IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 17, 2021, accepted February 18, 2021, date of publication March 8, 2021, date of current version March 15, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3064305

Robust Auditory Functions Based on Probabilistic Integration of MUSIC and CGMM

YOSHIAKI BANDO[1], (Member, IEEE), YOSHIKI MASUYAMA[1,2], (Graduate Student Member, IEEE), YOKO SASAKI[1], (Member, IEEE), AND MASAKI ONISHI[1]

[1]National Institute of Advanced Industrial Science and Technology, Tokyo 100-8921, Japan
[2]Department of Intermedia Art and Science, Waseda University, Tokyo 169-8050, Japan

Corresponding author: Yoshiaki Bando (y.bando@aist.go.jp)

This work was supported in part by the Japan Science and Technology Agency (JST) ACT-X under Grant JPMJAX200N, and in part by the New Energy and Industrial Technology Development Organization (NEDO).

ABSTRACT Sound source localization and separation are essential functions for robot audition to comprehend acoustic environments. The widely-used multiple signal classification (MUSIC) can precisely estimate the directions of arrival (DoAs) of multiple sound sources if its hyperparameters are selected appropriately depending on the surrounding environment. A popular separation method based on a complex Gaussian mixture model (CGMM), on the other hand, can extract multiple sources even in noisy environments if its latent variables are properly initialized to avoid bad local optima. To overcome the drawbacks of both the MUSIC and CGMM, we propose a robot audition framework that complementarily combines the MUSIC and CGMM in a probabilistic manner. Our method is based on a variant of the CGMM conditioned by the localization results of MUSIC. The hyperparameters of MUSIC are estimated by the type II maximum likelihood estimation of the CGMM, and the CGMM itself is efficiently initialized and regularized by using the localization results of MUSIC. Experimental results show that our method outperformed conventional localization and separation methods even when the number of sound sources is unknown. we also demonstrate that our method can work even with moving sound sources in real time.

INDEX TERMS Robot audition, multichannel signal processing, sound source localization, sound source separation, Bayesian signal processing.

I. INTRODUCTION

Robot audition, which computationally comprehends acoustic environments [1]–[4], is an essential function for robots working in our everyday lives. A service robot communicating with humans, for example, has to understand what the customer is saying in a crowded noisy shop. A rescue robot searching for victims by detecting faint voices or other sounds needs to understand acoustic scenes. Such a robot has to be equipped with a computational audition system enabling it to comprehend when, where, and which kind of a sound event happens.

The construction of a typical robot audition system is based on sound source localization and separation to recognize multiple sound sources from a mixture observation [5]–[8]. Such systems have been based on a cascading strategy; they firstly localize sound sources from a multichannel observation and then separate the sound sources by using the localization results. For the localization, multiple signal classification (MUSIC) [9], [10] and steered response power with phase transform (SRP-PHAT) [11] have been extensively utilized. For the separation, adaptive beamformers and blind source separation (BSS) methods constrained by the localization results have been widely used [12], [13]. These systems can work in real time on a low-resource computer (e.g., a laptop computer) by combining the individual modules. This combination has enabled various applications such as humanoid robots [14]–[16], search-and-rescue drones [17], and tele-existence robots [18].

A major problem of the cascading systems is that when the source localization fails, the subsequent separation also severely deteriorates. For example, MUSIC-based localization, which is known for its high spatial resolution, requires hyperparameters such as thresholding parameters and the

The associate editor coordinating the review of this manuscript and approving it for publication was Ananya Sen Gupta.

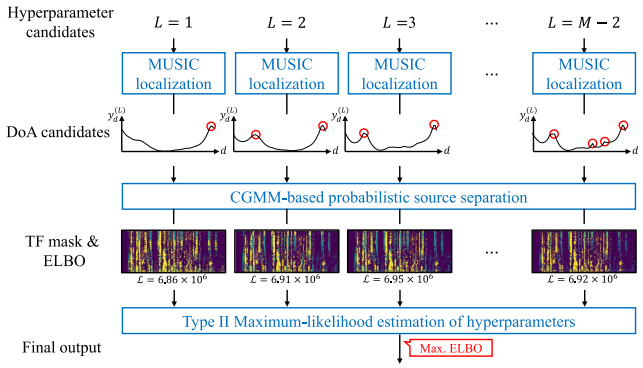

FIGURE 1. Overview of proposed MUSIC-CGMM.

number of sound sources. Because these parameters are critical to the performance, MUSIC (and the subsequent separation) often deteriorate in unknown environments where the hyperparameters cannot be optimized in advance.

To overcome the limitation of the cascading approach, statistical methods unifying sound source localization and separation have been studied [19], [20]. Complex Gaussian mixture models (CGMMs), for example, have been proposed to jointly localize and separate sound sources by estimating the posterior distributions of the latent directions of arrival (DoAs) and time-frequency (TF) mask of sources [19], [21]. This method can complementarily estimate the DoAs and TF mask by iteratively and alternately updating these variables. In practice, however, this approach is sensitive to the initial values for the iteration, and the estimation often gets stuck at a bad local optimum, resulting in a performance limitation.

In this paper we present a hybrid robot audition framework, called MUSIC-CGMM, that complementarily combines the conventional cascading and unified frameworks (Fig. 1). We take the full advantage of the MUSIC-based localization, which works very well if its hyperparameters are appropriately selected. More specifically, we formulate a variant of the CGMM conditioned by the outputs of MUSIC. The hyperparameters of MUSIC are automatically selected based on the likelihood of the probabilistic model. The CGMM is, on the other hand, efficiently initialized and regularized with the localization results of MUSIC. The inference is formulated as a variational expectation-maximization (VEM) algorithm [22] and implemented in a mini-batch manner to work with moving sound sources in real time.

The main contribution of this study is to mildly integrate source localization and separation in a probabilistic manner. Existing methods integrate them in cascaded [6]–[8] or fully integrated manner [19]–[21], which results in a performance limitation. The proposed method integrates the MUSIC localization and CGMM separation to complementarily solve their problems while keeping their strengths. We demonstrate the effectiveness of the proposed method with experimental evaluations using both simulated data and data recorded by a mobile robot. We also show our method works in real time on a laptop computer having a graphic processing unit (GPU).

The rest of this paper is organized as follows. Sec. II introduces the related work on robot audition systems and also discusses recent neural methods. Sec. III describes the proposed method combining the MUSIC and CGMM. Sec. IV describes the real-time extension of the proposed method. Sec. V evaluates the proposed method in a variety of conditions with numerically simulated data. Sec. VI reports experiments with audio data recorded by an autonomous mobile robot. Sec. VII concludes this paper.

II. RELATED WORK

This section introduces the existing studies of sound source localization and separation methods and reviews the robot audition frameworks combining these methods.

A. SOUND SOURCE LOCALIZATION

A most fundamental approach to sound source localization is the steered response power (SRP) (or equivalently beamforming) method [23]–[25]. Especially, SRP-PHAT [24] has been widely used because of its robustness against reverberation. By using pre-measured steering vectors for potential source directions, SRP-PHAT estimates the source existence in the directions. Due to the high computational cost of searching all the potential directions, a low-cost searching algorithm called SVD-PHAT has been proposed with a singular value decomposition (SVD) [11]. It has been also reported that SVD-PHAT can improve the robustness against directional noise by using spatial covariance subtraction [26].

Since the SRP methods may be degraded when the input mixture includes multiple sources, sub-space methods have been widely studied to localize multiple sound sources from their mixture signal [9], [10], [27], [28]. MUSIC [9] is a popular sub-space method that utilizes the standard eigenvalue decomposition (SEVD) to distinguish speech and noise subspaces. A MUSIC method based on the generalized eigenvalue decomposition (GEVD) has also been investigated for improving the robustness against directional noise [29]. The GEVD-MUSIC has been utilized for localizing a victim's voice from a noisy quadrocopter [30], [31]. A variant of the GEVD-MUSIC with the generalized SVD (GSVD) has also been reported to reduce the computational cost [10].

B. SOUND SOURCE SEPARATION

BSS has been studied to separate a multichannel mixture signal into latent source signals with few prior information about sound sources or microphones [32]–[35]. Independent vector analysis (IVA) [33], for example, is a popular BSS method used to separate source signals based on their statistical independence. Assuming the source spectrograms of IVA to be low-rank, independent low-rank matrix analysis (ILRMA) [34] has been proposed to improve the separation performance. Another kind of BSS is a clustering-based approach that estimates a TF mask to separate source signals. A complex Gaussian mixture model (CGMM) [36] was reported to robustly estimate the TF mask for real noisy speech signals provided in

CHiME-3 and -4 challenges [37], [38]. This model is also called a complex angular central GMM (CACGMM) [39] and was widely used in CHiME-5 and -6 challenges [40], [41].

Source separation based on deep learning is also actively studied in efforts to achieve excellent performance. For example, deep clustering [42] and permutation invariant training (PIT) [43] are the most popular methods. A fully convolutional time-domain audio separation network (Conv-TasNet) [44] with PIT has yielded significantly better separation performance than that of the conventional TF domain neural methods. Despite their advances, the performance often deteriorates in unknown environments that are not included in the training data. To solve this problem, BSS has been utilized to generate pseudo supervised data for the target environment to train the network [45]–[47].

