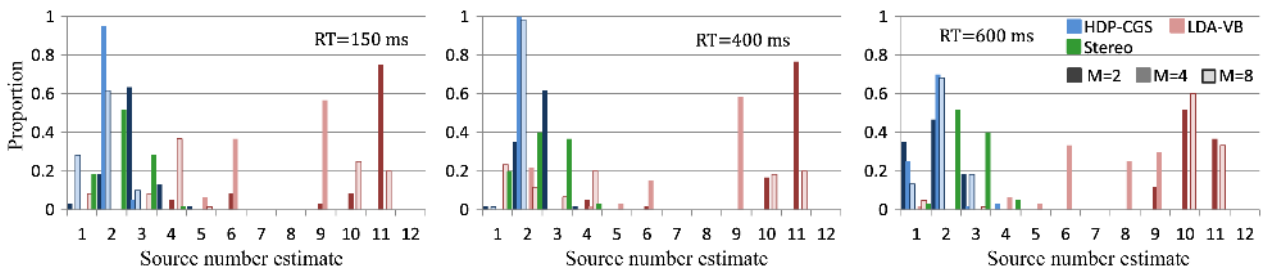


Fig. 18. Source number estimation results for recorded mixtures with two sources ($N = 2$).

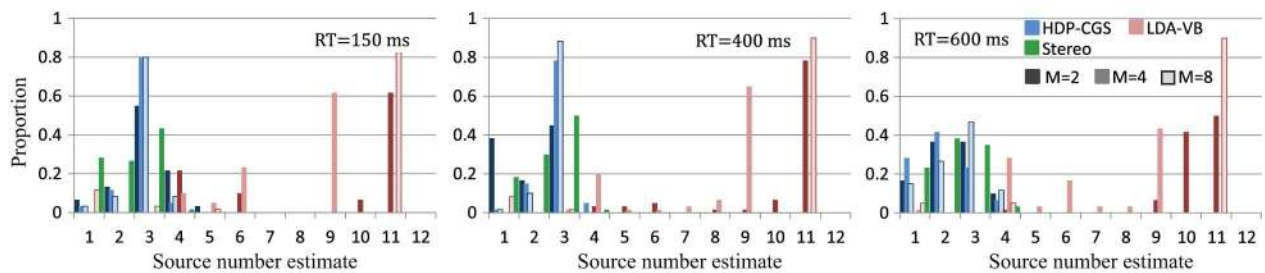


Fig. 19. Source number estimation results for recorded mixtures with three sources ($N = 3$).

cult to set a static threshold for the source counting in general setups. Three points in particular are observed. (1) VB tends to estimate more sources than CGS apparently because local optima obtained by VB have heavier tails in the posterior weights, which prevents correct source number estimation. (2) A larger number of microphones contributes to a better estimation with HDP-CGS. This means the number of microphones affects source number estimation as well as source separation quality.

HDP-CGS sometimes underestimates the source number when M is small and reverberation time is large. This is because the reverberation component is led to merge with most-weighted TF mask due to HDP prior that encourages a sparsity of activated masks. Thus, the ratio between the largest weight P_1' and the second largest weight P_2' is maximized, where the notation P' comes from Section III-C. On the other hand, Stereo can estimate a larger source number as long as the threshold for the TF mask weight is accurately configured. For the improvement of this underestimation of HDP-CGS, more sophisticated source number estimation mechanism may be necessary.

E. Discussion and Future Work

The experiments revealed that our method outperforms state-of-the-art methods in terms of separation quality. In addition, our method is capable of robust source number estimation from a multichannel mixture even in a reverberant environment thanks to the CGS.

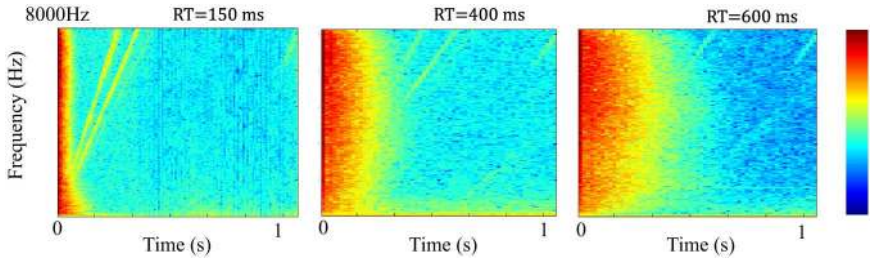


Fig. 20. Spectrograms of measured impulse response for each room.

