2012 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 7-12, 2012. Vilamoura, Algarve, Portugal

Unified Auditory Functions based on Bayesian Topic Model

Takuma Otsuka, Katsuhiko Ishiguro, Hiroshi Sawada and Hiroshi G. Okuno

***Abstract*— Existing auditory functions for robots such as sound source localization and separation have been implemented in a cascaded framework whose overall performance may be degraded by any failure in its subsystems. These approaches often require a careful and environment-dependent tuning for each subsystems to achieve better performance. This paper presents a unified framework for sound source localization and separation where the whole system is integrated as a Bayesian topic model. This method improves both localization and separation with a common configuration under various environments by iterative inference using Gibbs sampling. Experimental results from three environments of different reverberation times confirm that our method outperforms state-of-the-art sound source separation methods, especially in the reverberant environments, and shows localization performance comparable to that of the existing robot audition system.**

I. INTRODUCTION

Comprehending the surrounding environment through processing and analyzing of sensory input is one of the most essential functions for robots and intelligent systems [1], [2]. For example, automatic cleaning robots are equipped with haptic sensors or infrared sensors to avoid obstacles in a domestic room [3], and telepresence robots have cameras and microphones to deliver environmental awareness to their operators as remote sensing devices [4]. In many cases, these sensing functions aim to extract multiple pieces of information from the environment; for example, as an auditory function, a robot may estimate the direction of sound sources and retrieve each audio signal from the observed sound mixture. These functions are referred to as sound source localization and separation, respectively.

Such compound sensing functions are categorized into two frameworks to deal with multiple pieces of information. These two frameworks are different in their primal objectives: (1) a computational efficiency or tractability, and (2) an overall optimization of the system. Figure 1 illustrates two kinds of frameworks: A *cascaded framework* prioritizes the computational efficiency by sequentially extracting multiple pieces of information with different subsystems. While this strategy can accelerate the computation by focusing on each subsystem, the overall performance can be degraded due to any failure in subsystems because falsely extracted information is propagated into subsequent subsystems. A *unified framework* seeks for an overall optimized information extraction, sometimes at the cost of computation time due to the larger search space of the optimal solution. Especially when the targeted information is mutually dependent, the overall performance can be improved by unified systems.

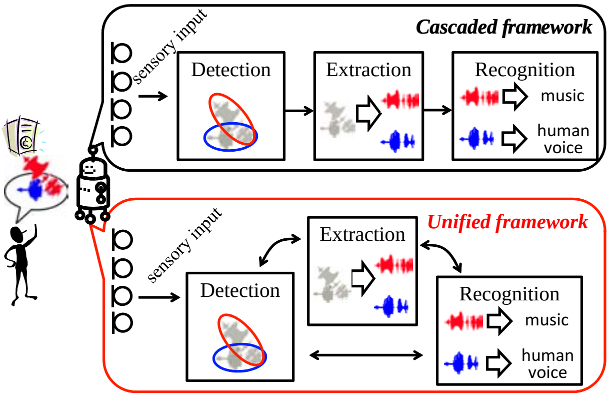

Fig. 1. An example of two frameworks: *cascaded framework* and *unified framework*. A robot is sensing a mixture of human voice and music. A typical cascaded framework may consist of irreversible subsystems, e.g., detection, extraction, and recognition. A failure in the detection of the human voice would result in incorrect recognition. In a unified framework, mutual optimization of each subsystem may improve the overall performance; the extraction of one sound may discover the other hidden sound to improve the detection subsystem, which leads to correct recognition of both sounds.

Both frameworks are necessary for robotic sensing depending on their applications. For instance, automatic mobile robots may use a cascaded framework to achieve a realtime processing for a quick motion path planning; while remote sensing probe robots working in a nuclear plant require an elaborate and precise information extraction by a unified framework at any computational cost.

This paper presents a unified framework for fundamental auditory functions: sound source localization and separation. Sound source localization provides spatial auditory information, which enables robots to construct an auditory map [5]. Sound source separation is essential in a noisy environment. For example, it enables robots to have a multiparty oral interaction with human beings [6] with microphones attached to the robot's own body or inform the telepresence robot operator of surrounding auditory events [7].

We formulate this joint problem of sound source localization and separation as a Bayesian topic model [8]; a sound source is separated by clustering the multichannel spectrogram in the time-frequency domain, and is localized by assigning the source to a certain direction. We construct a generative model of the multichannel spectrogram given multiple sound sources and derive the inference using Gibbs sampling [9] where our method copes with the uncertainty of the sound location and spectrogram simultaneously.

T. Otsuka and H. G. Okuno are with Graduate School of Informatics, Kyoto University, Sakyo, Kyoto, 606-8501, Japan {ohtsuka, okuno}@kuis.kyoto-u.ac.jp

K. Ishiguro and H. Sawada are with NTT Communication Science Laboratories, NTT Corporation, Seika-cho, Kyoto, 619-0237, Japan {ishiguro.katsuhiko, sawada.hiroshi}@lab.ntt.co.jp

978-1-4673-1735-1/12/S31.00 ©2012 IEEE

II. Unified Auditory Functions

This section presents our problem and explains the advantage of our method compared with existing methods in terms of the robustness against diverse auditory environments. Our sound source localization and separation problem is stated as follows:

Input: Multichannel audio signal,
Outputs: The direction of arrival and separated signal of each sound source,
Assumptions: The microphone array configuration is known as steering vectors, the number of sources is given, and the sound sources are still.

A steering vector conveys the time difference between sound arrivals at each sensor given a certain direction of the sound source and a frequency bin. The number of sources is important, especially in reverberant environments, where automatic estimation is still a challenge.