C. INTEGRATED SYSTEMS

Robot audition systems have been developed by combining the sound source localization and separation. ODAS [8], for example, is developed for low-cost embedded computers such as the Raspberry Pi computers. This system utilizes SRP-PHAT [25] for localization and separates sources by using a beamforming technique. HARK [5], [6] localizes sound sources based on MUSIC [10] and separate sources based on BSS constrained by the localization results [12]. HARK also performs automatic speech recognition to retrieve the contents of the separated signals. This system has been utilized for various robot systems including humanoid robots [16], autonomous mobile robots [48], and drones [17], [49]. These cascading systems are designed to work in real time by combining the individual modules of localization, separation, and recognition methods. Because each module is designed to perform a different task, the failure of a module is hard to be recovered by the subsequent modules.

To overcome the drawback of the cascading approach, several unified probabilistic methods have been developed [19], [21], [50]. A CGMM-like model inspired by latent Dirichlet allocation (LDA) [19], [51] was proposed to jointly estimate the TF mask and DoAs for source signals from a mixture recording. The estimation errors of the latent parameters can be recovered during the iterative inference because the model handles the dependency of the parameters. In addition, the number of sound sources is also automatically determined during the inference because this model is formulated in a Bayesian manner with a hierarchical Dirichlet process [52]. A CGMM with a complex inverse Wishart mixture model (CIWMM) [21] has also been proposed to jointly localize and separate sources with much less computational cost. These methods, however, have initialization sensitivity due to the dependencies between the TF mask and DoAs.

III. PROPOSED INTEGRATION OF MUSIC AND CGMM

To overcome the drawbacks of both the cascading and unified frameworks, our method combines MUSIC-based localization and CGMM-based separation in a probabilistic manner. We first describe the details of MUSIC and formulate a CGMM constrained by the results of MUSIC. We then explain the probabilistic inference to optimize the hyperparameters of MUSIC and to estimate the latent TF mask of the CGMM. Our framework enables MUSIC to select its hyperparameters from the model likelihood and the CGMM to stably separate sources with the prior information of MUSIC.

A. PROBLEM SPECIFICATION

The proposed MUSIC-CGMM estimates DoAs and TF mask of latent sound sources from a multichannel mixture recording as follows:

Input:
M-channel mixture signal $\mathbf{x}_{tf} \in \mathbb{C}^M$,
Output:
1. DoA $d_k \in \{1, \ldots, D\}$ of source $k \in \{1, \ldots, K_{\max}\}$,
2. Time-frequency mask $\hat{z}_{tfk} \in [0, 1]$,
Assumption:
1. The observation $\mathbf{X}$ includes $K_{\max}$ sources at most,
2. Pre-measured steering vectors $\mathbf{b}_{fd} \in \mathbb{C}^M$ are given,

where $t = 1, \ldots, T$ and $f = 1, \ldots F$ represent the indices for time frames and frequency bins, respectively. We utilize pre-measured steering vectors of potential DoAs $d = 1, \ldots, D$ to localize sound sources. These vectors are measured in an actual environment or calculated by using the plane-wave assumption from the array geometry. In this paper, we assume the potential DoAs d on a horizontal plane with an interval of 5° ($D = 72$). Note that because the steering vectors for an observation change from the pre-measured ones depending on the surrounding environment, we use the pre-measured vectors only as prior information.

B. MUSIC-BASED SOURCE LOCALIZATION

MUSIC-based localization utilizes the eigenvectors $\mathbf{e}_{fl} \in \mathbb{C}^M$ of the average spatial covariance matrix of an observation $\frac{1}{T}\sum_t \mathbf{x}_{tf}\mathbf{x}_{tf}^{\mathsf{H}}$. The eigenvectors can be split into two subsets respectively spanning a directional source space and a diffuse noise space by using their eigenvalues [9], [10]. Since the eigenvectors corresponding to the directional sources are orthogonal to those for the diffuse noise, a MUSIC spectrogram is calculated to indicate the source intensity at potential directions d as follows [10]:

$$y_{fd}^{(L)} = \frac{\mathbf{b}_{fd}^{\mathsf{H}}\mathbf{b}_{fd}}{\sum_{l=L+1}^{M} |\mathbf{b}_{fd}^{\mathsf{H}}\mathbf{e}_{fl}|}, \tag{1}$$

where $\mathbf{e}_{fl}$ is the eigenvector that has the l-th largest eigenvalue and L is a parameter indicating the number of sources in the observation.

The MUSIC spectrogram $y_{fd}^{(L)}$ is then merged into a MUSIC spectrum $y_d^{(L)}$ for all frequencies by taking the sum of $y_{fd}^{(L)}$ weighted by the eigenvalues $\lambda_{fl} \in \mathbb{R}_+$:

$$y_d^{(L)} = \sum_{f=1}^{F} \sqrt{\lambda_{f1}} \frac{\mathbf{b}_{fd}^{\mathsf{H}}\mathbf{b}_{fd}}{\sum_{l=L+1}^{M} |\mathbf{b}_{fd}^{\mathsf{H}}\mathbf{e}_{fl}|}. \tag{2}$$

The DoA candidates $d_k^{(L)}$ are finally obtained by taking the $K_{\max}$ peaks from this MUSIC spectrum $y_d^{(L)}$.

The MUSIC-based localization has generally been performed by thresholding the DoA candidates $d_k^{(L)}$ [10], [18], [31] to reject pseudo peaks. Since the peaks of the MUSIC spectrum includes pseudo sources caused by spatial aliasing and reflected sounds, the optimal thresholding parameter changes according to the array geometry and surrounding environments. In addition, because the optimum value of L changes for each frequency bin, this parameter depends on the characteristics of source signals. This dependency further makes it difficult to fix L in advance. Our MUSIC-CGMM robustly determines L and validates the candidates $d_k^{(L)}$ such that the log-marginal likelihood of the CGMM is maximized.

C. CGMM-BASED PROBABILISTIC MODEL

The CGMM [21], [36] represents the M-channel observed signal $\mathbf{x}_{tf} \in \mathbb{C}^M$ with $K_{\max}$ source signals $s_{tfk} \in \mathbb{C}$ and a noise signal $\mathbf{n}_{tf} \in \mathbb{C}^M$ as follows:

$$\mathbf{x}_{tf} = \sum_{k=1}^{K_{\max}} (z_{tfk}\mathbf{a}_{fk}) s_{tfk} + z_{tf0}\mathbf{n}_{tf}, \tag{3}$$

where $z_{tfk} \in \{0, 1\}$ is a TF mask introduced by assuming the source spectra to be sufficiently sparse and $\mathbf{a}_{fk} \in \mathbb{C}^M$ is the steering vector of source k. The source signals s_{tfk} and noise signal $\mathbf{n}_{tf}$ are assumed to follow the zero-mean complex Gaussian distributions:

$$s_{tfk} \sim \mathcal{N}_{\mathbb{C}}\left(0, \lambda_{tfk}\right) \qquad k = 1, \ldots, K_{\max}, \tag{4}$$

$$\mathbf{n}_{tf} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \lambda_{tf0}\mathbf{H}_{f0}\right), \tag{5}$$

where λ_{tfk} represents the power spectral density (PSD) of each signal and $\mathbf{H}_{f0}$ is the spatial covariance matrix (SCM) of the noise. From Eqs. (3)–(5), we obtain the likelihood function of $\mathbf{x}_{tf}$ as the following Gaussian mixture model:

$$\mathbf{x}_{tf} \sim \prod_{k=0}^{K_{\max}} \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \lambda_{tfk}\mathbf{H}_{fk}\right)^{z_{tfk}}, \tag{6}$$

where $\mathbf{H}_{fk} = \mathbb{E}[\mathbf{a}_{fk}\mathbf{a}_{fk}^{\mathsf{H}}]$ $(k = 1, \ldots, K_{\max})$ is the SCM of source k.

To constrain the directivity of each source, we assume the SCM $\mathbf{H}_{fk}$ to follow a complex inverse Wishart distribution (denoted as $\mathcal{IW}_{\mathbb{C}}$) [53]:

$$\mathbf{H}_{fk} \sim \mathcal{IW}_{\mathbb{C}}\left(\nu_k, \nu_k\mathbf{G}_{fk}^{(L)}\right), \tag{7}$$

where $\nu_k > M - 1$ is a hyperparameter that controls the strength of this prior distribution. To associate source k to the DoA candidate $d_k^{(L)}$ estimated by MUSIC, the SCM $\mathbf{G}_{fk}^{(L)}$ for source $k \in \{1, \ldots, K_{\max}\}$ is given as follows:

$$\mathbf{G}_{fk}^{(L)} = \mathbf{b}_{fd_k^{(L)}}\mathbf{b}_{fd_k^{(L)}}^{\mathsf{H}} + \epsilon I_M, \tag{8}$$

where $\epsilon \in \mathbb{R}_+$ is a hyperparameter representing the fluctuation of the source location. The SCM $\mathbf{G}_{f0}^{(L)}$ for noise, on the other hand, is given to represent the diffuse noise as follows:

$$\mathbf{G}_{f0}^{(L)} = I_M. \tag{9}$$

To encourage the shrinkage of sources corresponding to the incorrect DoA candidates, we put a Dirichlet-categorical prior [51] on z_{tfk}. We first put the categorical prior on the mask $\mathbf{z}_{tf} = [z_{tf1}, \ldots, z_{tfK}]^{\mathsf{T}}$ as follows:

$$\mathbf{z}_{tf} \sim \mathrm{Cat}\left(\pi_{t0}, \ldots, \pi_{tK}\right), \tag{10}$$

where $\pi_{tk} \in \mathbb{R}_+$ $(\sum_k \pi_{tk} = 1)$ is the prior probability that the source k is selected at time frame t. This parameter is further assumed to follow the Dirichlet distribution for encouraging the shrinkage:

$$\boldsymbol{\pi}_t \sim \mathrm{Dir}\left(\alpha_0, \ldots, \alpha_K\right), \tag{11}$$

where α_k is a hyperparameter whose smaller values encourage stronger shrinkage.