They also revealed that reverberation particularly affects the separation quality. We may incorporate a reverberation reduction technique such as [37] for further improved performance. Our method assumes non-moving sound sources. The use of a hidden Markov model would be a natural way to cope with moving sound sources [38] as it would make the direction indicator w_k a time-series sequence. For source number estimation, a model selection approach, such as [39], may be useful.

We used the measured impulse responses from the directions we consider as prior information about the microphone array we use. Reducing the necessary prior information about the microphone array can also be enumerated as the future directions. For example, the impulse responses can be simulated from the position of microphones or obtained through more casual and automatic calibration.

V. Conclusion

Our sound source localization and separation method using a microphone array achieves the decomposition function that is essential to CASA systems in a unified manner based on hierarchical Dirichlet process. Source separation experiments using simulated and recorded mixtures under various conditions demonstrated that our method outperforms state-of-the-art methods without a priori source number knowledge. The Bayesian nonparametrics-based framework contributes to the basis of CASA systems and robot audition architectures that work in our daily environments.

Appendix
Impulse Responses for Simulated Mixtures

This appendix describes the impulse responses used to generate the simulated mixtures used in the experiment. The impulse responses were measured by recording the time-stretched pulse (TSP) signal. The TSP signal was recorded with 16000 (Hz) sampling rate. The length of the TSP signal was set 16384 points, that is, an approximately 1 (s) signal.

Fig. 5 visualizes the impulse responses measured in three rooms with different reverberant conditions in the time-frequency domain. We can confirm that the energy of the impulse response is extended along the time axis with a larger reverberation time. We can also notice that the energy of the reverberation is more concentrated on the higher frequency region than the lower range, especially in RT 400 (ms) and RT 600 (ms) rooms. This may be because the frequency characteristics of the loudspeaker used for the TSP recording or the short length of the TSP signal attenuated the reverberation of the lower frequency range.

Acknowledgment

The authors would like to thank Dr. S. Araki for providing the implementation of her source number estimation method for our experiment. The authors are also grateful to the associate editor, Prof. S. Jensen, and the anonymous reviewers for the valuable comments and helpful suggestions.

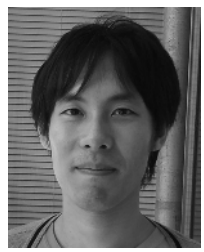
References

[1] D. F. Rosenthal and H. G. Okuno, *Computational Auditory Scene Analysis*. Mahwah, NJ, USA: Lawrence Erlbaum, 1998.
[2] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. New York, NY, USA: Wiley-IEEE Press, 2006.
[3] , P. Common and C. Jutten, Eds., *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. New York, NY, USA: Academic, 2010.
[4] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, ser. Springer Topics in Signal Processing. New York, NY, USA: Springer, 2008.
[5] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 19, no. 3, pp. 516–527, Mar. 2011.
[6] M. I. Mandel, D. P. W. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," *Adv. Neural Inf. Process. Syst.*, vol. 19, pp. 953–960, 2007.
[7] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "Design and implementation of robot audition system "HARK"," *Adv. Robot.*, vol. 24, no. 5–6, pp. 739–761, 2010.
[8] T. Otsuka, K. Ishiguro, H. Sawada, and H. G. Okuno, "Bayesian unification of sound source localization and separation with permutation resolution," in *Proc. AAAI Conf. Artif. Intel.*, 2012, pp. 2038–2045.
[9] T. Otsuka, K. Ishiguro, H. Sawada, and H. G. Okuno, "Unified auditory functions based on Bayesian topic model," in *Proc. IEEE/RSJ Int. Conf. Intel. Robot. Syst.*, 2012, pp. 2370–2376.
[10] K. Yamamoto, F. Asano , W. F. G. van Rooijen, E. Y. L. Ling, T. Yamada, and N. Kitawaki, "Estimation of the number of sound sources using support vector machines and its application to sound source separation," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2003, pp. V-485–V-488.
[11] B. Loesch and B. Yang, "Source number estimation and clustering for underdetermined blind source separation," in *Proc. Int. Workshop Acoust. Echo Noise Control*, 2008.
[12] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal Process.*, vol. 81, no. 11, pp. 2353–2362, 2001.
[13] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "A survey of convolutive blind source separation methods," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. New York, NY, USA: Springer, 2007.
[14] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 12, no. 5, pp. 530–538, Sep. 2004.
[15] I. Lee, T. Kim, and T.-W. Lee, "Fast fixed-point independent vector analysis algorithms for convolutive blind source separation," *Signal Process.*, vol. 87, no. 8, pp. 1859–1871, 2007.
[16] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2011, pp. 189–192.

