Multiple Acoustic Source Localization in Microphone Array Networks

Jielong Yang, Xionghu Zhong, Weiguang Chen, and Wenwu Wang

***Abstract*—The problem of multiple acoustic source localization using observations from a microphone array network is investigated in this article. Multiple source signals are assumed to be window-disjoint-orthogonal (WDO) on the time-frequency (TF) domain and time delay of arrival (TDOA) measurements are extracted at each TF bin. A Bayesian network model is then proposed to jointly assign the measurements to different sources and estimate the acoustic source locations. Considering that the WDO assumption is usually violated under reverberant and noisy environments, we construct a relational network by coding the distance information between the distributed microphone arrays such that adjacent arrays have higher probabilities of observing the same acoustic source, which is able to mitigate the miss detection issues in adverse environments. A Laplace approximate variational inference method is introduced to estimate the hidden variables in the proposed Bayesian network model. Both simulations and real data experiments are performed. The results show that our proposed method is able to achieve better source localization accuracy than existing methods.**

***Index Terms*—Acoustic Source Localization, Bayesian Network, Laplace Approximate Variational Inference, Time-Frequency Masking, Time-Delay of Arrival.**

I. Introduction

Acoustic source localization (ASL) in a room environment plays an important role in many speech and audio applications such as multimedia, hearing aids, hands-free speech communication, and teleconferencing systems as the location information can be fed into a higher processing stage for high-quality speech acquisition, enhancement of a specific speech signal in the presence of other competing talkers, or directing a camera towards the acoustic source [1]–[6]. However, it is a difficult task to provide an accurate position estimate since the received audio signal can be significantly distorted and its statistical properties can be changed drastically due to room reverberation and noise. The difficulty is further increased when multiple sources are simultaneously active in the localization scene. Distributed acoustic sensor networks composed of a number of randomly deployed microphones or microphone arrays have been increasingly attractive for ASL due to their higher flexibility and scalability, and better spatial coverage compared to a single microphone or microphone array.

In the past, methods based on time-delay of arrival (TDOA) measurements are extensively employed and studied for ASL [7]–[16] due to their simplicity and ease of access in many applications. TDOA measurements can be extracted, for example, by employing the generalized cross-correlation (GCC) function [17] or adaptive eigenvalue decomposition (AED) algorithm [18]. Since each TDOA yields half a hyperholoid of two sheets which, in the far field, can be approximated by an angular segment, multiple TDOA measurements from distributed microphone arrays are usually employed to triangulate a target position [19], [20]. Such a triangulation can be approximated by either using a linear intersection (LI) algorithm [21] or an extended Kalman filter (EKF) [13], [16]. In [22], the authors consider the 2D source localization problem using TDOA at a minimal element monitoring arrays in both Cartesian and polar coordinate systems. However, in the presence of noise and room reverberation, ghost peaks may present in the GCC function and spurious TDOA measurements may be collected and the subsequent triangulation methods can be seriously degraded. In [23], a TDOA denoising method is proposed such that better localization performance can be achieved by using TDOA measurements. In [24], a TDOA outlier removal method is proposed to enhance the localization accuracy. In [25], the authors show the sufficient and necessary conditions of the uniqueness of localizing a single acoustic source and propose a geometric formulation to estimate the sound source using observations from arbitrarily shaped microphone arrays. However, in [23], [24] and [25], only one source is active at each time instance.

In a real conversation, multiple talkers can also be simultaneously active and, under such a scenario, the received signal is a mixture of different speech sources. This significantly increases the complexity of the ASL problem since: *i*) TDOAs for multiple sources are no longer easily available; and *ii*) given the TDOA measurements for multiple sources, the measurement-to-source assignment is unknown.

Many methods are proposed to obtain TDOAs for multiple sources. Knowing that traditional GCC methods may not yield sharp peaks for TDOAs of multiple sources, a degenerate unmixing estimation technique (DUET) [26], [27] is introduced to

Manuscript received March 3, 2020; revised August 27, 2020; accepted November 4, 2020. Date of publication November 30, 2020; date of current version December 18, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61971186. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dilek Hakkani-Tur. *(Jielong Yang and Weiguang Chen contributed equally to this work.) (Corresponding author: Xionghu Zhong.)*

Jielong Yang is with the School of Artificial Intelligence, Jilin University, Changchun 410000, China (e-mail: jyang022@e.ntu.edu.sg).

Xionghu Zhong and Weiguang Chen are with the College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China (e-mail: xzhong@hnu.edu.cn; cwg@hnu.edu.cn).

Wenwu Wang is with the Department of Electrical and Electronic Engineering, University of Surrey, GU2 7XH Guildford, U.K. (e-mail: w.wang@surrey.ac.uk).

Digital Object Identifier 10.1109/TASLP.2020.3039390

extract the measurement set for multiple sources. In DUET, the source signals are assumed to be window-disjoint-orthogonal (WDO) in the time-frequency (TF) domain. Hence, the TF spectrogram of sources can be considered as separated and the phase difference of the arrived signal due to each source can be extracted. Mandel *et al.* [28] also built a probabilistic models for phase difference and attenuation ratio information and used an expectation-maximization (EM) algorithm to find the TDOAs of multiple sources. However, the EM algorithm needs a burn-in period to converge to the final estimates. Other multi-source TDOA estimation methods based on signal separation for localization problem can also be found in [29]–[35].

Various source localization and data association methods are also studied in the literature. In [36], [37], grid-based methods are proposed to localize multiple sound sources using the DOA estimates of each microphone array. In [38], a versatile blind signal processing framework is proposed, which provides a unified treatment of both blind signal separation problem and multichannel blind deconvolution problem. Blind signal separation is used in [39] to extract DOA measurements and an intersection point selection scheme is introduced to locate multiple sources. However, the method does not consider miss detection issue. In [40], a method called Acoustic Simultaneous Localization and Mapping (aSLAM) is proposed to simultaneously map the 3D positions of multiple sound sources and to passively localize a moving observer. In this method, the observer's spatio-temporal diversity is used to probabilistically triangulate the source positions. In [41], a multi-view soundfield imaging method is proposed, which generalize the previous method from a single array to multiple arrays. In [42], the authors develop a two stage method and the method uses DOA estimates of microphone arrays to first estimate association features that describe how the frequency components of the captured signals are distributed to the sources, then both DOA estimates and association features are used to localize sound sources. In [43], the authors regard the data association problem as a measurement set partition task, which is further transformed into a generalized multidimensional assignment problem. The methods in [42], [43] deal with both the source localization and the missed detection problem. However, all these methods in the literature do not consider the location relationships between microphone arrays.

In the presence of noise and reverberation, and in particular, when the source is located at far-field, signal to noise ratio (SNR) and signal to reverberation ratio (SRR) can be low and the spectrogram is usually smeared and blurred. The WDO assumption is thus violated. However, it is observed from Fig. 1 that adjacent microphones are highly likely to be able to detect the same source, and vice versa. Hence, the information of distance between each pair of sources is essentially important to measurement-to-source association. In this work, such information is coded and exploited such that adjacent arrays have higher detection probabilities to the same acoustic source, which is able to mitigate the miss detection issue in the presence of reverberation. A relational network-based Bayesian network model is constructed and a Laplace approximate variational inference method is then introduced to estimate the hidden variables indicating the measurement-to-source associations and the corresponding source positions.

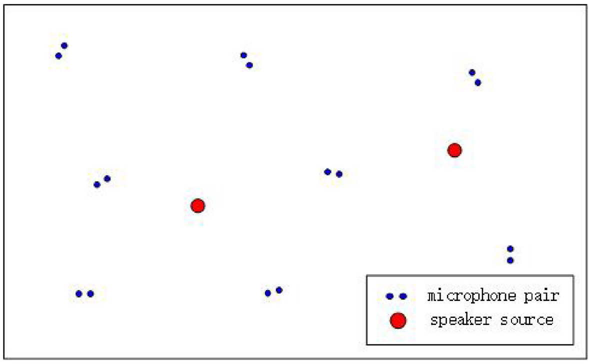

Fig. 1. Illustration of the localization scene using a microphone array network.