Our method incorporates steering vectors as prior knowledge of the microphone array. The steering vectors are assumed to be determined by (1) the relative position of microphones in a microphone array, (2) each frequency bin of the spectrogram in the time-frequency domain, and (3) the direction of arrival of a sound source. We use steering vectors measured in an anechoic chamber so as to make the system independent of reverberation that may vary among different environments[1]. These anechoic steering vectors are assumed to be environment-independent because a microphone array embedded with a robot and the window size of the Fourier transform are usually fixed. Since steering vectors are associated with the direction of a sound source, they are important information for sound source localization.

A. From Cascaded Framework to Unified Framework

The joint processing of sound source localization and separation has been separately tackled [10]–[13]. Thus, the compound problem has been tackled in a cascaded way. Since initial subsystems in a cascaded framework are often critical to its overall performance, a careful tuning dependent on the auditory environment becomes necessary.

A robot audition system HARK [13] provides both functions by first localizing multiple sound sources and then retrieving each sound source from the observed mixture on the basis of the preceding localization results where the method uses the steering vectors in both steps. While this strategy achieves a fast computation, the separation performance is severely affected by the localization quality; if the localization fails, the separation result is also degraded. Therefore, a careful parameter tuning is essential for the localization depending on the number of sound sources and the reverberation.

Some sound source separation methods dispense with steering vectors. For example, independent vector analysis (IVA) [14], [15] separates sound sources given only the multichannel mixture observation. However, the localization with IVA requires the steering vectors of the microphone array to seek the correlation between the separated signals and steering vectors of each direction. Thus, the joint processing again becomes cascaded; the localization performance is affected by the separation quality.

Furthermore, IVA has a limitation regarding the number of sources. IVA algorithm itself decomposes the given M-channel signal into M signals. When the number of sources is N, we have to select N individual sources in some way, e.g., by a channel reduction preprocessing using principal component analysis. In addition, IVA is incapable of coping with the situation where $N > M$. The auditory environment should be guaranteed to satisfy $N \leq M$ for the use of IVA.

Our method achieves a unified framework for the localization and separation problem by constructing a generative model of the multichannel mixture observation. The unified model efficiently extracts the information from the input mixture without switching environment-dependent parameters. The latent variables are inferred by Gibbs sampling where the latent variables are iteratively updated. The iterative inference procedures are qualitatively interpreted as follows: When the iteration is at a separation step, the sampling of the separation is carried out on the ground of the current localization estimate; whereas the localization sampling is based on the probabilistic hypothesis of the separation. This inference scheme based on Gibbs sampling constitutes considerable progress on our previous study [16] in that more parameters are automatically inferred from the observed data.

III. Our Method

Figure 2 outlines our method. First, the mixed signal to be observed is generated by adding sound sources as shown on the left in Fig. 2. A real-valued waveform in the time domain is converted into complex values in the time-frequency domain by a short-time Fourier transform (STFT). Then, a time-frequency mask (TF mask) is estimated for each source to retrieve it from the mixture.

Figure 2 shows power spectrograms on a linear scale to emphasize that the power is sparsely distributed in the time-frequency domain, that is, the power is nearly zero at most time-frequency points. Therefore, we can assume that only one sound source is dominant at each time-frequency point and that we are able to extract sound sources with TF masks.

The estimation of the TF masks is formulated as a clustering problem on the observed multi-channel signal in the time-frequency domain. Each time-frequency point stems from a certain source referred to as a class in the clustering context. Our method estimates the posterior probability to which class each time-frequency point belongs. Furthermore, our method handles more classes than the actual number of sound sources for the clustering to make our algorithm independent of the number of actual sound sources. We set parameters s.t. redundant classes shrink during the clustering and we obtain stable results regardless of the source number. Note that we can handle more sources than the number of

[1] In practice, steering vectors measured in a reverberant environment do not fully contain the information of the reverberation because the window size of the Fourier transform is usually much shorter than the reverberation.

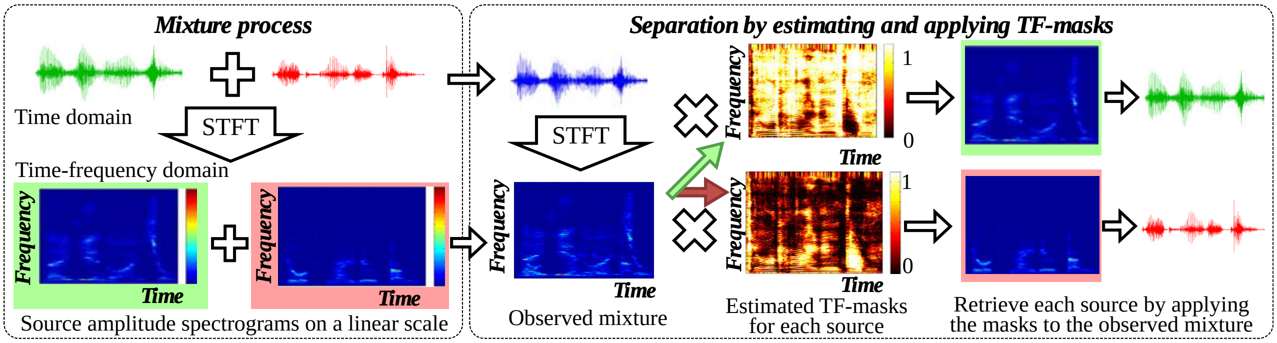
Mixture process
Time domain
STFT
Time-frequency domain
Frequency
Time
Source amplitude spectrograms on a linear scale
Separation by estimating and applying TF-masks
STFT
Observed mixture
1
0
Estimated TF-masks for each source
Retrieve each source by applying the masks to the observed mixture

Fig. 2. Illustration of mixture process and separation method based on time-frequency masking

microphones because the number of classes is not necessarily capped at the number of microphones.