D. INTEGRATED INFERENCE

To perform source localization and separation complementarily, we estimate the posterior probability of the TF mask $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\lambda}, \mathbf{G}^{(L)})$ while selecting the MUSIC's hyperparameter L that maximizes the log-marginal likelihood $\log p(\mathbf{X}|\boldsymbol{\lambda}, \mathbf{G}^{(L)})$. The conventional thresholding of the DoA candidates $d_k^{(L)}$ of MUSIC is replaced by that of the estimated TF mask to consider the spatial model of the CGMM.

Since the log-marginal likelihood $\log p(\mathbf{X}|\boldsymbol{\lambda}, \mathbf{G}^{(L)})$ and the posterior $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\lambda}, \mathbf{G}^{(L)})$ is hard to analytically calculate, we estimate them by using variational Bayesian inference [22]. More specifically, we estimate them by introducing the following variational posterior q:

$$p(\mathbf{Z}, \boldsymbol{\pi}, \mathbf{H} \mid \mathbf{X}, \boldsymbol{\lambda}, \mathbf{G}^{(L)}) \approx q(\mathbf{Z})q(\boldsymbol{\pi})q(\mathbf{H}). \tag{12}$$

Instead of the log-marginal likelihood, we maximize its lower bound called the evidence lower-bound (ELBO):

$$\mathcal{L}^{(L)} = \mathbb{E}_q\left[\log p(\mathbf{X} \mid \mathbf{H}, \lambda_{tfk}, \mathbf{Z})\right] - \mathcal{D}_{\mathrm{KL}}[q(\mathbf{Z}, \boldsymbol{\pi}, \mathbf{H}) \mid p(\mathbf{Z}, \boldsymbol{\pi}, \mathbf{H} \mid \mathbf{G}_{fk}^{(L)})]. \tag{13}$$

The maximization of the ELBO corresponds to the minimization of the Kullback-Leibler divergence between the variational posterior q and the true posterior p.

The whole inference procedure is organized in three steps as illustrated in Fig. 1:

1) The MUSIC source localization is performed to obtain $d_k^{(L)}$ for each hyperparameter candidate $L \in \mathbb{L} = \{1, \ldots, M-2\}$.
2) For each candidate L, the posterior q and PSD $\boldsymbol{\lambda}$ are estimated to maximize the ELBO $\mathcal{L}^{(L)}$.
3) The results whose L maximizes $\mathcal{L}^{(L)}$ are selected as the output of MUSIC-CGMM.

The hyperparameter of MUSIC L is selected so that the log-marginal likelihood of the CGMM is maximized. This inference is called a type II maximum likelihood estimation. The posterior $q(\mathbf{Z}, \boldsymbol{\pi}, \mathbf{H})$ and PSD $\boldsymbol{\lambda}$, on the other hand, are

obtained by the variational expectation-maximization (VEM) algorithm. More specifically, this inference iteratively and alternately updates these variables as follows:

$$\log q(z_{tfk}=1) \;=\; \hat{z}_{tfk} \leftarrow \langle \log \pi_{tk} \rangle - \langle \log |\mathbf{H}_{fk}| \rangle - M \log \lambda_{tfk} - \mathrm{tr}\left(\langle \mathbf{H}_{fk}^{-1} \rangle \mathbf{x}_{tf} \mathbf{x}_{tf}^{\mathsf{H}} \right) + \text{const.}, \tag{14}$$

$$q(\boldsymbol{\pi}_t) \leftarrow \mathrm{Dir}\left(\boldsymbol{\alpha} + \sum_{f=1}^{F} \hat{\mathbf{z}}_{tf} \right), \tag{15}$$

$$q(\mathbf{H}_{fk}) \leftarrow \mathcal{IW}_{\mathbb{C}}\left(\nu_k + \sum_{t=0}^{T} \hat{z}_{tfk}, \mathbf{G}_{fk}^{(L)} + \sum_{t=0}^{T} \mathbf{x}_{tf} \mathbf{x}_{tf}^{\mathsf{H}} \right), \tag{16}$$

$$\lambda_{tfk} \leftarrow \frac{1}{M} \mathrm{tr}\left(\langle \mathbf{H}_{fk}^{-1} \rangle \mathbf{x}_{tf} \mathbf{x}_{tf}^{\mathsf{H}} \right), \tag{17}$$

where $\langle \cdot \rangle$ represents the expectation by the posterior q. As in the variational Bayesian GMM algorithms [22], we accelerate the inference by stopping the parameter updating of a redundant source k as $q(z_{tfk}=1) \leftarrow 0$ when the average value of the TF mask $\hat{z}_{tfk}$ is lower than a small value ψ^z.

E. INITIALIZATION OF CGMM

To update the variational posteriors $q(\mathbf{Z})$, $q(\boldsymbol{\pi})$, and $q(\mathbf{H})$ and the PSD $\boldsymbol{\lambda}$, they need initial values for the iteration. In this paper we initialize only $q(\mathbf{Z})$ and $\boldsymbol{\lambda}$ because $q(\boldsymbol{\pi})$ and $q(\mathbf{H})$ can be initialized by Eqs. (15) and (16), respectively. The $q(z_{tfk}=1)$ (equivalently $\hat{z}_{tfk}$) is initialized by using the localization results of MUSIC d_k as follows:

$$\hat{z}_{tfk} \propto \mathcal{N}_{\mathbb{C}}\left(\mathbf{x}_{tf} \,\middle|\, \mathbf{0}, \mathbf{G}_{fd_k}^{(L)} \right) \quad k = 1, \ldots, K_{\max}, \tag{18}$$

$$\hat{z}_{tf0} = \begin{cases} 1, & \text{if } |x_{tfm}|^2 < \tilde{x}, \\ 0, & \text{otherwise}, \end{cases} \tag{19}$$

where $\tilde{x}$ is a median value of the power spectrogram $|x_{tfm}|^2$ for initializing the mask of noise ($k=0$). The PSD is simply initialized by the average power spectrogram as follows:

$$\lambda_{tfk} = \frac{1}{M} \sum_{m=1}^{M} |x_{tfm}|^2. \tag{20}$$

F. POST-PROCESSING

After obtaining the DoA d_k and TF mask $\hat{z}_{tfk}$ by the VEM algorithm, we perform several post-processing techniques for further improving the separation quality. First, we remove sources having small powers, which can be considered as negligible sources. Such a source is detected if the following equation is satisfied:

$$10 \log_{10}\left(\frac{\sum_{t,f} \hat{z}_{tfk} \sum_m |x_{tfm}|^2}{\sum_{t,f,m} |x_{tfm}|^2} \right) < \psi^{\mathrm{dB}}. \tag{21}$$

This equation represents whether the power ratio of the masked spectrogram and the observed mixture spectrogram is less than ψ^{dB} dB. We finally obtain distortion-less source signals $\hat{s}_{tfk}$ by using a minimum variance distortionless response (MVDR) beamformer [54], [55] as follows:

$$\hat{s}_{tfk} = \frac{1}{\mathrm{tr}\left\{ \left(\mathbf{R}_f^{\neg k} \right)^{-1} \mathbf{R}_f^k \right\}} \mathbf{e}_1^{\mathsf{T}} \left(\mathbf{R}_f^{\neg k} \right)^{-1} \mathbf{R}_f^k \mathbf{x}_{tf}, \tag{22}$$

$$\mathbf{R}_f^k = \frac{1}{\sum_t \hat{z}_{tfk}} \sum_t \hat{z}_{tfk} \mathbf{x}_{tf} \mathbf{x}_{tf}^{\mathsf{H}}, \tag{23}$$

$$\mathbf{R}_f^{\neg k} = \frac{1}{\sum_t \left(1 - \hat{z}_{tfk}\right)} \sum_t \left(1 - \hat{z}_{tfk}\right) \mathbf{x}_{tf} \mathbf{x}_{tf}^{\mathsf{H}}, \tag{24}$$

where $\mathbf{e}_m$ is the m-th identity vector, and $\mathbf{e}_1$ is utilized to extract the source image at the 1st channel. The scale of the source image is corrected with the blind analytic normalization (BAN) postfilter as in [40].

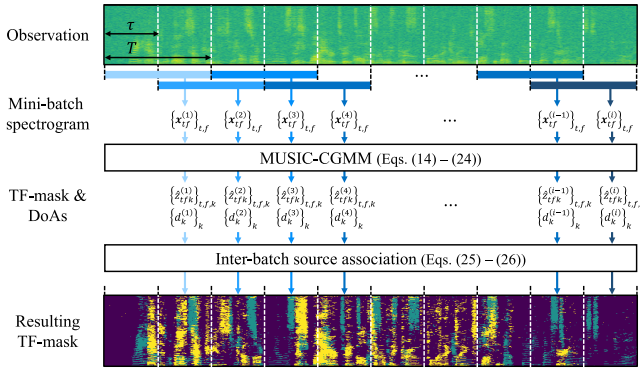

FIGURE 2. Overview of mini-batch inference.

IV. REAL-TIME INFERENCE

This section describes the mini-batch (streaming) inference of the MUSIC-CGMM called St-MUSIC-CGMM for a real-time robot audition system.