[17] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation maximization source separation and localization," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 18, no. 2, pp. 382–394, Feb. 2010.
[18] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.
[19] J. Taghia, N. Mohammadia, and A. Leijon, "A variational Bayes approach to the underdetermined blind source separation with automatic determination of the number of sources," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2012, pp. 253–256.
[20] S. Winter, H. Sawada, and S. Makino, "Geometrical interpretation of the PCA subspace approach for overdetermined blind source separation," *EURASIP J. Adv. Signal Process.*, vol. 2006, pp. 1–11, 2006, article ID 71632.
[21] S. Vaseghi and H. Jetelová, "Principal and independent component analysis in image processing," in *Proc. 14th Int. Conf. Mobile Comput. Netw.*, 2006, pp. 1–5.
[22] N. Kovacevic and A. R. McIntosh, "Groupwise independent component decomposition of EEG data and partial least square analysis," *Neuroimage*, vol. 35, no. 3, pp. 1103–1112, 2007.
[23] D. Knowles and Z. Ghahramani, "Infinite sparse factor analysis and infinite independent components analysis," in *Proc. Int. Conf. Ind. Compon. Anal. Signal Separat.*, 2007, pp. 381–388.
[24] K. Nagira, T. Takahashi, T. Ogata, and H. G. Okuno, "Complex extension of infinite sparse factor analysis for blind speech separation," in *Proc. Int. Conf. Latent Variable Anal. Signal Separat.*, 2012, pp. 388–396.
[25] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical dirichlet processes," *J. Amer. Statist. Assoc.*, vol. 101, no. 476, pp. 1566–1581, 2006.
[26] H. Kameoka, M. Sato, T. Ono, N. Ono, and S. Sagayama, "Blind separation of infinitely many sparse sources," in *Proc. Int. Workshop Acoust. Signal Enhance.*, 2012, pp. 1–4.
[27] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.
[28] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 18, no. 7, pp. 1830–1840, Jul. 2010.
[29] A. van den Bos, "The multivariate complex normal distribution–A generalization," *IEEE Trans. Inf. Theory*, vol. 41, no. 2, pp. 537–539, Mar. 1995.
[30] K. Conradsen, A. A. Nielsen, J. Schou, and H. Skiriver, "A test statistic in the complex Wishart distribution and its application to change detection in polarimetric SAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 1, pp. 4–19, Jan. 2003.
[31] M. Escobar and M. West, "Bayesian density estimation and inference using mixtures," *J. Amer. Statist. Assoc.*, vol. 90, pp. 577–588, 1995.
[32] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.
[33] H. Sawada, S. Araki, and S. Makino, "MLSP 2007data analysis competition: Frequency-domain blind source separation for convolutive mixtures of speech/audio signals," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, 2007, pp. 45–50.
[34] S. Araki, T. Nakatani, H. Sawada, and S. Makino, "Stereo source separation and source counting with MAP estimation with Dirichlet prior considering spatial aliasing problem," in *Proc. 8th Int. Conf. Ind. Compon. Anal. Signal Separat.*, 2009, pp. 742–750.
[35] M. Matsumoto, M. Tohyama, and H. Yanagawa, "A method of interpolating binaural impulse responses for moving sound images," *Acoust. Sci. Technol.*, vol. 24, no. 5, pp. 284–292, 2003.
[36] K. Nakamura, K. Nakadai, and H. G. Okuno, "A real-time super-resolution robot audition system that improves the robustness of simultaneous speech recognition," *Ad*, vol. 27, no. 12, pp. 933–945, 2013.
[37] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 19, no. 1, pp. 69–84, Jan. 2011.
[38] D. MacKay, Ensemble Learning for Hidden Markov Models Dept. of Phys., Cambridge Univ., Cambridge, U.K., Tech. Rep., 1997.
[39] R. Fujimaki and S. Morinaga, "Factorized asymptotic Bayesian inferencce for mixture modeling," in *Proc. Artif. Intell. Statist.*, 2012.