Our method resembles that of Mandel et al. [17] in that time-frequency points are clustered into each source to generate TF masks. While Mandel et al. [17] use only the phase of complex values and 2 microphones, our method uses full parts of the complex values and allows more than two microphones for better results.

Sections III-A–III-C explain the generative model, and Section III-D presents the inference steps. Table I shows the notations we use. A set of variables is denoted with a tilde without subscripts, e.g., $\tilde{\mathbf{x}} = \{\mathbf{x}_{tf} | 1 \le t \le T, 1 \le f \le F\}$.

TABLE I
NOTATIONS

Symbol	Meaning
t	Time frame ranging 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 t and f
β	Global class ratio
$\boldsymbol{\pi}_t$	Class ratio at time t
w_k	Direction indicator for class k
$\boldsymbol{\varphi}$	Direction ratio for all classes
λ_{tfk}	Inverse power of class k at t and f
$\mathbf{H}_{fd}$	Inverse covariance of direction d at frequency f
n_{tk}	The number of samples of class k at time frame t
m_d	The number of samples of class d

A. Time-varying Covariance Matrix-Based Observation

We employ the covariance model [18] for the likelihood function of the signal in the time-frequency domain; each sample follows a complex normal distribution with zero mean and time-varying covariance. Figure 3 shows a scatter plot of the two-channel observations of two sources in blue and red. These samples are generated as follows: let s_{tfk} and $\mathbf{q}_{fd}$ denote the signal of the kth class at time t and frequency f, and the steering vector from direction d where class k is located, respectively. Then, the signal is observed as $\mathbf{x}_{tf} = s_{tfk}\mathbf{q}_{fd}$, where the elements of $\mathbf{x}_{tf}$ are the observation of each microphone. The covariance is

$$\mathbb{E}[\mathbf{x}_{tf}\mathbf{x}_{tf}^H] = \mathbb{E}[|s_{tfk}|^2\mathbf{q}_{fd}\mathbf{q}_{fd}^H], \quad (1)$$

where $\cdot^H$ means a Hermitian transpose.

As shown in Figure 3, the covariance matrix of each source has an eigenvector with a salient eigenvalue. This vector corresponds to the steering vector associated with the direction in which the source is located. This is the mutual dependence between the localization and separation. That is, the clustering of each sample corresponds to the separation of sound sources, and the investigation of the eigenvectors of the clustered covariances means the localization of sources.

The covariance is factorized into a power term and a steering matrix. While the power of the signal $|s_{tfk}|^2$ is time-varying in Eq. (1), the steering term $\mathbf{q}_{fd}\mathbf{q}_{fd}^H$ is fixed over time since we assume steady sources. Because we can assume s_{tfk} and $\mathbf{q}_{fd}$ are independent, we introduce an inverse power $\lambda_{tfk} \approx |s_{tfk}|^{-2}$ and an inverse steering matrix $\mathbf{H}_{fd} \approx (\mathbf{q}_{fd}\mathbf{q}_{fd}^H + \varepsilon\mathbf{I}_M)^{-1}$, where $\mathbf{I}_M$ is the $M \times M$ identity matrix. The likelihood distribution is

$$p(\tilde{\mathbf{x}}|\tilde{z},\tilde{w},\tilde{\lambda},\tilde{\mathbf{H}}) = \prod_{tfkd} \mathcal{N}_{\mathbb{C}}(\mathbf{x}_{tf}|\mathbf{0},(\lambda_{tfk}\mathbf{H}_{fd})^{-1})^{z_{tf}^k w_k^d}, \quad (2)$$

where $\mathcal{N}_{\mathbb{C}}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1}) = \frac{|\boldsymbol{\Lambda}|}{(2\pi)^M}\exp\left(-\mathbf{x}^H\boldsymbol{\Lambda}\mathbf{x}\right)$ is the probability density function (pdf) of the complex normal distribution [19] with a mean $\boldsymbol{\mu}$ and precision $\boldsymbol{\Lambda}$. $\prod_{tfkd}$ means a product over all ranges of t, f, k, and d. Note that z_{tf} and w_k indicate the class of $\mathbf{x}_{tf}$ and the direction of class k, and range from 1 to K and from 1 to D, respectively, where $z_{tf}^k = 1$ iff. $z_{tf} = k$ and $z_{tf}^k = 0$ when $z_{tf} \neq k$. Similarly, $w_k^d = 1$ iff. $w_k = d$. By placing these binary variables in the exponential part of the likelihood function and calculating the product over all possible k and d, Eq. (2) provides the likelihood given the class and its direction. $|\boldsymbol{\Lambda}|$ is the determinant of the matrix $\boldsymbol{\Lambda}$.

We adopt conjugate priors for parameters λ_{tfk} and $\mathbf{H}_{fd}$:

$$p(\tilde{\lambda}) = \prod_{tfk} \mathcal{G}(\lambda_{tfk}|a_0,b_{tf}), \quad (3)$$

$$p(\tilde{\mathbf{H}}) = \prod_{fd} \mathcal{W}_{\mathbb{C}}(\mathbf{H}_{fd}|\nu_0,\mathbf{G}_{fd}), \quad (4)$$

where $\mathcal{G}(\lambda|a,b) \propto \lambda^{a-1}e^{-b\lambda}$ denotes the pdf of a gamma distribution with a shape a and inverse scale b, and $\mathcal{W}_{\mathbb{C}}(\mathbf{H}|\nu,\mathbf{G}) = \frac{|\mathbf{H}|^{\nu-M}\exp\left\{-\mathrm{tr}(\mathbf{H}\mathbf{G}^{-1})\right\}}{|\mathbf{G}|^{\nu}\pi^{M(M-1)/2}\prod_{i=0}^{M-1}\Gamma(\nu-i)}$ is the pdf of a complex Wishart distribution [20]. $\mathrm{tr}(\mathbf{A})$ is the trace of $\mathbf{A}$ and $\Gamma(x)$ is the gamma function. The configuration of the hyperparameters is explained in Section III-D.