A. PROBLEM SPECIFICATION

The problem setting of the mini-batch inference is defined as follows:

Input:
M-channel mixture signal $\mathbf{x}_{tf}^{(i)} \in \mathbb{C}^M$ of mini-batch i,
Output:
1. DoA $d_k^{(i)} \in \{1, \ldots, D\}$ of source k,
2. Time-frequency mask $\hat{z}_{tfk}^{(i)} \in [0, 1]$,
Assumptions:
1. $\mathbf{X}^{(i)}$ includes $K_{\max}$ sources at most,
2. Pre-measured steering vectors $\mathbf{b}_{fd} \in \mathbb{C}^M$ are given,

where i indicates the index of a mini-batch having T frames with τ-frame shifting interval.

B. MINI-BATCH INFERENCE

The mini-batch inference incrementally outputs $L^{(i)}$, $d_k^{(i)}$, and $\hat{z}_{tfk}^{(i)}$ for each batch i by using the MUSIC-CGMM (Fig. 2). In this scenario there is ambiguity of source indices k between the adjacent batches. We solve this problem by associating sources between two batches with the following two criteria.

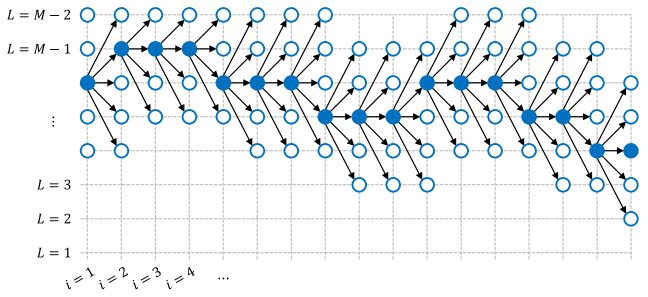

FIGURE 3. Pruning of hyperparameter candidates when both ΔL^- and ΔL^+ are set to 2. Blue-edged circles are evaluated values and blue-filled circles are optimum values at corresponding mini-batches.

One evaluates the similarity of the TF mask to maintain the consistency of source spectra. The other is the DoA difference to reject associations of sources far from each other. More specifically, we find a permutation $\mathcal{K}^{(i)}$ by minimizing the absolute differences of TF masks as follows:

$$\underset{\mathcal{K}^{(i)}}{\operatorname{argmin}} \sum_{t=1}^{T-\tau} \sum_{(k_1,k_2)\in\mathcal{K}^{(i)}} \left| \hat{z}^{(n-1)}_{(t+\tau)fk_1} - \hat{z}^{(i)}_{tfk_2} \right|, \tag{25}$$

where the permutation $\mathcal{K}^{(i)}$ satisfies that the DoA differences of sources are less than a threshold $\psi^d > 0$:

$$\mathcal{K}^{(i)} \subset \left\{ (k_1, k_2) \,\middle|\, \left| d^{(n-1)}_{k_1} - d^{(i)}_{k_2} \right| < \psi^d \right\}. \tag{26}$$

If the current batch i has a sound source that is not associated with the previous batch $i-1$, we consider that the source has newly appeared. Conversely, if a source in the previous batch $i-1$ is not associated with the current batch i, then the source is considered to have disappeared.

To localize and separate sound sources in real time, the proposed mini-batch inference is accelerated by a pruning technique as shown in Fig. 3. The original MUSIC-CGMM has to evaluate all the possible candidates of $L^{(i)} \in \mathbb{L} = \{1, \ldots, M-2\}$ per mini-batch i. This is problematic when the number of microphones is relatively large. To solve this problem, we introduce a pruning of the candidates. By assuming that the number of sources does not change drastically, the searching range of $L^{(i)}$ is limited to the values only around the last estimate $\hat{L}^{(i-1)}$:

$$\mathbb{L}^{(i)} = \{\hat{L}^{(n-1)} - \Delta L^-, \ldots, \hat{L}^{(n-1)} + \Delta L^+\} \wedge \mathbb{L}, \tag{27}$$

where ΔL^- and ΔL^+ are hyperparameters representing the lower and upper ranges for searching $L^{(i)}$, respectively.

V. EXPERIMENTS WITH SIMULATED DATA

To evaluate the proposed MUSIC-CGMM in various conditions, we first conducted experiments with mixture signals simulated numerically.

A. DATASET AND EVALUATION CRITERIA

The mixture signals were generated by convoluting room impulse responses (RIRs) to the dry speech signals from the WSJ0 English speech corpus [56]. The RIRs were numerically generated by using the image method [57]. As shown in Fig. 4, a microphone array was assumed in a rectangular room whose dimensions were 10 m $\times$10 m $\times$3 m . The reverberation time (RT_{60}) of the room was set to 0.2 s. The sources were placed at random positions such that the horizontal angle differences between the array and any pair of sources was at least 15°. The number of sources N in a mixture was changed randomly among $\{1, 2, 3\}$. The speech signals were mixed at random powers uniformly chosen between -2.5 dB and 2.5 dB. To simulate diffuse noise, we added Gaussian noise to the speech mixture with a signal-to-noise ratio (SNR) randomly chosen between 5 dB and 25 dB. We evaluated three circular microphone arrays 8 cm in diameter having $M \in \{4, 6, 8\}$ microphones. For each condition of M, we generated 1000 mixtures with a sampling rate of 16 kHz. In total, the evaluation dataset had 3000 mixture signals.

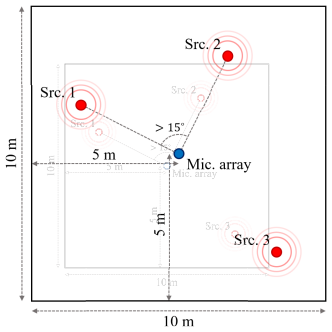

FIGURE 4. Configuration of microphone array and sources.

The performance of localization and separation was evaluated by the following criteria. The localization was evaluated by precision, recall, and their F-measure [49]. The precision $\mathcal{P}$ and recall $\mathcal{R}$ were calculated for each mixture as follows:

$$\mathcal{P} = \frac{N^{(correct)}}{N^{(estimated)}}, \qquad \mathcal{R} = \frac{N^{(correct)}}{N^{(true)}}, \tag{28}$$

where $N^{(true)}$, $N^{(estimated)}$, and $N^{(correct)}$ represent the number of true, estimated, and correctly-estimated sources, respectively. The $N^{(correct)}$ was calculated as the number of sources whose (average) estimated DoAs had errors less than 5° from those of the true sources. The F-measure $\mathcal{F}$ was calculated as the harmonic mean of the precision and recall. On the other hand, the source separation was evaluated by the scale-invariant source-to-distortion ratio (SI-SDR) [58] in dB and the perceptual evaluation of speech quality (PESQ) [59] ranging from -0.5 to 4.5, which are widely-used criteria for source separation. When the evaluated methods estimate a lower number of sources than the actual number, the SI-SDR and PESQ for the missing sources cannot be measured. In work reported in this paper, we regarded such SI-SDR and PESQ as $-\infty$ and evaluated the median values of SI-SDR and PESQ for all the mixtures in each condition.

B. EVALUATION FOR BATCH METHOD

This subsection reports the experimental results for the batch MUSIC-CGMM described in Sec. III.

TABLE 1. Localization performance for batch methods in average F-measure, precision, and recall.

Method	L	$M=4$			$M=6$			$M=8$		
		$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$
SRP-PHAT	–	0.72	0.84	0.67	0.80	0.87	0.76	0.83	0.87	0.81
SEVD-MUSIC	1	0.77	0.82	0.76	0.82	0.86	0.80	0.85	0.88	0.82
SEVD-MUSIC	2	0.79	0.81	0.79	0.83	0.86	0.82	0.84	0.87	0.83
SEVD-MUSIC	3	–	–	–	0.87	0.90	0.85	0.87	0.90	0.86
SEVD-MUSIC	4	–	–	–	0.84	0.85	0.84	0.85	0.87	0.85
SEVD-MUSIC	5	–	–	–	–	–	–	0.83	0.84	0.85
SEVD-MUSIC	6	–	–	–	–	–	–	0.81	0.82	0.83
PF-CGMM	–	0.62	0.58	0.72	0.67	0.62	0.76	0.67	0.61	0.77
MUSIC-CGMM	–	**0.84**	**0.85**	**0.83**	**0.88**	**0.90**	**0.88**	**0.90**	**0.91**	**0.90**

TABLE 2. Separation performance for batch methods in median SI-SDR.

Method	$M=4$			$M=6$			$M=8$		
	$N=1$	$N=2$	$N=3$	$N=1$	$N=2$	$N=3$	$N=1$	$N=2$	$N=3$
AuxIVA	9.11	3.57	0.13	8.65	3.48	0.37	8.40	3.41	0.20
ILRMA	14.01	2.48	−1.38	14.23	2.87	−0.69	14.22	3.17	−0.82
CACGMM	6.86	5.39	3.25	4.99	3.47	1.69	3.82	2.51	0.83
PF-CGMM	8.99	6.58	4.67	9.68	6.78	4.43	9.04	6.36	4.01
MUSIC-CGMM w/o MVDR	**14.81**	**10.14**	**6.37**	**16.19**	**10.52**	**6.99**	**16.54**	**10.57**	**7.19**
MUSIC-CGMM	12.63	7.67	3.20	13.70	8.47	4.50	14.04	8.93	5.18

1) EXPERIMENTAL CONDITION

The hyperparameters for MUSIC-CGMM were as follows. We obtained multichannel spectrograms by the short-time Fourier transform (STFT) with the frame length of 1024 samples and the shifting interval of 128 samples. To calculate the MUSIC spectrum, we cut off the frequency bins under 500 Hz to suppress false peaks. The maximum number of sources $K_{\max}$ was set to 8. The hyperparameters of the CGMM ν_0, ν_k ($k = 1, \ldots, K_{\max}$), ϵ, and α_k were set to 1.0, 5.0, 10^{-3}, and 5.0, respectively. The thresholding parameter ψ^z was set to 0.01. We iterated the proposed VEM update rules 20 times at each L. The parameter for the post-processing ψ^{dB} was set to -10 dB. These hyperparameters were determined empirically. Note that these parameters were not finely optimized for the evaluation data.