It is worth mentioning that several source localization methods focusing on the DOA estimation rather than estimation of the exact Cartesian (x, y) positions of the sources have been developed. In [44]–[47], dynamic sources are considered and a random finite set based Bayesian filtering approach was presented to track the sources. In [48], binaural cues, interal time difference and intensity difference were extracted from a microphone pair, and these observations are compared with predicted reference values obtained from simulations using prior knowledge of a catalogue head-related transfer functions (HRTFs). These reference values are obtained based on the binaural response of a KEMAR dummy head. The target space is modeled as a set of subspaces and switches among them with predefined jump probabilities. In [49], a distributed algorithm is proposed to estimate DOAs of multiple speech sources. In [50], an independent component analysis based approach was introduced to demix the speech mixtures from multiple sources and a probability hypothesis density filter was employed to track the DOAs of the sources.

The main contributions of this paper are: *i)* We develop a Bayesian-network-based learning method to jointly associate the TDOAs from each spectrogram bin to different sources and estimate the source locations; *ii)* our method considers both the missed detection problem and the data association problem; and *iii)* our method incorporates the distance information among arrays to improve the localization performance. The rest of this paper is organized as follows: in Section II, the DUET-based TDOA measurement extraction is introduced and ASL framework is formulated; in Section III, the inference algorithm is presented; the performance of the proposed approach is studied in Section IV. Finally, conclusions are drawn and directions for future work are discussed in Section V. A list of notations is summarized in below to illustrate the meaning of variables and symbols in the measurement extraction and localization algorithms.

Notations: We use boldfaced characters to represent vectors and matrices. Suppose that $\mathbf{A}$ is a matrix, then $\mathbf{A}(m, \cdot)$, $\mathbf{A}(\cdot, m)$, and $\mathbf{A}(m, n)$ denote its m-th row, m-th column, and (m, n)-th element, respectively. The vector $(x_1, \ldots, x_N)$ is abbreviated as $(x_i)_{i=1}^N$, or $(x_i)_i$ if i is running over the vector index. We use $\mathrm{Cat}(p_1, \ldots, p_K)$, $\mathrm{Dir}(\frac{\alpha}{K_s}, \ldots, \frac{\alpha}{K_s})$, $\mathrm{Unif}(a, b)$, $\mathrm{Unif}(1, \ldots, R)$, $\mathcal{B}e(g_0,\ h_0)$ and $\mathcal{N}\left(\mathbf{M},\ \mathbf{V}\right)$ to represent the categorical distribution with category probabilities $p_1, \ldots, p_K$, the Dirichlet distribution with concentration parameters $\frac{\alpha}{K_s}, \ldots, \frac{\alpha}{K_s}$, the uniform distribution over the interval (a, b), the uniform distribution over the discrete set $\{1, \ldots, R\}$, the beta distribution with shape parameters (g_0, h_0), and the normal distribution with mean $\mathbf{M}$ and covariance $\mathbf{V}$, respectively. We use $\Gamma(\cdot)$ and $\Psi(\cdot)$ to denote the gamma function and digamma function, respectively. The notation $\sim$ means equality in distribution. The notation $p(y \mid x)$ denotes a conditional probability density function of a random variable y conditioned on x. $\mathbb{E}$ is the expectation operator and $\mathbb{E}_q$ is expectation with respect to the probability distribution q. We use $I(a, b)$ and $I(a > b)$ to denote the indicator function. $I(a, b) = 1$ if $a = b$ and 0 otherwise. $I(a > b) = 1$ if $a > b$ and 0 otherwise. The notation $\|\cdot\|$ denotes the l_2 norm, and $\mathrm{ones}(1, K)$ is a $1 \times K$ vector with all entries equal to one.

II. Problem Formulation and Model

In this section, the problem of multiple acoustic source localization is formulated. TDOA measurements at each TF bin across different microphone arrays are estimated and a Bayesian network is then developed to jointly assign the measurements to the corresponding sources and estimate the position of each source.

A. Measurement Extraction Over Distributed Arrays

Assume that N microphone arrays are deployed to receive the speech signals emitted by K speakers at a discrete time step t. Let ω be a TF bin index, and $S_k = (S_{k,\omega})_\omega$ denotes the short time Fourier transform (STFT) of the k-th source signal. Ignoring the effect of noise and reverberation, the signal model in the TF domain for the i-th microphone of the n-th array is

$$Z_{n,i}(\omega) = \sum_{k=1}^{K} a_{n,i}(k) e^{-j\omega\tau_{n,i}(k)} S_{k,\omega}, \tag{1}$$

where $a_{n,i}(k) = \frac{1}{4\pi r_{n,i}(k)}$ represents the attenuation with $r_{n,i}(k)$ denoting the corresponding distance from source k to the i-th microphone of the n-th array, and $\tau_{n,i}(k)$ represents the time-delay of the k-th source signal at the i-th microphone of n-th microphone pair. According to the WDO assumption [26], the TF bins are disjoint. Hence, each TF bin carries either information regarding one of the sources, or simply noise.

Here we consider the case where each array has two microphones. The solution for arrays with more than two microphones can be extended in a straightforward manner. The ratio of the TF bins across a microphone pair is given by

$$R_n(\omega) = \frac{Z_{n,1}(\omega)}{Z_{n,2}(\omega)} = a_n(\omega) e^{-j\omega y_n(\omega)}, \tag{2}$$

where $a_n(\omega)$ and $y_n(\omega)$ are the gain-ratio (GR) and time-delay of arrival (TDOA) estimates for TF bin ω respectively. Suppose that the k-th source is active on ω (the contribution of other sources on this TF bin is thus nil), the GR and TDOA are given respectively as

$$\begin{aligned} a_n(\omega) &= |R_n(\omega)| = \frac{a_{n,1}(k)}{a_{n,2}(k)} \triangleq a_n(k), \\ y_n(\omega) &= \frac{\angle R_n(\omega)}{-\omega} = \tau_{n,1}(k) - \tau_{n,2}(k) \triangleq \tau_n(k), \end{aligned} \tag{3}$$

with $|\cdot|$ and $\angle\cdot$ denoting the amplitude and the phase of the estimates respectively, and $a_n(k)$ and $\tau_n(k)$ are the GR and TDOA information of the k-th source, respectively. Note that the TF bin index ω can be omitted in (3) as the GRs and TDOAs are determined by the geometry of the source and the microphone arrays, and thus the same across different TF bins associated to a source.

Based on the GR and TDOA parameters, a histogram of all TF bins can be generated and the TF bins for each source can thus be clustered and separated in the TF domain. TDOAs for multiple sources can hence be associated and the position of each source can be triangulated accordingly. Assume that at the n-th, for $n = 1, \ldots, N$ microphone array, a set of TDOAs $\mathbf{y}_n(\omega) = \{y_n(1), \ldots, y_n(\Omega)\}$ is obtained by using DUET. Such a TDOA set contains the source generated TDOAs as well as false TDOAs when reverberation and noise are considered.