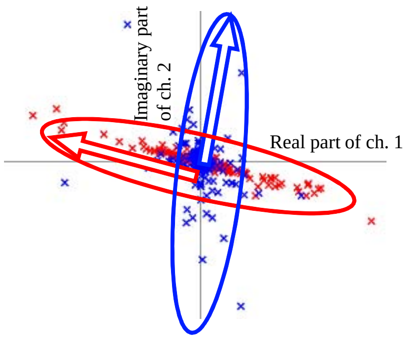

Fig. 3. Plot of complex-valued multichannel signals at 3000 (Hz). The colors represent respective sound sources.

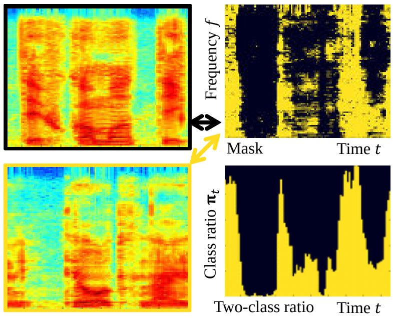

Fig. 4. TF mask for two sources (top right) and their ratio for each time frame (bottom right). Left: original log-scale power spectrograms of the two sources in black and yellow.

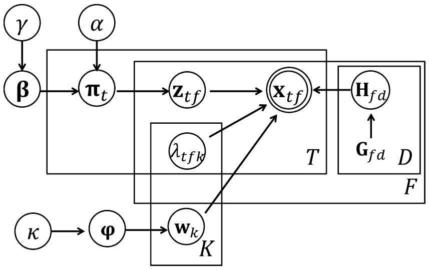

Fig. 5. Graphical representation of our model with the plate notation. Rectangle boxes indicate the repetition of the variables; the label T means the index t ranges from 1 to T inside the box.

B. Permutation Resolution by Latent Dirichlet Allocation

Since the clustering is independently carried out for each frequency bin, we must identify a class from each frequency bin that corresponds to the same sound source. This is called permutation resolution [21]. We use a topic model called latent Dirichlet allocation (LDA) [8] to incorporate the permutation resolution into our unified framework.

The bottom right image in Figure 4 shows how dominant each source is at time frames. The black source is dominant in some time frames whereas the yellow source is dominant in others. We can expect to resolve the permutation by preferring one or several classes for each time frame in a way similar to that used by Sawada et al. [22] to seek the synchronization of the sound dominance over frequency bins.

LDA is used to introduce the ratio of classes. In the context of document analysis, LDA infers the topic of documents containing many words from a document set by assigning each word to a certain topic. We regard the topic as a sound source, the document as a time frame, and the words as frequency bins.

Let $\boldsymbol{\pi}_t$ denote the class ratio at time t. The class indicator variable z_{tf} in Eq. (2) determines to which class $\mathbf{x}_{tf}$ belongs in accordance with $\boldsymbol{\pi}_t$ as:

$$p(\tilde{z}|\tilde{\boldsymbol{\pi}}) = \prod_{tfk} \pi_{tk}^{z_{tf}^{k}}, \tag{5}$$

where $\boldsymbol{\pi}_t$ follows a conjugate prior Dirichlet distribution:

$$\begin{aligned} p(\tilde{\boldsymbol{\pi}}|\boldsymbol{\beta}) &= \prod_t \mathcal{D}(\boldsymbol{\pi}_t|\alpha\boldsymbol{\beta}) \\ &= \prod_t \frac{\Gamma(\alpha\beta_\cdot)}{\prod_k \Gamma(\alpha\beta_k)} \prod_k \pi_{tk}^{\alpha\beta_k - 1}, \end{aligned} \tag{6}$$

where the subscript $\cdot$ denotes the summation over the specified index, i.e., $\beta_\cdot = \sum_k \beta_k$.

The global class ratio $\boldsymbol{\beta}$ again follows Dirichlet distribution given a positive parameter γ.

$$p(\boldsymbol{\beta}|\gamma) = \mathcal{D}(\boldsymbol{\beta}|\gamma\mathbf{1}_K) = \frac{\Gamma(K\gamma)}{\prod_k \Gamma(\gamma)} \prod_k \beta_k^{\gamma-1}, \tag{7}$$

where $\mathbf{1}_K$ is a K-dimensional vector whose elements are all 1. The parameters γ and α follows gamma distributions; $\mathcal{G}(\gamma|a_\gamma, b_\gamma)$ and $\mathcal{G}(\alpha|a_\alpha, b_\alpha)$, respectively.

C. Latent Variable for Localization

A discrete variable w_k is introduced to localize each sound source; when $w_k = d$, it indicates that the class k is located in the direction d. The directions of the sound sources are made discrete in this model to simplify the inference process.

The indicator w_k is dependent on the direction ratio $\boldsymbol{\varphi}$ that follows a Dirichlet distribution:

$$p(\tilde{w}|\boldsymbol{\varphi}) = \prod_{kd} \varphi_d^{w_k^d}, \tag{8}$$

$$p(\boldsymbol{\varphi}|\kappa) = \mathcal{D}(\boldsymbol{\varphi}|\kappa\mathbf{1}_D) = \frac{\Gamma(D\kappa)}{\prod_d \Gamma(\kappa)} \prod_d \varphi_d^{\kappa-1}. \tag{9}$$

Note that we use a symmetric Dirichlet distribution with a concentration parameter κ because we have no prior knowledge about the spatial position of the sound sources. Similarly to γ, κ follows a gamma distribution $\mathcal{G}(\kappa|a_\kappa, b_\kappa)$.