MUSIC-CGMM was compared with the following batch localization and separation methods. As baseline localization methods, SRP-PHAT [24] and SEVD-MUSIC [29] were evaluated. Although these methods were proposed more than 10 years ago, they are still actively utilized in recent studies [23]. Both the SRP-PHAT and SEVD-MUSIC have a thresholding parameter to reject pseudo peaks, which has to be selected appropriately. This parameter was optimized such that the average F-measure for all the mixtures was maximized for each condition of M. Since SEVD-MUSIC requires the number of sources L in advance, we evaluated all the candidates $\mathbb{L} = \{1, 2, \ldots, M-2\}$. As baseline separation methods, we evaluated three blind source separation methods: AuxIVA [33], ILRMA [34], and CACGMM [39]. AuxIVA and ILRMA assume a determined condition that the number of sources equals that of microphones. As in [60], we performed these methods with all the microphones and then selected sources that maximize the average SI-SDR for a mixture signal. The number of bases for ILRMA, which is a hyperparameter to control the low-rankness of source spectra, was set to 8. Similarly, because the CACGMM requires the number of sound sources K in advance, we performed the CACGMM with $K = 8$ to extract N sources that maximize the average SI-SDR for a mixture signal. As a method for joint localization and separation, we evaluated the permutation-free CGMM (PF-CGMM) that has a CIWMM prior on the SCMs [21]. For fair comparison, we put a categorical-Dirichlet prior on the TF mask (Eqs. (10)–(11)) to automatically estimate the number of sources and estimated the parameters by a VEM algorithm. We utilized the initialization method proposed in [19]. The hyperparameters were set to the same values as in MUSIC-CGMM.

2) EXPERIMENTAL RESULTS

Table 1 shows the localization performance in average F-measures, precisions, and recalls. First, SEVD-MUSIC outperformed SRP-PHAT by selecting the best value of L for each condition ($L = 2$ in $M = 4$, $L = 3$ in $M = 6$, and $L = 3$ in $M = 8$). The performance of SEVD-MUSIC, however, significantly deteriorated when the parameter was not appropriately selected, and thus the parameter tuning is essential for MUSIC. In contrast, the proposed MUSIC-CGMM outperformed both the SRP-PHAT and SEVD-MUSIC. This is because our method can adaptively select the best parameter for each observation. In addition, the localization performance of MUSIC-CGMM was significantly better than that of PF-CGMM. Since the localization and separation

TABLE 3. Separation performance for batch methods in median PESQ.

Method	$M=4$			$M=6$			$M=8$		
	$N=1$	$N=2$	$N=3$	$N=1$	$N=2$	$N=3$	$N=1$	$N=2$	$N=3$
AuxIVA	2.77	2.29	1.99	2.80	2.35	2.12	2.81	2.44	2.18
ILRMA	**3.20**	2.26	1.95	**3.33**	2.40	2.11	**3.42**	2.45	2.17
CACGMM	2.59	2.30	2.08	2.28	2.11	1.93	2.15	2.01	1.85
PF-CGMM	2.76	2.34	2.00	2.69	2.23	1.86	2.63	2.15	1.79
MUSIC-CGMM w/o MVDR	3.04	**2.54**	**2.09**	3.03	2.47	2.04	3.00	2.42	1.99
MUSIC-CGMM	3.02	2.45	2.08	3.16	**2.57**	**2.20**	3.20	**2.63**	**2.23**

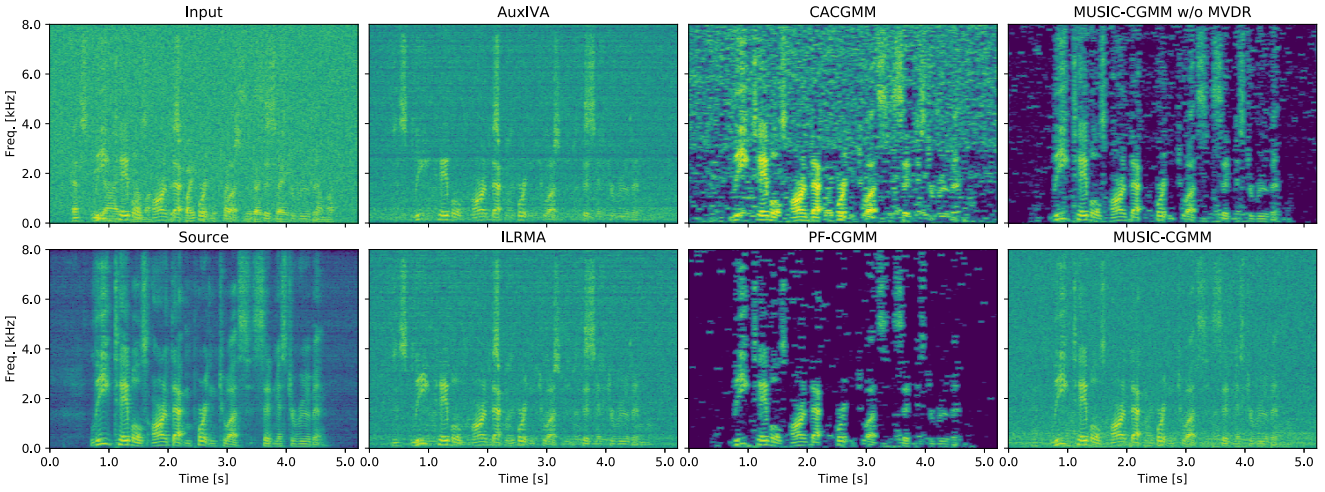

FIGURE 5. Excerpts of spectrograms separated by batch methods for an 8-channel mixture signal including two sources.

are mutually dependent, PF-CGMM easily gets stuck at a local optimum. Our MUSIC-CGMM successfully avoided this problem by utilizing the MUSIC localization.

Tables 2 and 3 summarize the separation performance in median SI-SDR and PESQ, respectively. Compared with the blind separation methods (AuxIVA, ILRMA, and CACGMM), PF-CGMM, which used the pre-measured steering vectors, achieved better performance. The performance was further improved by our MUSIC-CGMM as in the localization results. As shown in Fig. 5, AuxIVA and ILRMA have estimated different sources above and below about 3 kHz, which is called the frequency permutation ambiguity problem. CACGMM solved this problem with an external permutation solver [61], and PF-CGMM and MUSIC-CGMM resolved this problem by using the pre-measured steering vectors. The beamforming-based post-processing degraded the SI-SDRs from the version without it (MUSIC-CGMM w/o MVDR), while the PESQs were improved when the number of microphones M was 6 or more. We can see that the spectrograms of the mask-based methods (CACGMM, PF-CGMM, and MUSIC-CGMM w/o MVDR) have heavy salt-and-pepper-like noise on the separated spectrograms. Such artificial distortion often disturbs the automatic recognition of the separated signals. This distortion was recovered by the post-processing (MUSIC-CGMM), as can be seen from its spectrogram and the improvement in PESQ.

C. EVALUATION FOR STREAMING METHOD

Here we report experimental results for the mini-batch method (St-MUSIC-CGMM) proposed in Sec. IV.

1) EXPERIMENTAL CONDITION

The hyperparameters of the proposed St-MUSIC-CGMM were set as follows. The batch size T was set to 200 frames (1.6 s), and the shifting interval τ was set to 100 frames (0.8 s). The DoA difference tolerance ψ^d was set to 45°. The pruning parameters ΔL^- and ΔL^+ were set to 3 and 2, respectively. The other hyperparameters were set to the same values as for the batch MUSIC-CGMM in the previous section. As in the batch method, we experimentally determined these hyperparameters by hand.

The proposed method was compared with the real-time robot audition system called HARK. This software provides an all-in-one system integrating source localization, tracking, and separation (and also speech recognition) in a cascading manner. In this paper we report our evaluation of a standard HARK configuration that consists of MUSIC-based localization and geometric high-order decorrelation-based source separation (GHDSS) [12]. It also performs a post-processing based on spectral subtraction called histogram-based recursive level estimation (HRLE). The most sensitive hyperparameters of HARK were the number of sound sources L

TABLE 4. Localization performance for streaming methods in average F-measure, precision, and recall.

Method	L	$M=4$			$M=6$			$M=8$		
		$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$
HARK	1	**0.85**	**0.84**	**0.89**	0.87	0.85	0.91	0.87	0.86	0.90
HARK	2	0.74	0.69	0.84	0.82	0.80	0.87	0.83	0.80	0.89
HARK	3	–	–	–	0.74	0.70	0.82	0.77	0.74	0.86
HARK	4	–	–	–	0.56	0.49	0.73	0.69	0.63	0.83
HARK	5	–	–	–	–	–	–	0.63	0.56	0.83
HARK	6	–	–	–	–	–	–	0.50	0.43	0.73
St-MUSIC-CGMM w/o L-pruning	–	**0.85**	**0.84**	0.88	**0.89**	**0.87**	**0.93**	**0.91**	0.89	**0.93**
St-MUSIC-CGMM	–	**0.85**	**0.84**	0.88	**0.89**	**0.87**	0.92	**0.91**	**0.90**	**0.93**

TABLE 5. Separation performance for streaming methods in median SI-SDR.