Takuma Otsuka received the B.E. and M. S. in Informatics from Kyoto University, Kyoto, Japan, in 2009 and 2011, respectively.

Since 2011, he has been a Ph.D. candidate at Graduate school of Informatics, Kyoto University, and a recipient of the JSPS Research Fellowship for Young Scientists (DC1). His research interests include statistical signal processing, robot audition, statistical pattern recognition, and machine learning. He received the Best paper award of IEA/AIE-2010, NEC C&C Young Researcher Paper Award in 2010, and the Best paper award of IWSEC-2013. Mr. Otsuka is a member of RSJ and IPSJ.



Katsuhiko Ishiguro (M'09) received the B.Eng. and M.Inf. degrees from the University of Tokyo, Tokyo, Japan, in 2000 and 2004, respectively, and the Ph.D. degree from University of Tsukuba, Ibaraki, Japan, in 2010.

He has been a Researcher at the NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan, since 2006. His research interests include probabilistic models for data mining of structured data such as relational data and time series, statistical pattern recognition for multimedia data, and cognitive robotics. Dr. Ishiguro is a member of the IEICE and IPSJ.

Hiroshi Sawada (M'02–SM'04) received the B.E., M.E. and Ph.D. degrees in information science from Kyoto University, Kyoto, Japan, in 1991, 1993 and 2001, respectively.

He joined NTT Corporation in 1993. From 2009 to 2013, he was the group leader of Learning and Intelligent Systems Research Group at the NTT Communication Science Laboratories, Kyoto, Japan. He is now a senior research engineer, supervisor at the NTT Service Evolution Laboratories, Yokosuka, Japan. His research interests include statistical signal processing, audio source separation, array signal processing, machine learning, latent variable model, graph-based data structure, and computer architecture.

From 2006 to 2009, he served as an associate editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH & LANGUAGE PROCESSING. He is a member of the Audio and Acoustic Signal Processing Technical Committee of the IEEE SP Society. He received the 9th TELECOM System Technology Award for Student from the Telecommunications Advancement Foundation in 1994, the Best Paper Award of the IEEE Circuit and System Society in 2000, the MLSP Data Analysis Competition Award in 2007, and the SPIE ICA Unsupervised Learning Pioneer Award in 2013. Dr. Sawada is a member of the IEICE and the ASJ.

Hiroshi G. Okuno (M'03–SM'06–F'12) received the B.A. and Ph.D. degrees from the University of Tokyo, Tokyo, Japan, in 1972 and 1996, respectively.

He worked for NTT, JST, and the Tokyo University of Science. He is currently a Professor in the Graduate School of Informatics, Kyoto University, Kyoto, Japan. He was Visiting Scholar at Stanford University, Stanford, CA, from 1986 to 1988. He has done research in programming languages, parallel processing, and reasoning mechanisms in AI. He is currently engaged in computational auditory scene analysis, music scene analysis, and robot audition.

He coedited Computational Auditory Scene Analysis (Lawrence Erlbaum Assoc., 1998), Advanced Lisp Technology (Taylor and Francis, 2002), and New Trends in Applied Artificial Intelligence (IEA/AIE) (Springer, 2007). Prof. Okuno received various awards including the 1990 Best Paper Award of the JSAI, the Best Paper Award of IEA/AIE-2001, 2005, and 2013, and was an IEEE/RSJ IROS-2001 and 2006 Best Paper Nomination Finalist. He is a fellow of Japanese Society for Artificial Intelligence, and a member of the AAAI, ACM, ASJ and other 5 societies.