D. Inference

Figure 5 depicts the probabilistic dependency; the double-circled $\mathbf{x}_{tf}$ is the observation, the circled symbols are latent probability variables, and the plain symbols are fixed values. The inference procedures are summarized in Algorithm 1.

Some Dirichlet parameters $\boldsymbol{\pi}$ and $\boldsymbol{\varphi}$ are integrated out from the posterior distribution to accelerate the convergence of the sampling. The integration is analytically tractable thanks to the conjugacy of Dirichlet distribution. From Eqs. (5, 6, 8, 9), we obtain the integrated distribution as

$$\begin{aligned} p(\tilde{z}, \tilde{w}|\boldsymbol{\beta}, \alpha, \gamma, \kappa) = &\prod_t \left(\frac{\Gamma(\alpha)}{\Gamma(\alpha + n_{t\cdot})} \prod_k \frac{\Gamma(\alpha\beta_k + n_{tk})}{\Gamma(\alpha\beta_k)} \right) \\ &\frac{\Gamma(D\kappa)}{\Gamma(D\kappa + m_\cdot)} \prod_d \frac{\Gamma(\kappa + m_d)}{\Gamma(\kappa)}, \end{aligned} \tag{10}$$

where n and m mean that by definition, $n_{tk} = \sum_f z_{tf}^k$ and $m_d = \sum_k w_k^d$. $n_{t\cdot} = \sum_k n_{tk}$ and $m_\cdot = \sum_d m_d$. Here, the latent variables are drawn from the posterior distribution shown in Eq. (11) by Gibbs sampling.

$$p(\tilde{z}, \tilde{w}, \tilde{\lambda}, \tilde{\mathbf{H}}, \boldsymbol{\beta}, \gamma, \alpha, \kappa|\tilde{\mathbf{x}}). \tag{11}$$

a) Sampling indicators: The class of each time-frequency point z_{tf} is iteratively sampled from

$$\begin{aligned} &p(z_{tf} = k'|\tilde{\mathbf{x}}, \tilde{z}^{-tf}, \tilde{w}, \tilde{\lambda}, \tilde{\mathbf{H}}, \boldsymbol{\beta}, \gamma, \alpha, \kappa) \propto \\ &(\alpha\beta_{k'} + n_{tk'}^{-tf}) \prod_d \left\{ \lambda_{tfk'}^M |\mathbf{H}_{fd}| \exp\left(-\mathbf{x}_{tf}^H \lambda_{tfk'} \mathbf{H}_{fd} \mathbf{x}_{tf}\right) \right\}^{w_k^d}, \end{aligned} \tag{12}$$

where $\tilde{z}^{-tf}$ denotes a set of z except z_{tf} and n_{tk}^{-tf} is the count of class k samples at time t excluding z_{tf}. The first term in the rhs. of Eq. (12) derives from Eq. (10).

The direction of each class w_k is similarly sampled from

$$p(w_k = d' | \tilde{\mathbf{x}}, \tilde{z}, \tilde{w}^{-k}, \tilde{\lambda}, \tilde{\mathbf{H}}, \boldsymbol{\beta}, \gamma, \alpha, \kappa) \propto \tag{13}$$
$$(\kappa + m_{d'}^{-k}) \prod_{tfk} \left\{ \lambda_{tfk}^{M} |\mathbf{H}_{fd'}| \exp\left(-\mathbf{x}_{tf}^{H} \lambda_{tfk} \mathbf{H}_{fd'} \mathbf{x}_{tf}\right) \right\}^{z_{tf}^{k}},$$

where $\tilde{w}^{-k}$ denotes a set of w except w_k and m_d^{-k} is the count of direction d excluding w_k. The first term in the rhs. of Eq. (13) also derives from Eq. (10).

b) Sampling parameters: Thanks to the conjugacy of the gamma and Wishart distribution, the parameters λ and $\mathbf{H}$ are straightforwardly sampled from the posterior distribution: $\lambda_{tfk} \sim \mathcal{G}(\lambda | \hat{a}_{tfk}, \hat{b}_{tfk})$ and $\mathbf{H}_{fd} \sim \mathcal{W}_{\mathbb{C}}(\mathbf{H} | \hat{\nu}_{fd}, \hat{\mathbf{G}}_{fd})$, where

$$\hat{a}_{tfk} = a_0 + z_{tf}^{k} M, \qquad \hat{b}_{tfk} = b_{tf} + z_{tf}^{k} \mathbf{x}_{tf}^{H} \mathbf{H}_{fd} \mathbf{x}_{tf},$$
$$\hat{\nu}_{fd} = \nu_0 + \sum_{tk} z_{tf}^{k} w_k^{d}, \quad \hat{\mathbf{G}}_{fd}^{-1} = \mathbf{G}_{fd}^{-1} + \sum_{tk} z_{tf}^{k} w_k^{d} \lambda_{tfk} \mathbf{x}_{tf} \mathbf{x}_{tf}^{H}.$$

The other parameters $\boldsymbol{\beta}, \alpha, \gamma$, or κ are sampled by incorporating auxiliary variables. The sampling of $\boldsymbol{\beta}$ is explained by Teh et al. [23]. The parameters α, γ, κ are drawn from respective posterior gamma distributions using auxiliary variables as explained by Escobar and West [24].