Let $\mathbf{l}_k$ denote the location of the k-th source. For the measurement generated by the k-th source, its relationship to the location of the source is given by

$$y_n(\omega^k) = \frac{\|\mathbf{l}_k - \boldsymbol{p}_{n,1}\| - \|\mathbf{l}_k - \boldsymbol{p}_{n,2}\|}{c}. \tag{4}$$

where $\boldsymbol{p}_{n,i}$, $i \in \{1, 2\}$ is the position of the i-th microphone of the n-th array and ω^k represents that TF bin ω is associated to source k. Equation (4) shows that the source positions can be estimated by using correctly assigned measurements. However, in the presence of noise and reverberation, the spectrogram is smeared and blurred. The WDO assumption is violated and such a clustering-based method is no longer valid, i.e., it is very difficult to associate the TDOAs due to the same source and consequently, the location estimates can be significantly deviated from the ground truth.

B. Bayesian Network Model for Measurement-to-Source Association

Consider N microphone arrays monitoring K sound sources. The ground truth Cartesian coordinates of sources are $\mathbf{l} = \{\mathbf{l}_k\}_{k=1}^{K}$, which is the random variable we want to estimate. Let $\mathbf{l}' = \{\mathbf{l}'_n\}_{n=1}^{N}$ be the locations of arrays. The locations of the arrays and the number of sources are known. We generate F frequency bins at each time-frame and use observations of T time frames and thus in total $\mathbf{\Omega} \triangleq T \times F$ time frequency bins are considered.

Let $\mathbf{y} = \{y_n(\omega)\}_{n,\omega}$ be the collection of the measurements where $y_n(\omega)$ is the observation corresponding to the ω-th time frequency bin of the received signal of array n, and $s_n(\omega)$ be the index of the source signal that ω-th time frequency bin of array n mainly comes from. Then we have the following observation model

$$p(y_n(\omega) \mid s_n(\omega) = k, \mathbf{l}_k) = \mathcal{N}\Big(f(\mathbf{l}_k, \mathbf{l}'_n),\ \sigma^2\Big), \tag{5}$$

where σ is the known observation variance and $f(\mathbf{l}_k, \mathbf{l}'_n)$ is a known nonlinear function and depends on the structure of the arrays and the observation $y_n(\omega)$. Two examples of $f(\mathbf{l}_k, \mathbf{l}'_n)$ are as follows:

1) If $y_n(\omega)$ is the DOA of the ω-th TF bin of the signal received by array n, then $f(\mathbf{l}_k, \mathbf{l}'_n) \triangleq \arctan(\frac{l_k(2) - l'_n(2)}{l_k(1) - l'_n(1)})$.
2) If the number of microphones in each array is 2 and $y_n(\omega)$ is the TDOA of the ω-th TF bin of array n. Then $f(\mathbf{l}_k, \mathbf{l}'_n) \triangleq \frac{\|\mathbf{l}_k - \mathbf{l}'_n\|}{c}$, where c is the sound speed.

The signal received by array n may come from multiple sound sources. Let $\boldsymbol{\pi}_n = \{\pi_{n,k}\}_k$, where $\pi_{n,k}$ is the probability of $s_n(\omega) = k$, namely

$$s_n(\omega) \sim \mathrm{Cat}\left(\boldsymbol{\pi}_n\right). \tag{6}$$

In (6), $s_n(\omega)$ is a cluster index and in our method. We cluster the observation $y_n(\omega)$ by estimating $s_n(\omega)$. In our model, we not only cluster time-frequency bins of the same array, but also cluster time-frequency bins across different arrays. To achieve this, we use the same hierarchical model for any $n \in \{1, 2, \ldots, N\}$, given as

$$\boldsymbol{\pi}_n \sim \mathrm{Dir}\left(\widetilde{\boldsymbol{\pi}} + \alpha\right), \tag{7}$$

where α is a known hyper parameter and

$$\widetilde{\boldsymbol{\pi}} \triangleq \{\widetilde{\pi}_k\}_{k=1}^K \sim \mathrm{LogNormal}\left(\mathbf{M},\ \mathbf{V}\right) \tag{8}$$

with $\mathbf{M}$ and $\mathbf{V}$ being known hyper parameters. In (8), the log normal distribution ensures that all the elements of $\widetilde{\boldsymbol{\pi}}$ are positive. In (7), we use the Dirichlet distribution since the support of a Dirichlet distribution can be regarded as the probabilities of categorical events. Besides, the Dirichlet distribution is the conjugate prior distribution of the categorical distribution in (7), which helps to compute the posterior distribution in Bayesian inference.

Apart from observations $\mathbf{y}$, we also use the geometry relationships among arrays as arrays have higher probability to observe the same source when they are closely located. Let $z_{n\to m}$ be the cluster (i.e., source) index array n belongs to under the influence of array m. Let $\mathbf{D}(n, m)$ be the distance between array n and array m. We assume $z_{n\to m} \mid \boldsymbol{\pi}_n \sim \mathrm{Cat}(\boldsymbol{\pi}_n)$, $z_{m\to n} \mid \boldsymbol{\pi}_m \sim \mathrm{Cat}(\boldsymbol{\pi}_m)$, and $\beta_k \sim \mathcal{B}e(g_0,\ h_0)$, where $\mathcal{B}e(g_0,\ h_0)$ is the beta distribution with parameters $g_0, h_0 > 0, \forall\, k = 1, \ldots, K$, and

$$p(\mathbf{D}(n, m) < d \mid z_{n\to m}, z_{m\to n}, \beta_{z_{n\to m}}) = \begin{cases} \beta_{z_{n\to m}}, & \text{if } z_{n\to m} = z_{m\to n}, \\ \epsilon, & \text{if } z_{n\to m} \neq z_{m\to n}, \end{cases} \tag{9}$$

with ϵ being a small constant and d is the threshold below which we believe that two arrays are near to each other. In (9), when array n and array m observe the same source (i.e., $z_{n\to m} = z_{m\to n}$), the probability of $\mathbf{D}(n, m) < d$ (i.e., array n and array m are near) is $\beta_{z_{n\to m}}$, which is much larger than ϵ. Here ϵ is the probability of $\mathbf{D}(n, m) < d$ when array n and array m observe different sources. After integrating out $z_{n\to m}, z_{m\to n}$, we obtain

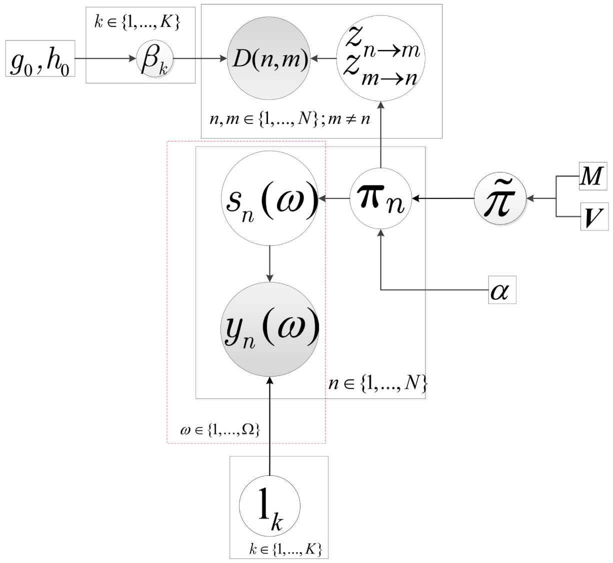

Fig. 2. Our proposed Bayesian network model.

$$p(\mathbf{D}(n, m) < d \mid \boldsymbol{\pi}_n, \boldsymbol{\pi}_m, \boldsymbol{\beta}) = \sum_{k=1}^{K} \pi_{n,k}\pi_{m,k}\beta_k. \tag{10}$$

If $\mathbf{D}(n, m) < d$, we say array n and array m are near to each other. From (10), it can be observed that the probability that array n and array m are close to each other will be high when $\boldsymbol{\pi}_n$ and $\boldsymbol{\pi}_m$ have larger cosine similarity, which means that σ_n and σ_m have high probability of being the same prior, i.e., array n and array m have high probability of observing the same acoustic source. The general Bayesian network model is shown in Fig. 2. The notations and their corresponding meanings are shown in Table I.