Method	L	$M=4$			$M=6$			$M=8$		
		$N=1$	$N=2$	$N=3$	$N=1$	$N=2$	$N=3$	$N=1$	$N=2$	$N=3$
HARK	1	**15.02**	4.36	−0.10	15.29	5.06	0.24	15.24	5.21	0.37
HARK	2	13.02	7.24	2.34	14.40	7.92	3.39	14.43	8.22	4.04
HARK	3	–	–	–	14.07	7.88	3.72	14.10	8.09	4.31
HARK	4	–	–	–	12.27	6.76	2.13	13.56	7.95	4.14
HARK	5	–	–	–	–	–	–	13.24	7.77	4.08
HARK	6	–	–	–	–	–	–	12.23	6.95	2.41
St-MUSIC-CGMM w/o L-pruning	–	12.59	7.15	2.05	13.66	8.46	4.80	13.19	8.99	5.58
St-MUSIC-CGMM w/o MVDR	–	14.70	**7.39**	**2.61**	**16.38**	**9.36**	**5.97**	**16.48**	**10.11**	**6.75**
St-MUSIC-CGMM	–	12.59	7.15	2.05	13.65	8.46	4.71	13.28	9.00	5.56

TABLE 6. Separation performance for streaming methods in median PESQ.

Method	L	$M=4$			$M=6$			$M=8$		
		$N=1$	$N=2$	$N=3$	$N=1$	$N=2$	$N=3$	$N=1$	$N=2$	$N=3$
HARK	1	**3.34**	2.20	1.73	**3.53**	2.26	1.78	**3.59**	2.30	1.77
HARK	2	3.20	**2.47**	2.02	3.38	2.55	2.11	3.47	2.60	2.16
HARK	3	–	–	–	3.32	2.55	2.15	3.43	2.59	2.20
HARK	4	–	–	–	3.25	2.47	2.08	3.40	2.58	2.20
HARK	5	–	–	–	–	–	–	3.38	2.58	2.20
HARK	6	–	–	–	–	–	–	3.34	2.49	2.13
St-MUSIC-CGMM w/o L-pruning	–	3.04	2.40	**2.09**	3.21	**2.57**	**2.23**	3.21	**2.61**	**2.29**
St-MUSIC-CGMM w/o MVDR	–	2.79	2.06	1.61	2.85	2.18	1.78	2.92	2.28	1.85
St-MUSIC-CGMM	–	3.04	2.40	**2.09**	3.21	**2.57**	2.22	3.21	**2.62**	2.27

and the thresholding parameter for the (SEVD-)MUSIC. We evaluated HARK with all of $L \in \mathbb{L}$ and optimized the thresholding parameter for each L with a grid search such that the average F-measure was maximized.

2) EXPERIMENTAL RESULTS

Tables 4–6 show the localization and separation performance of our St-MUSIC-CGMM and HARK. As in the evaluation of the batch methods, the localization and separation performance by HARK (SEVD-MUSIC) significantly changed according to the number of sources L. In addition, the optimum thresholding parameter for HARK, as listed in Table 7, also changed depending on the conditions, which means that it takes a lot of time to tune the HARK system. In contrast, the proposed St-MUSIC-CGMM robustly localized and separated sources with the same hyperparameters over all the conditions. Compared with the version without the pruning (St-MUSIC-CGMM w/o L-pruning), we can see that both the localization and separation performance of St-MUSIC-CGMM was not significantly degraded in this evaluation.

TABLE 7. Optimum thresholding parameters for HARK.

L	$M=4$	$M=6$	$M=8$
1	29.5	27.5	28.5
2	41.0	38.5	36.0
3	–	44.5	42.5
4	–	50.5	45.0
5	–	–	47.0
6	–	–	53.0

VI. EXPERIMENTS USING RECORDED DATA

We report the experimental results using more realistic data with moving sources and real-world ambient noise.

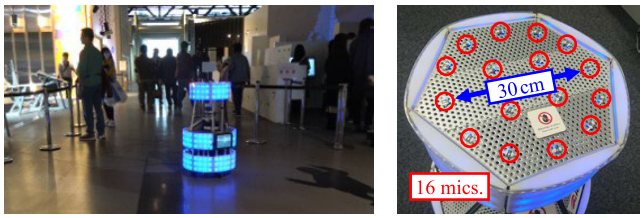

FIGURE 6. Our robot called Peacock demonstrating in Miraikan (left) and microphone array on its top (right).

A. EXPERIMENTAL CONDITION

The audio signals were collected by an autonomous mobile robot called Peacock [62], [63] (Fig. 6). The robot had a 16-channel microphone array on its top. As detailed in [64], we recorded environmental audio signals in the National Museum of Emerging Science and Innovation (Miraikan) with more than 1000 daily visitors. The recording was conducted by using a 24-bit A/D converter RASP-ZX (Systems In Frontier Inc.) with a sampling rate of 16 kHz. We used a 30-minute clip in the whole recording as ambient noise for this evaluation. As target sources, we collected 10 speech recordings in our experimental room, where the reverberation time (RT_{60}) was 0.82 s. As depicted in Fig. 7, we recorded three stationary sources and seven moving sources. These signals were recorded individually to evaluate the separation performance. We played back each speech signal randomly selected from the WSJ0 corpus by using a loudspeaker MS101-III (YAMAHA). The loudspeaker was lifted by hand about 1.1 m from the floor and moved around the robot to simulate moving sources. The source trajectories were captured by a light-detection-and-ranging (LiDAR) sensor VLP-32MR (Velodyne Lidar, Inc.) mounted on the robot. The mixture signals were generated by using these audio recordings as follows. For each signal, we mixed two target signals randomly selected from the 10 recordings such that the horizontal DoA difference of two sources was at least 15° in any time frames. The power of each speech was chosen uniformly between −2.5 dB and 2.5 dB. Finally, the random excerpt from the noise recording was added to the mixture with the SNR of 10 dB. We generated 100 mixture signals in this evaluation.

We evaluated our St-MUSIC-CGMM, which can handle source movements, and the online cascading system of HARK. The hyperparameters were set to the same values as in the previous evaluation. To reduce the computational time for the proposed method, MUSIC localization was performed for all the $M = 16$ channels and the CGMM separation was conducted with the half of them (i.e., $M = 8$). The number of sound sources L for MUSIC in the HARK was set to 2, and the thresholding parameter was set to 30. These parameters were experimentally determined by hand.

B. EXPERIMENTAL RESULTS AND DISCUSSIONS

As shown in Table 8, even when the sound sources were moving, the proposed St-MUSIC-CGMM outperformed the HARK system in both the median SI-SDR and PESQ.

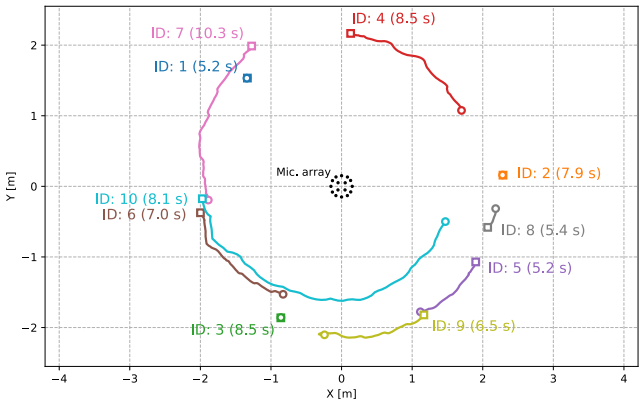

FIGURE 7. Trajectories of three stationary sources (ID: 1, 2, and 3) and seven moving sources. Squares and circles represent start and end locations, respectively.

TABLE 8. Separation performance for real-time methods in median SI-SDR and PESQ.

Method	SI-SDR	PESQ
HARK	2.16	2.07
St-MUSIC-CGMM w/o L-pruning	2.55	**2.14**
St-MUSIC-CGMM w/o MVDR	**4.78**	1.25
St-MUSIC-CGMM	2.57	**2.14**

The proposed streaming method was implemented by using Python and CuPy to utilize the general-purpose computing on GPUs (GPGPU). We measured the elapsed time to perform St-MUSIC-CGMM on a laptop computer ROG Zephyrus GX501GI (AsusTek Computer Inc.), which has GeForce GTX 1080 Max-Q (NVIDIA). To process 10.30 seconds of a mixture signal, the proposed method took 5.84 seconds, and the version without the pruning (St-MUSIC-CGMM w/o L-pruning) took 17.08 seconds. We can see that the computational time was reduced more than 60% with almost no loss in the separation performance.