c) Hyperparameters: Hyperparameters of λ and $\mathbf{H}$ are set as follows: $a_0 = 1$, $b_{tf} = \mathbf{x}_{tf}^{H} \mathbf{x}_{tf}/M$, $\nu_0 = M$, $\mathbf{G}_{fd} = (\mathbf{q}_{fd} \mathbf{q}_{fd}^{H} + \varepsilon \mathbf{I}_M)^{-1}$. The gamma parameter b_{tf} reflects the power of the observation and the Wishart parameter $\mathbf{G}_{fd}$ 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 $\varepsilon = 0.001$ to allow the inverse operation. Hyperparameters of the gamma parameters α, γ, κ are commonly set as: $a_{\{\alpha,\gamma,\kappa\}} = 1$ and $b_{\{\alpha,\gamma,\kappa\}} = 1$.

d) Initialization: The inference begins by initializing w_k then z_{tf} by the following distributions.

$$p_{init}(w_k = d) \propto \begin{cases} 1 & (k-1)D/K \le d < kD/K, \\ 0 & \text{otherwise}, \end{cases} \tag{14}$$
$$p_{init}(z_{tf} = k | \tilde{w}) \propto \exp\left(-\mathbf{x}_{tf}^{H} \mathbf{G}_{f w_k} \mathbf{x}_{tf}\right) \tag{15}$$

Here, Eq. (14) means that w_k is selected at random among the designated directions for class k, where D/K directions are equally assigned to each class. Equation (15) means that the initialization of z_{tf} is weighted by the direction of each class chosen at random by Eq. (14). All the other parameters are initialized by drawing from respective prior distributions.

e) Localization and separation: After obtaining the samples of latent variables by Algorithm 1, N sources are localized and separated using the posterior probability of direction and class indicators w_k and z_{tf}. Let ξ_{tfk} denote the posterior probability of time-frequency point t, f being class k, and η_{kd} denote that of class k being located in direction d. When we have I samples of $z_{tf}^{(i)}$ with i being the sample index, ξ_{tfk} is calculated as $\xi_{tfk} = \sum_{i=1}^{I} z_{tf}^{k(i)}/I$, where $z_{tf}^{k(i)} = 1$ iff. $z_{tf}^{(i)} = k$. Similarly, $\eta_{kd} = \sum_{i=1}^{I} w_k^{d(i)}/I$.

Algorithm 1 Inference procedures

```
Initialize $\tilde{w}$ and $\tilde{z}$ by Eqs. (14, 15)
Initialize the other latent parameters by their prior distributions
repeat
    Draw $\tilde{\lambda}$ from $\mathcal{G}(\lambda|\hat{a}_{tfk}, \hat{b}_{tfk})$
    Draw $\tilde{\mathbf{H}}$ from $\mathcal{W}_{\mathbb{C}}(\mathbf{H}|\hat{\nu}_{fd}, \hat{\mathbf{G}}_{fd})$
    for all $1 \le t \le T$ and $1 \le f \le F$ do
        Draw $z_{tf}$ from Eq. (12)
    end for
    for all $1 \le k \le K$ do
        Draw $w_k$ from Eq. (13)
    end for
    Draw $\boldsymbol{\beta}$
    Draw $\alpha, \gamma$, and $\kappa$
until Samples converge to the posterior in Eq. (11)
```

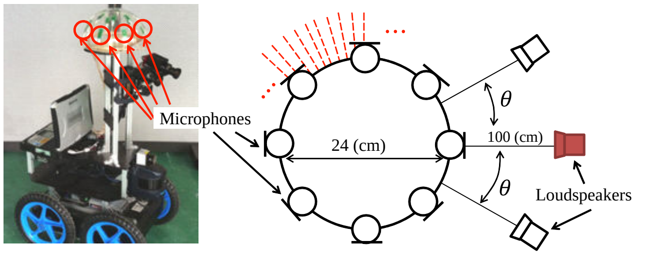

Fig. 6. Left: our robot with a circular microphone array embedded in its head part. Right: experimental setup of the speaker locations and of the measurement of steering vectors in red dotted lines from top view.

Before each sound source is extracted, class index k is sorted in descending order of the total weight of each class calculated as $\sum_{tf} \xi_{tfk}$, so as to extract more weighted classes as the sound sources. Here, the spectrogram of the nth sound source $\hat{\mathbf{x}}_{tf}^{n}$ is extracted as follows:

$$\hat{\mathbf{x}}_{tf}^{n} = \frac{\xi_{tfn}}{\sum_{k'=1}^{N} \xi_{tfk'}} \mathbf{x}_{tf}. \tag{16}$$

The posterior ξ_{tfk} are normalized within a given number of sources N to obtain better separation quality. The direction of the nth sound $\hat{d}_n$ is obtained as

$$\hat{d}_n = \underset{d'}{\operatorname{argmax}}\, \eta_{kd'}. \tag{17}$$

IV. Experimental Results

This section presents localization and separation results obtained with mixture audio signals captured by an eight-channel microphone array embedded in a robot head, as shown in Figure 6 on the left. The localization results of our method are compared with those of robot audition software HARK [13]. The separation results of our method are compared with those of HARK and IVA [15]. While our method is capable of dealing with the cases where $N > M$, this paper omits the results with $N > M$ because the methods for comparison do not cope with these cases. Experiments where $N = 3$ and $M = 2$ are presented in our previous work [16].