III. Inference Algorithm

Our proposed model tries to associate the observations in each array and across different arrays and estimate the locations of sound sources. In this section, we will present our inference algorithm for our Bayesian network model. Usually two kinds of methods are used to infer the parameters of the Bayesian network model, namely Markov Chain Monte Carlo (MCMC) and variational inference method. MCMC is a sampling-based method and can achieve global optimal estimation given infinite number of iterations but it is computationally expensive. For sound source localization applications, the variational inference method is employed due to its properties of guaranteed and fast convergence [51]. However, in our model, priors of some variables are not conjugate to their corresponding likelihood distributions, the traditional variational inference method can thus not be directly used and specific approximations are required.

The hidden random variables we need to estimate are $\mathbf{z} \triangleq (z_{n\to m})_{n,m,n\neq m}$, $\boldsymbol{\beta} \triangleq (\beta_k)_k$, $\boldsymbol{\pi} \triangleq (\boldsymbol{\pi}_n)_n$, $\widetilde{\boldsymbol{\pi}} \triangleq (\widetilde{\pi}_k)_k$, $\mathbf{l} \triangleq (\mathbf{l}_k)_k$, and $\mathbf{s} = (s_n(\omega))_{n,\omega}$. Let $\boldsymbol{\Upsilon} \triangleq \{\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\pi}, \widetilde{\boldsymbol{\pi}}, \mathbf{l}, \mathbf{s}\}$. We aim

TABLE I
SUMMARY OF COMMONLY-USED SYMBOLS

Algorithm	DUET Measurement Extraction
input	STFT of the received microphone signals $\mathbf{S}$
output	TDOAs over TF bins and microphones $\mathbf{y} = (y_n(\omega))_{n,\omega}$
ω	TF bin index, in total Ω bins
n, N	Microphone array index n; in total N arrays
k, K	Source index k; in total K sources
Algorithm	Proposed data association algorithm
input	TDOAs over TF bins and microphones $\mathbf{y} = (y_n(\omega))_{n,\omega}$
output	Estimated source locations $\mathbf{l} = (\mathbf{l}_k)_k$
$\mathbf{D}(n, m)$	Distance between array n and array m
$z_{n\to m}$	The group membership indicator of array n when influenced by array m.
β_k	Parameter to denote how dense the k-th group is.
$\boldsymbol{\pi}_n$	The prior distribution parameters of $s_n(\omega)$ and $z_{n\to m}$.
$\widetilde{\boldsymbol{\pi}}$	The prior distribution parameters of $\{\boldsymbol{\pi}_n\}_{n=1}^{N}$.
$\mathbf{l}_k$	The location of source k.
$y_n(\omega)$	The observation of array n that is in the ω-th TF bin.
$s_n(\omega)$	The source index corresponding to $y_n(\omega)$.
$\alpha, (g_0, h_0)$	Known hyper-parameters.

at obtaining the joint posterior distribution of hidden variables $p(\boldsymbol{\Upsilon} \mid \mathbf{y}, \mathbf{D})$. In the variational inference method, we aim to find a distribution $q(\boldsymbol{\Upsilon})$ from a distribution family $\mathcal{F}$ to minimize the KL-divergence between $q(\boldsymbol{\Upsilon})$ and $p(\boldsymbol{\Upsilon} \mid \mathbf{y}, \mathbf{D})$. $q(\boldsymbol{\Upsilon})$ is called the joint variational distribution. Following [52], we choose $\mathcal{F}$ to be the mean-field distribution family so that the model is efficient to infer, though we need to sacrifice the optimality. Distributions in $\mathcal{F}$ are distinguished by variational parameters (i.e., parameters of the joint variational distribution) and the optimal distribution $q(\boldsymbol{\Upsilon})$ is found by iteratively updating the variational parameters. We assign a variational parameter for each of the hidden variables in $\boldsymbol{\Upsilon}$, they are

$$\begin{aligned}
\boldsymbol{\Lambda} = \{&\boldsymbol{\phi} \triangleq (\boldsymbol{\phi}_{n\to m,k})_{n,m,n\neq m,k},\\
&\boldsymbol{\lambda} \triangleq (\boldsymbol{\lambda}_k)_k,\\
&\boldsymbol{\gamma} \triangleq (\gamma_{n,k})_{n,k},\\
&\boldsymbol{\xi} \triangleq (\xi_k)_k,\\
&\boldsymbol{\mu} \triangleq (\boldsymbol{\mu}_k)_k,\\
&\boldsymbol{\psi} \triangleq (\psi_{n,k}(\omega))_{n,k,\omega}\},
\end{aligned}$$

respectively. We aim to find

$$q^*(\boldsymbol{\Upsilon}) = \underset{q(\boldsymbol{\Upsilon})\in\mathcal{F}}{\arg\min}\, \mathcal{D}_{KL}(q(\boldsymbol{\Upsilon}) \,||\, p(\boldsymbol{\Upsilon} \mid \mathbf{y}, \mathbf{D})), \tag{11}$$

where $\mathcal{D}_{KL}(\cdot \,||\, \cdot)$ is the KL divergence. From [52], solving (11) is equivalent to maximizing the evidence lower bound

$$\mathcal{L}(q) \triangleq \mathbb{E}_{q(\boldsymbol{\Upsilon})}[\log p(\boldsymbol{\Upsilon}, \mathbf{y}, \mathbf{D})] - \mathbb{E}_{q(\boldsymbol{\Upsilon})}[\log q(\boldsymbol{\Upsilon})]. \tag{12}$$

We solve this problem by iteratively updating the variational parameters according to the updating equations shown below. The updating of the variational parameters are derived in the appendix. The pseudo code of our algorithm is shown in Algorithm 1 and the computation complexity of our algorithm in each iteration is $\mathcal{O}(N^2K)$, where N and K represent the number of arrays and sound sources, respectively.

Algorithm 1: Proposed Multiple ASL Method (i-th Iteration).

```
Input: Variational parameters in the $(i-1)$-th iteration,
  observations $\mathbf{y}$, and distance matrix $\mathbf{D}$.
Output: Variational parameters in the $i$-th iteration.
  for each array $n$ in $\{1, \ldots, N\}$ do
    for each array pair $(n, m)$ in $\{(n, m)\}_{m=1}^{N}$do
      Update $\boldsymbol{\phi}_{n\to m}$ and $\boldsymbol{\phi}_{m\to n}$ using (16) and (17).
    end for
    Update $\boldsymbol{\psi}_n$ using (18).
  Update $\boldsymbol{\gamma}_n$ using (19).
  end for
  Update $\boldsymbol{\xi}$ using (24).
  Update $\boldsymbol{\lambda}$ using (13) and (14).
  Update $\boldsymbol{\mu}$ using (22).
  return $\boldsymbol{\phi}$, $\boldsymbol{\psi}$, $\boldsymbol{\gamma}$, $\boldsymbol{\xi}$, $\boldsymbol{\lambda}$, and $\boldsymbol{\mu}$.
```