Fig. 8 depicts the DoA trajectories estimated by St-MUSIC-CGMM and HARK. As shown in the top three rows, our St-MUSIC-CGMM successfully localized the moving sources when the movements were relatively small, although small ghost sources were occasionally localized as in the fourth result. Such a ghost source could be rejected by a post processing with recognition of the estimated source. However, as shown in the bottom row, when the source movement was too fast, our method failed to track the source and split one sound source into multiple sources. This is due to the limitation of the statistical approach that assumes the stationarity of the source location in a mini-batch. This problem could be resolved by the following two approaches. One is a multi-modal approach that uses visual information to reject or merge the incorrect sources from the appearances of sound sources [64]–[66]. The other solution is to use a hybrid approach combining the statistical and neural methods [46], [47], [67]. It has been reported that several limitation of statistical methods can be overcome by imitating

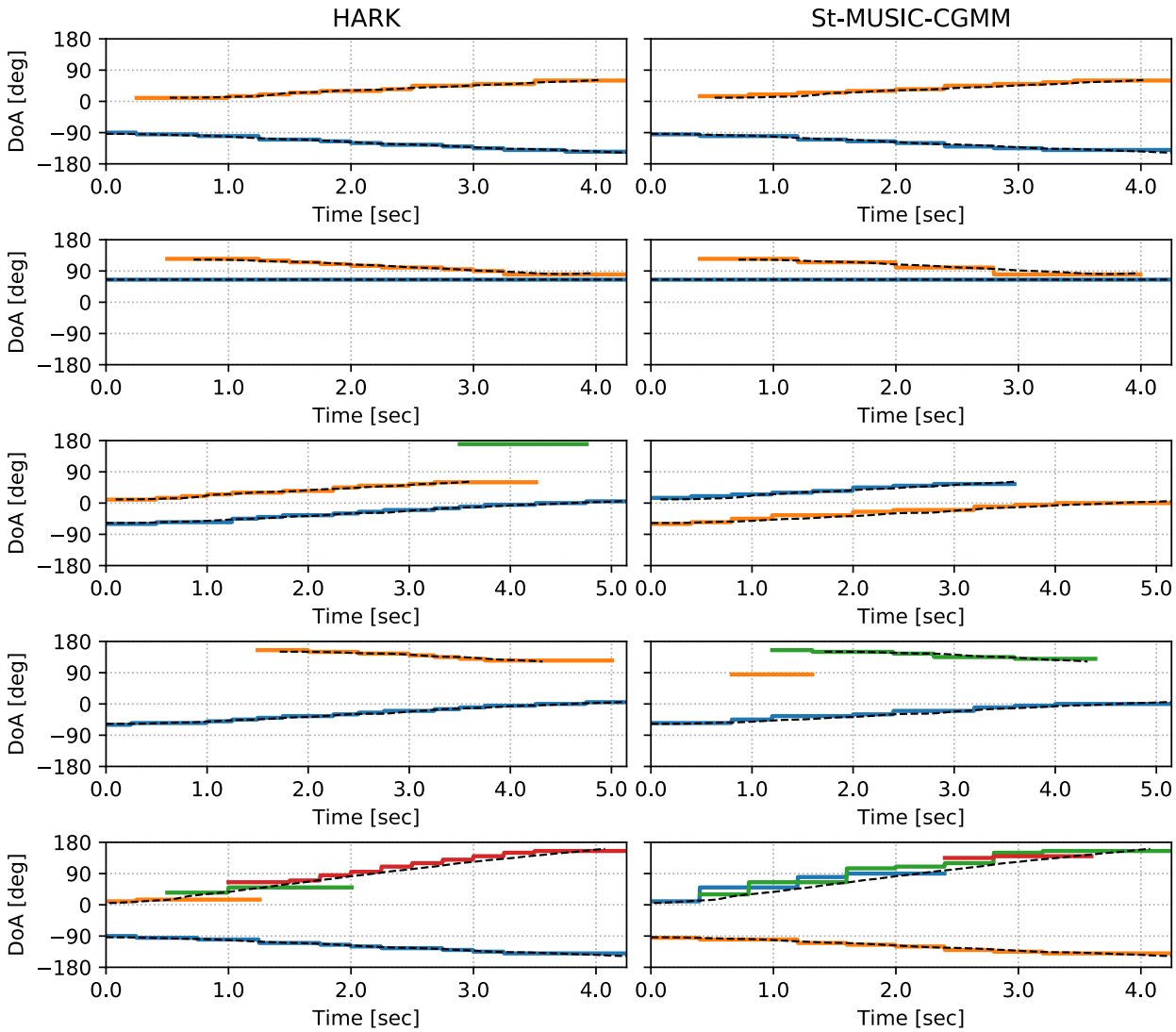

FIGURE 8. Excerpts of estimated DoAs of moving sources by HARK and proposed St-MUSIC-CGMM. Each color represents individual source. Dashed lines represent oracle trajectories of target sources.

the inference results of a statistical model with a neural network [45], [47]. While the statistical method usually estimates the latent variables optimized for only a single observation, the network can find the common characteristics of the statistical method over multiple observations. If the collected data includes a relatively small amount of fast-moving sources, such a hybrid approach will improve the separation performance.

VII. CONCLUSION

This paper presented a hybrid robot audition method, called MUSIC-CGMM, that complementarily combines the MUSIC-based localization and CGMM-based separation. While MUSIC requires hyperparameter tuning to achieve good performance, our method automatically optimizes the hyperparameter by evaluating the likelihood function of a CGMM. The CGMM-based separation, on the other hand, is efficiently initialized and regularized with the localization results of MUSIC to avoid bad local optima. The experimental results showed that the proposed method outperformed the conventional methods even when the number of sound sources was unknown. We also demonstrated that our method can localize and separate moving sources in a mini-batch manner. This mini-batch inference was implemented on a GPU-embedded laptop computer to work in real time.

Our future work includes integrating MUSIC-CGMM with sound event detection to recognize separated source signals. The proposed method is easily deployable robot audition that can localize and separate sound sources with the automatic tuning of the MUSIC's hyperparameters. The integration with source recognition will enable various robotic applications such as service robots communicating with humans in noisy crowded environments.

ACKNOWLEDGMENT

The authors would like to thank Dr. Yu Hoshina and Mr. Yusuke Date for their support in the experiment in Miraikan. They also thank Mr. Kazuki Kudo for valuable discussions.

REFERENCES

[1] H. G. Okuno and K. Nakadai, "Robot audition: Its rise and perspectives," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5610–5614.

[2] C. Evers, A. H. Moore, P. A. Naylor, J. Sheaffer, and B. Rafaely, "Bearing-only acoustic tracking of moving speakers for robot audition," in *Proc. IEEE Int. Conf. Digit. Signal Process. (DSP)*, Jul. 2015, pp. 1206–1210.

[3] C. Evers, Y. Dorfan, S. Gannot, and P. A. Naylor, "Source tracking using moving microphone arrays for robot audition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 6145–6149.

[4] K. Nakadai, G. Ince, K. Nakamura, and H. Nakajima, "Robot audition for dynamic environments," in *Proc. IEEE Int. Conf. Signal Process., Commun. Comput. (ICSPCC)*, Aug. 2012, pp. 125–130.

[5] K. Nakadai, H. G. Okuno, and T. Mizumoto, "Development, deployment and applications of robot audition open source software HARK," *J. Robot. Mechatron.*, vol. 29, no. 1, pp. 16–25, Feb. 2017.

[6] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "Design and implementation of robot audition system HARK-open source software for listening to three simultaneous speakers," *Adv. Robot.*, vol. 24, nos. 5–6, pp. 739–761, 2010.

[7] F. Grondin, D. Létourneau, F. Ferland, V. Rousseau, and F. Michaud, "The ManyEars open framework," *Auto. Robots*, vol. 34, no. 3, pp. 217–232, Apr. 2013.

[8] *ODAS*. Accessed: Jan. 15, 2021. [Online]. Available: https://github.com/introlab/odas

[9] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. AP-34, no. 3, pp. 276–280, Mar. 1986.

[10] K. Nakamura, K. Nakadai, and G. Ince, "Real-time super-resolution sound source localization for robots," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2012, pp. 694–699.

[11] F. Grondin and J. Glass, "SVD-PHAT: A fast sound source localization method," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 4140–4144.

[12] H. Nakajima, K. Nakadai, Y. Hasegawa, and H. Tsujino, "Blind source separation with parameter-free adaptive step-size method for robot audition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 6, pp. 1476–1485, Aug. 2010.

[13] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, vol. 1. Berlin, Germany: Springer, 2008.

[14] K. Inoue, P. Milhorat, D. Lala, T. Zhao, and T. Kawahara, "Talking with ERICA, an autonomous android," in *Proc. Annu. Meeting Special Interest Group Discourse Dialogue (SIGDIAL)*, 2016, pp. 212–215.

[15] A. Deleforge and R. Horaud, "The cocktail party robot: Sound source separation and localisation with an active binaural head," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interact. (HRI)*, Mar. 2012, pp. 431–438.

[16] I. Nishimuta, N. Hirayama, K. Yoshii, K. Itoyama, and H. G. Okuno, "A robot quizmaster that can localize, separate, and recognize simultaneous utterances for a fastest-voice-first quiz game," in *Proc. IEEE-RAS Int. Conf. Humanoid Robots*, Nov. 2014, pp. 967–972.

[17] M. Wakabayashi, H. G. Okuno, and M. Kumon, "Multiple sound source position estimation by drone audition based on data association between sound source localization and identification," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 782–789, Apr. 2020.

[18] T. Mizumoto, K. Nakadai, T. Yoshida, R. Takeda, T. Otsuka, T. Takahashi, and H. G. Okuno, "Design and implementation of selectable sound separation on the texai telepresence system using HARK," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2011, pp. 2130–2137.