A. Experimental Setups

Figure 6 shows an eight-channel circular microphone array and the location of the speakers, i.e., $M = 8$. Steering vectors are measured with a 5° resolution around the microphone

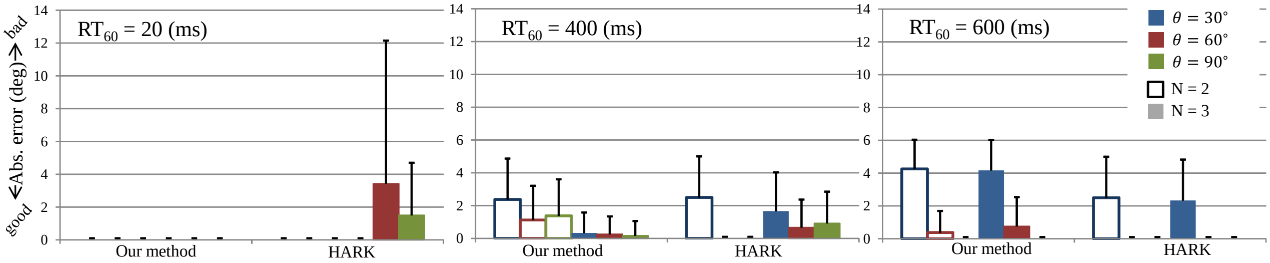

Fig. 7. Absolute localization errors of two methods in degree. Smaller values mean better localization results. The bars are the mean values and the segments are the standard deviations. The color of the bar represents the interval between the sources and the pattern represents the number of sources.

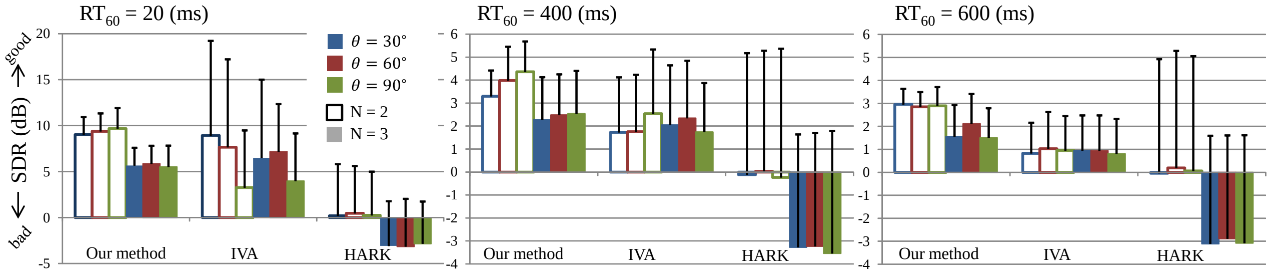

Fig. 8. Separation scores of three methods in signal to distortion ratio (SDR). Larger values mean better separation results. The bars are the mean values and the segments are the standard deviations. The color of the bar represents the source interval and the pattern represents the number of sources.

array in an anechoic chamber, as drawn with dotted red lines in Fig. 6. Mixture signals consist of two or three speakers, i.e., $N = 2$ or 3. These signals are convolutive simulations of three rooms; the anechoic chamber and two kinds of lecture rooms whose reverberation times are $RT_{60} = 20, 400, 600$ (ms), respectively.

As illustrated in Figure 6, two or three speakers are placed 100 (cm) from the array at an interval $\theta = 30, 60, 90°$. When two sources are present, the central source, shown in red in Fig. 6, is omitted. Thus, the interval becomes $2 \times \theta$ when $N = 2$. Under all conditions, the clustering is carried out with $K = 36$. The sampling process in Algorithm 1 is iterated 70 times. The samples drawn in the last 50 iterations are used for the localization and separation discarding the first 20 iterations considered to contain the initialization bias.

For each condition, 20 mixtures are generated from JNAS phonetically balanced Japanese utterances. The reverberant environment is simulated by a convolution operation; impulse responses of each environment measured by the microphone array are convoluted into the clean speach signals. The speakers on the two sides in Fig. 6 are male, and the center speaker is female. The sampling rate of the audio signals is 16000 (Hz) and an STFT is carried out with a 512 (pt) window and a 128 (pt) shift size.

While our method uses the same configurations for all audio signals, HARK changes some parameters depending on conditions. Since the localization step is critical in the HARK system, some thresholds for the localization are manually modified so that the localization is optimized for each reverberation time and the number of speakers N. To extract N sources by IVA, the M-channel signal are preprocessed into the N-channel signal by principal component analysis. Then, IVA is applied to this N-channel signal.

B. Results

Figure 7 shows the absolute localization errors of our method and HARK for three reverberation times. The bars are the mean errors for 20 utterances under each condition and the segments are their standard deviation. The bar color represents the speaker intervals, and the pattern represents the number of sources.

While, no error is found in the anechoic chamber where $RT_{60} = 20$ (ms) with our method, some errors are reported with HARK when $N = 3$. The error in the anechoic environment with HARK is caused by falsely detected sources. The error in the reverberant rooms for both our method and HARK is less than the localization resolution 5°. Note that the localization performance of our method is comparable to that of HARK even though HARK requires a manual parameter tuning specific to the environment.

Figure 8 shows the separation results of our method, IVA, and HARK. The separation performance is evaluated in terms of the signal to distortion ratio (SDR) [25], which measures the retrieval quality of sound sources from their mixture. The larger the SDR value, the better the separation results. Our method outperforms the other methods, especially in the reverberant environments. The standard deviations of our method are less than those of IVA and HARK. This means IVA and HARK tend to extract lower numbers, only one in many cases, of sound sources than our method. In other words, our method is capable of extracting all sources with equal quality even in reverberant environments whereas IVA and HARK fails to separate some sources from their mixture.

Comparison of the separation scores with our method and those with HARK confirms the superiority of our unified framework to the cascaded approach. However, while our method constantly outperforms the separation performances of existing methods, localization errors with our method are worse than those of HARK when $RT_{60} = 600$ (ms). These results imply that the clustering of z_{tf} for the separation is correct whereas the assignment of w_k is sometimes affected by the mismatch between the observed signal with reverberation and the anechoic steering vectors.