A. Hyper Parameters $\boldsymbol{\beta}$

We denote $\boldsymbol{\lambda}_k$ as (G_k, H_k) and let the variational distribution of β_k be $q(\beta_k) \triangleq \mathcal{B}e(G_k,\ H_k)$. From [52], we have

$$\begin{aligned}
G_k &= \mathbb{E}_{q(\mathbf{z})}\Bigg[\sum_{(n,m)} I(\mathbf{D}(n, m) < d) I(z_{n\to m}, k) I(z_{m\to n}, k)\\
&\qquad + g_0\Bigg]\\
&= \sum_{(n,m)} I(\mathbf{D}(n, m) < d)\phi_{n\to m,k}\phi_{m\to n,k} + g_0,
\end{aligned} \tag{13}$$

$$\begin{aligned} H_k &= \mathbb{E}_{q(\mathbf{z})}\left[\sum_{(n,m)} I(\mathbf{D}(n,m) \geq d) I(z_{n\to m}, k) I(z_{m\to n}, k) \right. \\ &\qquad \left. + h_0 \right] \\ &= \sum_{(n,m)} I(\mathbf{D}(n,m) \geq d) \phi_{n\to m,k} \phi_{m\to n,k} + h_0, \end{aligned} \tag{14}$$

where $\phi_{n\to m,k}$ is defined in (15) as $q(z_{n\to m} = k) \triangleq \phi_{n\to m,k}$. We use β_k to denote how dense the k-th group is. From (13) and (14), it is observed that the variational distribution of $\boldsymbol{\beta}$ is related to the group membership of arrays $\mathbf{z}$ and the distance relationships among arrays $\mathbf{D}$. We also have

$$\begin{aligned} \mathbb{E}_{q(\beta_k)}[\log(\beta_k)] &= \Psi(G_k) - \Psi(G_k + H_k), \text{ and} \\ \mathbb{E}_{q(\beta_k)}[\log(1-\beta_k)] &= \Psi(H_k) - \Psi(G_k + H_k), \end{aligned}$$

which will be used in Section III-B.

B. Group Membership Indicators z

We let the variational distribution of the group membership index z be

$$q(z_{n\to m} = k) \triangleq \phi_{n\to m,k}. \tag{15}$$

From equation (17) in [52], we have

$$\begin{aligned} &\phi_{n\to m,k} \mid \mathbf{D}(n,m) < d \\ &\quad \propto \exp\{\phi_{m\to n,k} \mathbb{E}_{q(\beta_k)}[\log(\beta_k)] + (1 - \phi_{m\to n,k}) \log \epsilon \\ &\qquad + \mathbb{E}_{q(\pi_{n,k})}[\log(\pi_{n,k})]\}, \end{aligned} \tag{16}$$

where ϵ is a small constant. Thus, $\log \epsilon < \mathbb{E}_{q(\beta_k)}[\log(\beta_k)] < 0$ and $\phi_{n\to m,k}$ (i.e., $q(z_{n\to m} = k)$) increases with $\phi_{m\to n,k}$ when array n and array m are close. Equation (16) holds because of the mean field assumption and $\mathbb{E}[I(z_{m\to n}, k)] = \phi_{m\to n,k}$. Similarly, we have

$$\begin{aligned} &\phi_{n\to m,k} \mid \mathbf{D}(n,m) \geq d \\ &\quad \propto \exp\{\phi_{m\to n,k} \mathbb{E}_{q(\beta_k)}[\log(1-\beta_k)] \\ &\qquad + (1 - \phi_{m\to n,k}) \log(1-\epsilon) \\ &\qquad + \mathbb{E}_{q(\pi_{n,k})}[\log(\pi_{n,k})]\}. \end{aligned} \tag{17}$$

Usually the term $(1 - \phi_{m\to n,k}) \log(1-\epsilon)$ in (17) can be ignored when ϵ is small.

C. Source Indices s

Let the variational distribution of community index s be $q(s_n(\omega) = k) = \psi_{n,k}(\omega)$. Then, we have

$$\begin{aligned} &\psi_{n,k}(\omega) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}[V'_{n,k} + (M'_{n,k} - y_n(\omega))^2] + \mathbb{E}_{q(\boldsymbol{\pi}_n)}[\log(\pi_{n,k})]\right) \end{aligned} \tag{18}$$

where $\mathbb{E}_{q(\boldsymbol{\pi}_n)}[\log(\pi_{n,k})]$ is computed using (20), $M'_{n,k} \triangleq \mathbb{E}_{q(\mathbf{l}_k)}[f(\mathbf{l}_k, \mathbf{l}'_n)]$, and $V'_{n,k} \triangleq \mathbb{E}_{q(\mathbf{l}_k)}[(f(\mathbf{l}_k, \mathbf{l}'_n) - M'_{n,k})^2]$. Both $M'_{n,k}$ and $V'_{n,k}$ are computed with the Monte Carlo method and the samples of $\mathbf{l}_k$ are drawn from (21).

D. Source Weights $\boldsymbol{\pi}$

Let the variational distribution of the mixture weight $\boldsymbol{\pi}_n$ be $q(\boldsymbol{\pi}_n) \triangleq \mathrm{Dir}(\boldsymbol{\gamma}_n)$, where γ_n is a K dimensional vector, K is the number of sources, then we have

$$\begin{aligned} \gamma_{n,k} &= \mathbb{E}_{q(\widetilde{\pi}_k)}[\widetilde{\pi}_k] + \sum_{m=1, m\neq n}^{N} \phi_{n\to m,k} + \sum_{\omega=1}^{\Omega} \psi_{n,k}(\omega) \\ &= \xi_k + \sum_{m=1, m\neq n}^{N} \phi_{n\to m,k} + \sum_{\omega=1}^{\Omega} \psi_{n,k}(\omega), \end{aligned} \tag{19}$$

where ξ_k is obtained from (24). Equation (24) integrates $\sum_m \phi_{n\to m,k}$ (which is related to the clustering of array network) and $\sum_{\omega=1}^{\Omega} \psi_{n,k}(\omega)$ (which is related to the clustering of observations). Besides, according to the property of the Dirichlet distribution, we have

$$\mathbb{E}_q[\log(\pi_{n,k})] = \Psi(\gamma_{n,k}) - \Psi\left(\sum_{k=1}^{K} \gamma_{n,k}\right). \tag{20}$$

E. Position of Sources

We use Laplace approximation method proposed in [53] to find a Gaussian approximation of the variational distribution $q(\mathbf{l}_k)$, given as

$$q(\mathbf{l}_k) \approx \mathcal{N}\left(\boldsymbol{\mu}_k, -\frac{1}{\nabla^2 \log\left(q(\boldsymbol{\mu}_k)\right)}\right), \tag{21}$$

Laplace approximations adopts a Taylor approximation around the maximum a posterior (MAP) point of the target distribution. Thus $\boldsymbol{\mu}_k$ is given by

$$\boldsymbol{\mu}_k = \arg\max_{\mathbf{l}_k} \log q(\mathbf{l}_k), \tag{22}$$

which can be solved using the gradient descent algorithm.

F. Prior Distribution Parameter $\widetilde{\boldsymbol{\pi}}$

We approximate $q(\widetilde{\boldsymbol{\pi}})$ using a normal distribution and obtain

$$q(\widetilde{\boldsymbol{\pi}}) \approx \mathcal{N}\left(\boldsymbol{\xi}, -\frac{1}{\nabla^2 \log\left(q(\boldsymbol{\xi})\right)}\right), \tag{23}$$

where $\boldsymbol{\xi}$ is given by

$$\boldsymbol{\xi} = \arg\max_{\widetilde{\boldsymbol{\pi}}} \log q(\widetilde{\boldsymbol{\pi}}), \tag{24}$$

which can also be solved using the gradient descent algorithm.

IV. Simulation and Experiment Results

In this section, we evaluate our method on both simulated datasets and real recordings, and compare it with the state-of-art method proposed in [42].