[19] T. Otsuka, K. Ishiguro, H. Sawada, and H. G. Okuno, "Bayesian nonparametrics for microphone array processing," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 2, pp. 493–504, Feb. 2014.

[20] K. Itakura, Y. Bando, E. Nakamura, K. Itoyama, K. Yoshii, and T. Kawahara, "Bayesian multichannel audio source separation based on integrated source and spatial models," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 4, pp. 831–846, Apr. 2018.

[21] J. Azcarreta, N. Ito, S. Araki, and T. Nakatani, "Permutation-free cgmm: Complex Gaussian mixture model with inverse wishart mixture model based spatial prior for Permutation-free source separation and source counting," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 51–55.

[22] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Germany: Springer, 2006.

[23] C. Evers, H. W. Löllmann, H. Mellmann, A. Schmidt, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA challenge: Acoustic source localization and tracking," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1620–1643, 2020.

[24] C. Zhang, D. Florêncio, and Z. Zhang, "Why does PHAT work well in lownoise, reverberative environments?" in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2008, pp. 2565–2568.

[25] F. Grondin and F. Michaud, "Lightweight and optimized sound source localization and tracking methods for open and closed microphone array configurations," *Robot. Auto. Syst.*, vol. 113, pp. 63–80, 2019.

[26] F. Grondin and J. Glass, "Fast and robust 3-D sound source localization with DSVD-PHAT," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 5352–5357.

[27] H. Teutsch and W. Kellermann, "Detection and localization of multiple wideband acoustic sources based on wavefield decomposition using spherical apertures," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2008, pp. 5276–5279.

[28] H. Teutsch and W. Kellermann, "EB-ESPRIT: 2D localization of multiple wideband acoustic sources using eigen-beams," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 3, Mar. 2005, pp. 89–92.

[29] K. Nakamura, K. Nakadai, F. Asano, Y. Hasegawa, and H. Tsujino, "Intelligent sound source localization for dynamic environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2009, pp. 664–669.

[30] M. Strauss, P. Mordel, V. Miguet, and A. Deleforge, "DREGON: Dataset and methods for UAV-embedded sound source localization," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 1–8.

[31] K. Okutani, T. Yoshida, K. Nakamura, and K. Nakadai, "Outdoor auditory scene analysis using a moving microphone array embedded in a quadrocopter," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2012, pp. 3288–3293.

[32] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, "A review of blind source separation methods: Two converging routes to ILRMA originating from ICA and NMF," *APSIPA Trans. Signal Inf. Process.*, vol. 8, no. E12, pp. 1–14, 2019.

[33] 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. (WASPAA)*, Oct. 2011, pp. 189–192.

[34] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 9, pp. 1626–1641, Sep. 2016.

[35] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 3, pp. 550–563, Mar. 2010.

[36] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5210–5214.

[37] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. J. Fabian, M. Espi, T. Higuchi, S. Araki, and T. Nakatani, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand. (ASRU)*, Dec. 2015, pp. 436–443.

[38] J. Du, Y.-H. Tu, L. Sun, F. Ma, H.-K. Wang, J. Pan, C. Liu, J.-D. Chen, and C.-H. Lee, "The USTC-iFlytek system for CHiME-4 challenge," in *Proc. Int. Workshop Speech Process. Everyday Environ.*, 2016, pp. 36–38.

[39] N. Ito, S. Araki, and T. Nakatani, "Complex angular central Gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *Proc. 24th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2016, pp. 1153–1157.

[40] C. Boeddeker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann, and R. Haeb-Umbach, "Front-end processing for the CHiME-5 dinner party scenario," in *Proc. CHiME5 Workshop*, 2018, pp. 1–6.

[41] J. Deadman and J. Barker, "Simulating realistically-spatialised simultaneous speech using video-driven speaker detection and the CHiME-5 dataset," in *Proc. Interspeech*, Oct. 2020, pp. 349–353.

[42] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 31–35.
[43] M. Kolbaek, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.
[44] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal Time–Frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.
[45] L. Drude, D. Hasenklever, and R. Haeb-Umbach, "Unsupervised training of a deep clustering model for multichannel blind source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 695–699.
[46] L. Drude, J. Heymann, and R. Haeb-Umbach, "Unsupervised training of neural mask-based beamforming," in *Proc. Interspeech*, Sep. 2019, pp. 1253–1257.
[47] M. Togami, Y. Masuyama, T. Komatsu, and Y. Nakagome, "Unsupervised training for deep speech source separation with kullback-leibler divergence based probabilistic loss function," in *Proc. ICASSP IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 56–60.
[48] Y. Sasaki, R. Tanabe, and H. Takernura, "Online spatial sound perception using microphone array on mobile robot," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 2478–2484.
[49] K. Furukawa, K. Okutani, K. Nagira, T. Otsuka, K. Itoyama, K. Nakadai, and H. G. Okuno, "Noise correlation matrix estimation for improving sound source localization by multirotor UAV," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Nov. 2013, pp. 3943–3948.
[50] J. Nikunen and T. Virtanen, "Multichannel audio separation by direction of arrival based spatial covariance model and non-negative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 6677–6681.
[51] T. Otsuka, K. Ishiguro, H. Sawada, and H. G. Okuno, "Unified auditory functions based on Bayesian topic model," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2012, pp. 2370–2376.
[52] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *J. Amer. Stat. Assoc.*, vol. 101, no. 476, pp. 1566–1581, 2006.
[53] N. Q. K. Duong, E. Vincent, and R. Gribonval, "An acoustically-motivated spatial prior for under-determined reverberant source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 9–12.
[54] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 2, pp. 260–276, Feb. 2010.
[55] T. Higuchi, N. Ito, S. Araki, T. Yoshioka, M. Delcroix, and T. Nakatani, "Online MVDR beamformer based on complex Gaussian mixture model with spatial prior for noise robust ASR," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 4, pp. 780–793, Apr. 2017.
[56] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete," Linguistic data consortium, Philadelphia, PA, USA, 2007.
[57] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.
[58] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR—Half-baked or well done?" in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 626–630.
[59] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 2, May 2001, pp. 749–752.
[60] R. Scheibler and N. Ono, "Independent vector analysis with more microphones than sources," in *Proc. IEEE Workshop Appl. Signal Process. to Audio Acoust. (WASPAA)*, Oct. 2019, pp. 185–189.
[61] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2007, pp. 3247–3250.
[62] A. Kanezaki, J. Nitta, and Y. Sasaki, "GOSELO: Goal-directed obstacle and self-location map for robot navigation using reactive neural networks," *IEEE Robot. Autom. Lett.*, vol. 3, no. 2, pp. 696–703, Dec. 2017.
[63] Y. Sasaki and J. Nitta, "Long-term demonstration experiment of autonomous mobile robot in a science museum," in *Proc. IEEE Int. Symp. Robot. Intell. Sensors (IRIS)*, Oct. 2017, pp. 304–310.
[64] Y. Masuyama, Y. Bando, K. Yatabe, Y. Sasaki, M. Onishi, and Y. Oikawa, "Self-supervised neural audio-visual sound source localization via probabilistic spatial modeling," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 4848–4854.
[65] S. M. Naqvi, M. Yu, and J. A. Chambers, "A multimodal approach to blind source separation of moving sources," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 5, pp. 895–910, Oct. 2010.
[66] I. Hara, F. Asano, H. Asoh, J. Ogata, N. Ichimura, Y. Kawai, F. Kanehiro, H. Hirukawa, and K. Yamamoto, "Robust speech interface based on audio and video information fusion for humanoid HRP-2," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, vol. 3, Oct. 2004, pp. 2404–2410.
[67] Y. Bando, Y. Sasaki, and K. Yoshii, "Deep Bayesian unsupervised source separation based on a complex Gaussian mixture model," in *Proc. IEEE 29th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Oct. 2019, pp. 1–6.

YOSHIAKI BANDO (Member, IEEE) received the M.S. and Ph.D. degrees in informatics from Kyoto University, Kyoto, Japan, in 2015 and 2018, respectively. He is currently a Researcher with the Artificial Intelligence Research Center (AIRC), National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan. He is also a Visiting Researcher with the RIKEN Center for Advanced Intelligence Project (AIP), Tokyo. His research interests include microphone array signal processing, deep Bayesian learning, robot audition, and field robotics. He is a member of RSJ and IPSJ.

YOSHIKI MASUYAMA (Graduate Student Member, IEEE) received the B.E. degree from Waseda University, in 2019. He is currently pursuing the M.E. degree with the Department of Intermedia Art and Science, Waseda University. He is a Student Member of ASJ and IPSJ.

YOKO SASAKI (Member, IEEE) received the Ph.D. degree in mechanical engineering from the Tokyo University of Science, in 2009, for her research on auditory systems for mobile robots. She joined the Digital Human Research Center, National Institute of Advanced Industrial Science and Technology (AIST), where she is currently a Senior Researcher with the Artificial Intelligent Research Center. Her research interests include developing autonomous mobile robots and sensing and navigation technology, especially for home environments.

MASAKI ONISHI received the M.Eng. and Dr.Eng. degrees from Osaka Prefecture University, in 1999 and 2002, respectively. From 2002 to 2006, he was a Research Scientist with the Bio-Mimetic Control Research Center, RIKEN. Since 2006, he has been a Research Scientist with the National Institute of Advanced Industrial Science and Technology (AIST), where he is currently the Team Leader of the Social Intelligence Research Team, Artificial Intelligence Research Center. His research interests include video surveillance, crowd simulation, and automated machine learning.