While the experiment used only human voice signals, our method is capable of handling other type of sound sources. This is because the method has no assumption with regard to the shape of the spectrogram. In fact, we confirmed that our method is able to separate a music audio signal and a frogs' chorus sound from their mixture.

Our method has two limitations: (1) SDR scores are degraded for larger number of sources N. This is a limitation of TF mask-based separation methods: larger N causes more conflicts of energy of multiple sources at each time-frequency point, hence the separation is affected. (2) While HARK is capable of realtime computation, our method takes approximately one minute given a five-second audio signal in our implementation where our system is implemented with C++ on Linux with 2.40 GHz processors.

Future work includes accelerating the inference by, for example, checking the convergence of the sampling to omit redundant iterations. Alleviation of the assumptions of our method is also enumerated as future work; e.g., the inference of the source number N from the observation data using a nonparametric Bayesian model [23], or the tracking and separation of moving sound sources. Furthermore, the implementation of our method as a module of the open source software HARK [13] is another future work so as to boost the robot audition research.

V. Conclusion

This paper presented a unified framework for sound source localization and separation to optimize both processes at the same time. A generative model for the joint problem was constructed and the inference was carried out in a Bayesian manner using Gibbs sampling. Experimental results confirmed that the our unified method outperforms state-of-the-art methods in terms of the separation performance without environment-specific parameter tunings.

References

[1] S. Thrun, "Probabilistic Robotics," *Comm. of the ACM*, vol. 45, no. 3, pp. 52–57, 2002.

[2] T. Nakamura, T. Nagai, and N. Iwahashi, "Multimodal Categorization by Hierarchical Dirichlet Process," in *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2011, pp. 1520–1525.

[3] J. Forlizzi and C. DiSalvo, "Service Robots in the Domestic Environment: A Study of the Roomba Vacuum in the Home," in *Proc. of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*, 2006, pp. 258–265.

[4] Willow Garage Inc., "Texai Remote Presence System," http://www.willowgarage.com/pages/texai/overview.

[5] Y. Sasaki, S. Kagami, and H. Mizoguchi, "Online Short-Term Multiple Sound Source Mapping for a Mobile Robot by Robust Motion Triangulation," *Advanced Robotics*, vol. 23, no. 1–2, pp. 145–164, 2009.

[6] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Active Audition for Humanoid," in *Proc. of 17th National Conference on Artificial Intelligence*, 2000, pp. 832–839.

[7] T. Mizumoto, T. Yoshida, K. Nakadai, R. Takeda, T. Otsuka, T. Takahashi, and H. G. Okuno, "Design and Implementation of Selectable Sound Separation on a Texai Telepresence System using HARK," in *Proc. of IEEE/RAS International Conference on Robotics and Automation*, 2011, pp. 2130–2137.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[9] T. Griffiths and M. Steyvers, "Finding Scientific Topics," *Proc. of the National Academy of Sciences*, vol. 101, pp. 5228–5235, 2004.

[10] A. Badali, J.-M. Valin, F. Michaud, and P. Aarabi, "Evaluating Real-time Audio Localization Algorithms for Artificial Audition in Robotics," in *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009, pp. 2033–2038.

[11] F. Asano and H. Asoh, "Joint Estimation of Sound Source Location and Noise Covariance in Spatially Colored Noise," in *Proc. of 19th European Signal Processing Conference*, 2011, pp. 2009–2013.

[12] S. Brière, D. Létourneau, M. Fréchette, J.-M. Valin, and F. Michaud, "Embedded and Integrated Audition for a Mobile Robot," in *Proc. AAAI Fall Symposium Workshop Aurally Informed Performance: Integrating Machine Listening and Auditory Presentation in Robotic Systems*, 2006, pp. 6–10.

[13] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, H. Yuji, and H. Tsujino, "Design and Implementation of Robot Audition System "HARK"," *Advanced Robotics*, vol. 24, no. 5–6, pp. 739–761, 2010.

[14] I. Lee, T. Kim, and T.-W. Lee, "Fast Fixed-point Independent Vector Analysis Algorithms for Convolutive Blind Source Separation," *Signal Processing*, vol. 87, no. 8, pp. 1859–1871, 2007.

[15] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2011, pp. 189–192.

[16] T. Otsuka, K. Ishiguro, H. Sawada, and H. G. Okuno, "Bayesian Unification of Sound Source Localization and Separation with Permutation Resolution," in *Proc. of AAAI Conference on Artificial Intelligence*, 2012, to appear.

[17] M. I. Mandel, D. P. W. Ellis, and T. Jebara, "An EM Algorithm for Localizing Multiple Sound Sources in Reverberant Environments," *Advances in Neural Information Processing Systems*, vol. 19, 2007.

[18] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-Determined Reverberant Audio Source Separation Using a Full-Rank Spatial Covariance Model," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[19] A. van den Bos, "The Multivariate Complex Normal Distribution–A Generalization," *IEEE Trans. on Information Theory*, vol. 41, no. 2, pp. 537–539, 1995.

[20] 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. on Geoscience and Remote Sensing*, vol. 41, no. 1, pp. 4–19, 2003.

[21] 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. on Audio, Speech, and Language Processing*, vol. 12, no. 5, pp. 530–538, 2004.

[22] H. Sawada, S. Araki, and S. Makino, "Underdetermined Convolutive Blind Source Separation via Frequency Bin-Wise Clustering and Permutation Alignment," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516–527, 2011.

[23] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[24] M. Escobar and M. West, "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, vol. 90, pp. 577–588, 1995.

[25] E. Vincent, R. Gribonval, and C. Févotte, "Performance Measurement in Blind Audio Source Separation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.