TABLE II
SIMULATION SETUP AND PARAMETER INITIALIZATION OF ALGORITHM 1

Hyper-parameter and initial value	Value
Number of iterations	30
g_0 in $\beta_k \sim \mathcal{B}e\,(g_0,\ h_0)$	0.1
h_0 in $\beta_k \sim \mathcal{B}e\,(g_0,\ h_0)$	0.1
Number of sources K	2 or 3
Initial value of α in (7)	$0.1 \times \text{ones}(1, K)$
Initial value of $\boldsymbol{\psi}$ in (18)	$\text{ones}(N, K, \Omega)/K$
Initial value of $\mathbb{E}_q[\log(\pi_{n,k})]$ in (18)	$\Psi(\gamma(n,k)) - \Psi(\sum_k \gamma(n,k))$, where $\Psi(\cdot)$ is the digamma function and $\gamma(n,k) \sim \boldsymbol{\tau}\mathcal{N}\,(1,\ 0.001)$ with $\boldsymbol{\tau}\mathcal{N}\,(\cdot,\ \cdot)$ being a truncated normal distribution.
Initial value of $\widetilde{\boldsymbol{\pi}}$	$0.1 \times \text{ones}(1, K)$
ϵ	$\exp(-5)$
σ^2	0.1 when $K = 3$ and 0.2 when $K = 2$

A. 2D Localization in Anechoic Environments

In this simulation, 2D localization and anechoic environment are considered. Our objective is to evaluate the data association accuracy and the localization accuracy of our method against the baseline method proposed in [42] in ideal conditions.

1) Simulation Settings: To generate the dataset, we consider a 2D space with length and width both 20 decimeters. 40 microphone arrays are randomly deployed in the space and each array is composed of two microphones. The distance between two microphones in an array is 2 decimeters. The position of each array is fixed at the central position of the corresponding microphone pair. We consider two or three sound source scenarios and the positions of the sound sources are also randomly generated. The observation of each array is divided into 40 TF bins and each TF bin belongs to an acoustic source. In this subsection, the total number of TF bins is 40 and in the simulation of our method and the baseline method, we use all the TF bins. Here we let the total number of TF bins be 40 to show the effectiveness of our method when limited number of observations are available. In Section IV-B, we will show our method performs better than the baseline method when the number of TF bins is much larger. The number and the positions of arrays are assumed to be known and the number of acoustic sources is also known.

When performing simulations, The estimated locations of acoustic sources are randomly initialized in the space in our simulation. The values of the hyperparameters and parameter initialization in our model are shown in Table II. The localization performance is evaluated using the Mean Localization Error (MLoE), which is defined as follows:

$$MLoE = \frac{1}{K}\sum_{k=1}^{K} \min_{j\in\{1,\ldots,K\}} \left\|\hat{\mathbf{l}}_j - \mathbf{l}_k\right\|, \tag{25}$$

where $\{\hat{\mathbf{l}}_j\}_{j=1}^{K}$ and $\{\mathbf{l}_k\}_{k=1}^{K}$ are the estimations and the ground truths of the source locations. For an element $\mathbf{l}_k$ in $\{\mathbf{l}_k\}_{k=1}^{K}$, we choose the closest element to $\mathbf{l}_k$ in $\{\hat{\mathbf{l}}_j\}_{j=1}^{K}$ as its estimation and calculate the mean square error over K sources.

Following [42], the mean association error (MAsE) is employed to evaluate the performance of the data association. MAsE counts the percentage of wrong pairwise associations between all pairs of arrays. In essence, the lower the MAsE is, the less impact an erroneous pair will have on the data-association and thus to the localization error.

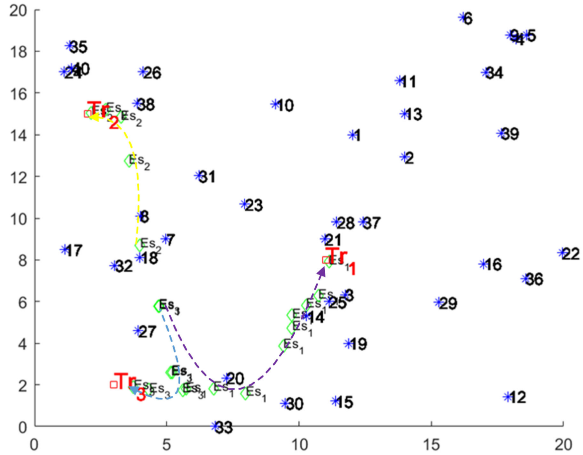

Fig. 3. Demonstration of the estimates approaching the ground truth in a single implementation. The symbols '*', 'Tr$_k$' and 'Es$_k$' denote the sensors, the ground truth location of the k-th source and the estimated location of the k-th source in our algorithm. The three arrows show the trajectories of our estimation results in different iterations.

2) Single Experiment Result: In this experiment, we demonstrate how the estimates approach the ground truth in a single implementation. Fig. 3 shows the experiment setup and the result. We use '*', 'Tr$_k$' and 'Es$_k$' to denote the sensors, the true location of the k-th source and the estimated location of the k-th source in our algorithm. The three arrows show the trajectories of our estimation results in different iterations. It can be observed that even the sources are randomly initialized to the same position, the algorithm can still converge to the ground truth quickly.

3) Mean Localization Error: We conduct 50 Monte Carlo experiments for both two and three source scenarios and show the localization performances of our method and the baseline method in Fig. 4. The performances are evaluated using Eq. (25). The boxes in the figure show the MLoE distributions of our method and the baseline method. It can be observed that our proposed method achieves lower MLoE median and lower MLoE variance in both two and three source scenarios. In our method, the MLoE medians are 0.07 decimeter and 0.14

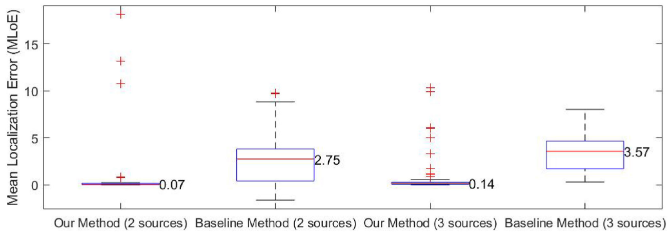

Fig. 4. Mean localization error comparison between our method and the baseline method under two sources and three sources scenarios. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles separately. The outliers are indicated by '+'.

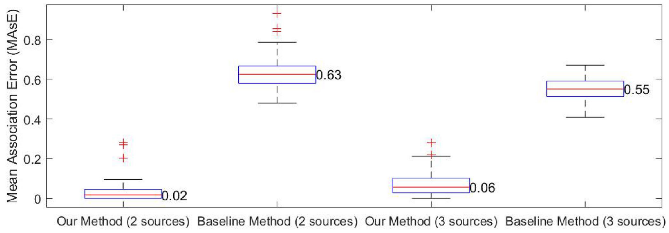

Fig. 5. Mean measurement to source association error comparison between our method and the baseline method under two sources and three sources scenarios. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles separately. The outliers are indicated by '+'.

decimeter for two sources and three sources separately. In the baseline method, however, the medians are 2.75 decimeter and 3.57 decimeter for corresponding cases. The performance of our method is better than that of the baseline method due to two reasons. First, the distance information between arrays are employed and close arrays have high probability of observing the same sound sources in our model. The other reason is that we only use 40 TF bins and the baseline method cannot obtain accurate histogram based features using such a small number of observations. The result also shows the robustness of our method when the number of TF bins is small.

4) Mean Association Error: The performance of data association is shown in Fig. 5. The mean association error (MAsE) is computed across different arrays. It can be observed that the MAsE medians of our method are 2% and 6% for two sources and three sources separately. In comparison, the MAsE medians of the baseline method are 63% and 55% correspondingly. The MAsE variances of our method are also significantly lower than those of the baseline method.

B. 3D localization with Reverberation and Noise

In this dataset, we consider localization of the sources in 3D space with consideration of room reverberation. By considering various reverberation time and signal-to-noise ratio, the robustness of our method against model mismatch can be studied. We do not evaluate the data association performance of our method and the baseline method on this simulation as the ground truth of the data associations is not available due to reverberation and noise.

1) Data Generation Process: We use open source software Pyroomacoustics[1] to generate this simulation dataset. This software provides room impulse response (RIR) simulations via the imaging method. The length, width, and height of the room is set to be [15, 10, 3]m. 20 arrays are randomly deployed in the room with height fixed to 1.5 m. The source positions are randomly generated in the localization scene; the distance between the wall and the sources are assumed to be larger than 1m. We perform simulations for both two and three source scenarios. Various reverberation time (RT60) (i.e., 250 ms, 400 ms, 600 ms) and signal to noise ratio (SNR) (i.e., 0 dB, 10 dB, 20 dB) are considered in our experiments.

Pyroomacoustics simulates the sound propagation in the room using our above settings. We obtain the audio signal received by each microphone, which is a mixture of multiple speech signals due to different sources. The audio signals are sampled at 16kHz. We compute short time Fourier transform (STFT) of the received signal of each microphone and obtained the phase of each time frequency bin. For each TF bin, we extract the TDOA and then convert it to the distance difference. The number of time frames is 108 with 50% overlap between adjacent frames. At each array, 108×512 TF bins are available to obtain the observations in our model. We select the TF bins with amplitude higher than a predefined threshold so that non-informative bins due to noise and reverberation can be removed. The localization scenes are illustrated in Fig. 6.

2) Mean Localization Error Under Different Reverberation Time (RT60): The estimated locations of acoustic sources are randomly initialized in the space in our experiment. The values of the hyperparameters and initial values of some variables are the same as those shown in Table II except that the number of iterations is set to 100. The observation association becomes more difficult when the reverberation is considered. We set the SNR in our simulation to be 20 dB and conduct 20 MC experiments for different RT60 values, namely $\{250, 400, 600\}$ms. The results are shown in Fig. 7. It can be observed that the proposed method has lower MLoE median, and is more robust to the reverberation compared with the baseline method. It is worth mentioning that both the baseline method [42] and our proposed method are two stage methods, i.e., extracting the measurements first and then performing the association. Hence, the association algorithms in our method and in the baseline method can directly deal with both DOAs and TDOAs. Besides, when implementing the baseline method, we use the same observation association and localization methods as those in [42].

3) Mean Localization Error Under Different SNRs: We use the same settings as in Section IV-B2. In this subsection, we consider the influence of different SNR values on the performance of our method and the baseline method. We set the RT60 in our simulations to be 250ms and conducted 20 MC experiments for different SNRs in $\{0, 10, 20\}$dB. The results are shown in Fig. 8. It shows that our method has better performance than the baseline method when SNR are 10 dB and 20 dB. The performance of our method is worse than the baseline method when SNR is 0dB which is an extreme challenge environment that both methods present high localization error.

[1] [Online]. Available: https://pypi.org/project/pyroomacoustics/0.4.1/

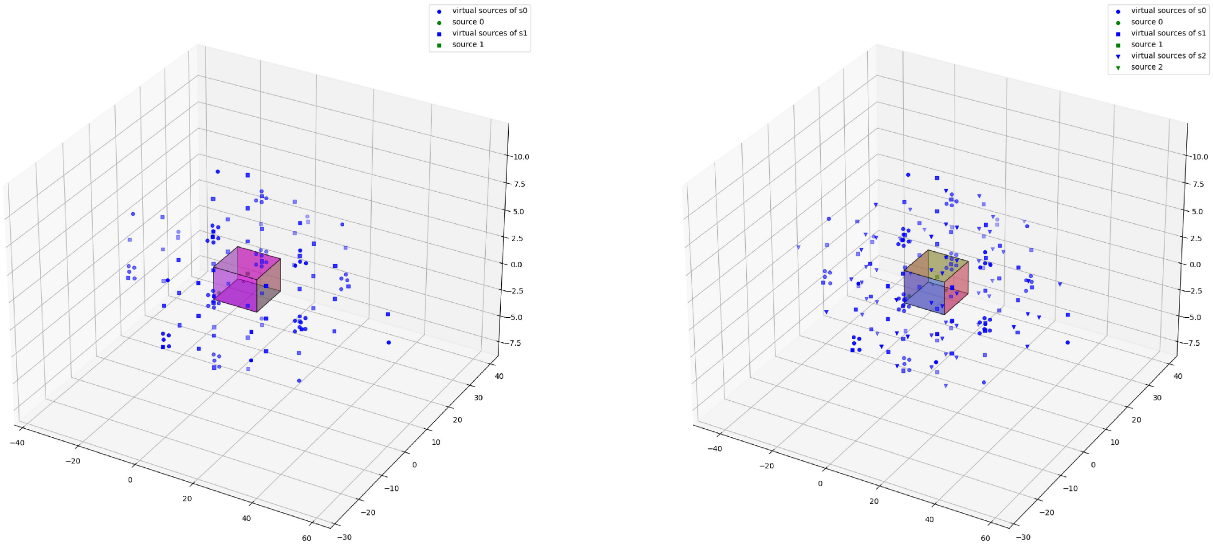

Fig. 6. Rooms with two (left) and three (right) sources. The boxes represent the boundary of the room. The markers represent either a real sound source or a virtual source generated by using the imaging method.

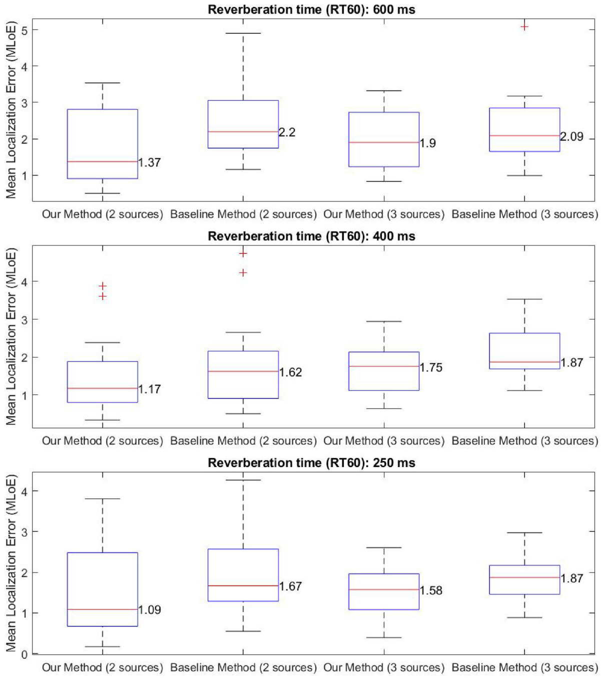

Fig. 7. Mean Localization Error (MLoE) under different reverberation time values. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles separately. The outliers are indicated by '+'.

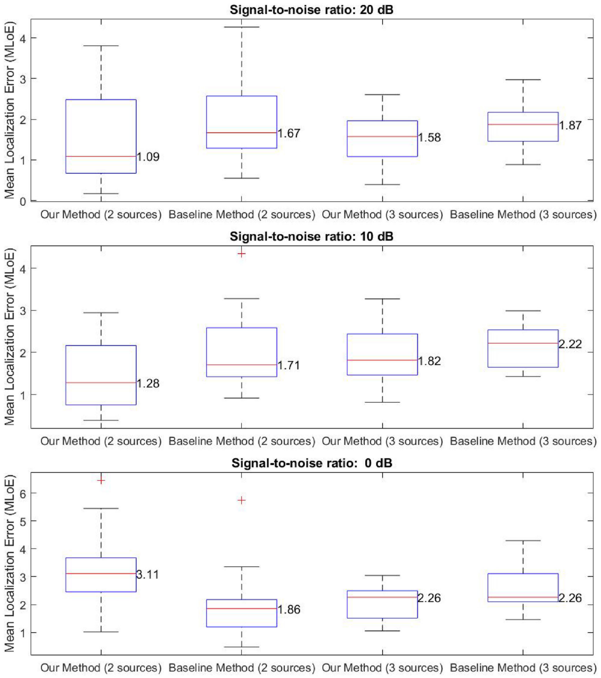

Fig. 8. Mean Localization Error (MLoE) under different SNR values. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles separately. The outliers are indicated by '+'.

C. *Real Data Experiment*

To further demonstrate the performance of our method in practice, an experiment with real recordings in a lecture room is conducted.

1) Recording Environment: The experiment settings are illustrated in Fig. 9. The data set was recorded in a lecture room of which the size is $[10.5, 8.9, 3.9]$ m. The room has two wooden doors and tiled floor. Three walls are concrete blocks and the other one is mainly glass windows. There are desks and seats in the room. The measured reverberation time of the room is about 880 ms. Six omni-directional microphone arrays are placed on the table with a height of 0.79 m. Two different types of microphone arrays are employed for recording: three uniform linear arrays with 4 cm spacing and 4 microphones, and three uniform circular arrays with 4.67 cm radius and also 4

TABLE III
MLoE FOR THE BASELINE AND PROPOSED METHOD IN REAL DATA EXPERIMENT

Method	**Number of Sources**	**MLoE**					
		rec 1	**rec 2**	**rec 3**	**rec 4**	**rec 5**	**Average**
Proposed	2	2.50	2.41	2.27	2.83	2.06	2.42
	3	2.57	2.77	2.43	2.81	2.59	2.63
Baseline	2	3.86	3.49	3.63	3.88	3.39	3.65
	3	3.35	3.50	4.67	3.39	2.97	3.57

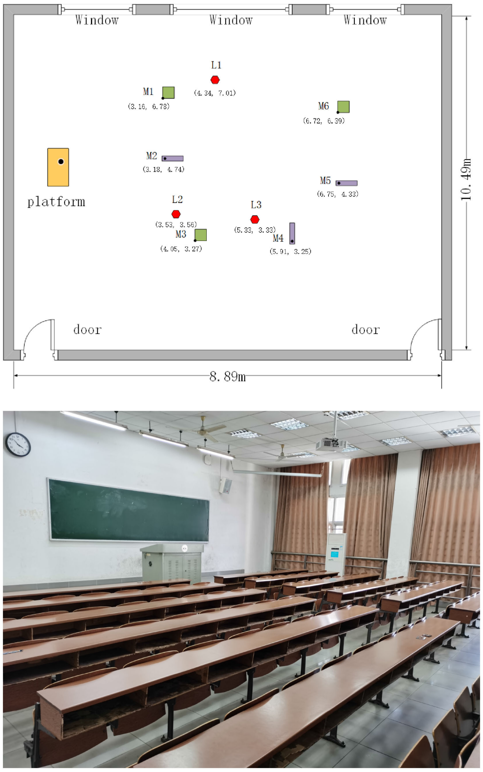

Fig. 9. Real data experiment setting. Top: a schematic of the recording setup; M1, M3 and M6 are three circular microphone arrays. M2, M4 and M5 are three linear microphone arrays. The position of a microphone (marked by a dark dot) in the array is given next to it; L1, L2 and L3 are the positions of three speakers; the coordinate are shown in meter. Bottom: Real lecture room environment.

microphones. The signals received by the diagonal microphones in the circular array and the signals from the microphones at two ends of the linear array are employed for experiments. Three different talkers (1 male and 2 female) are sitting at the given position with height of $[1.50, 1.39, 1.39]$ m to speak as acoustic sources. Five different recordings for both two speaker and three speaker simultaneously talking scenarios are considered.

2) Mean Localization Error: The values of the hyperparameters and initial values of variables are the same as those in Section IV-B2. The experiment results are shown in Table III. It can be observed that our method has better performance than the baseline method on all real recordings.

V. CONCLUSION

In this paper, we propose a Bayesian network model to jointly infer the measurement-source association and locations of multiple speaker sources. The proposed approach is able to incorporate the information of distances between microphone arrays to reduce the performance degradation due to reverberation and noise. Experiments on both simulated environments and real recordings are performed. The mean association error and the mean location error are employed to evaluate the performance of the proposed method. The results show the advantage of our method in assigning TDOAs and localizing sound sources under different environments. However, the number of sources are assumed to be known and fixed in this paper. In our future work, dynamic source and joint detection and localization problem will be considered and Bayesian network based tracking algorithm will be studied.

APPENDIX

In the appendix, we show the derivation details of the variational parameter updating equations.

A. Hyper Parameters β

We use $I(\cdot)$ and $I(\cdot,\cdot)$ to denote the indicator functions and they are defined at the end of Section I. The posterior distribution of β_k is

$$
\begin{aligned}
&p(\beta_k \mid \mathbf{D}, \mathbf{z})\\
&\propto \prod_{(n,m)} p(\mathbf{D}(n,m) \mid \beta_k, z_{n\to m} = k, z_{n\leftarrow m} = k)p(\beta_k)\\
&\propto \prod_{(n,m)} \beta_k^{I(\mathbf{D}(n,m)<d)I(z_{n\to m},k)I(z_{m\to n},k)}\\
&\quad (1-\beta_k)^{I(\mathbf{D}(n,m)\geq d)I(z_{n\to m},k)I(z_{m\to n},k)}\mathcal{B}e\left(g_0,\ h_0\right)\\
&= \beta_k^{\sum_{(n,m)} I(\mathbf{D}(n,m)<d)I(z_{n\to m},k)I(z_{m\to n},k)+g_0-1}\\
&\quad (1-\beta_k)^{\sum_{(n,m)} I(\mathbf{D}(n,m)\geq d)I(z_{n\to m},k)I(z_{m\to n},k)+h_0-1}\\
&= \mathcal{B}e\left(\sum_{(n,m)} I(\mathbf{D}(n,m) < d)I(z_{n\to m},k)I(z_{m\to n},k) + g_0,\right.\\
&\qquad \left.\sum_{(n,m)} I(\mathbf{D}(n,m) \geq d)I(z_{n\to m},k)I(z_{m\to n},k) + h_0\right).
\end{aligned}
$$

This posterior distribution is also a beta distribution. The beta distribution is an exponential family distribution and its natural parameter is given by

$$\left(\sum_{(n,m)} I\left(\mathbf{D}(n,m) < d\right) I(z_{n\to m},k) I(z_{m\to n},k) + g_0,\right.$$
$$\left.\sum_{(n,m)} I(\mathbf{D}(n,m) \geq d) I(z_{n\to m},k) I(z_{m\to n},k) + h_0\right).$$

We further denote $\boldsymbol{\lambda}_k$ as (G_k, H_k) and let the variational distribution of β_k be $q(\beta_k) \triangleq \mathcal{B}e(G_k,\ H_k)$. From [42], we obtain (13) and (14).

B. Group Membership Indicators $\mathbf{z}$

The posterior distribution of $z_{n\to m}$ is

$$\begin{aligned} p(z_{n\to m} = k \mid \pi_n, z_{m\to n}, \mathbf{D}(n,m) < d, \beta_k) &\\ \propto p(\mathbf{D}(n,m) < d \mid z_{n\to m} = k, \pi_n, z_{m\to n}, \beta_k)&\\ p(z_{n\to m} = k \mid \pi_n)&\\ = \beta_k^{I(z_{m\to n},k)} \epsilon^{(1-I(z_{m\to n},k))} \pi_{n,k}\ .& \end{aligned}$$

Similarly, we can also derive

$$\begin{aligned} &p(z_{n\to m} = k \mid \pi_n, z_{m\to n}, \mathbf{D}(n,m) \geq d, \beta_k)\\ &\quad \propto (1-\beta_k)^{I(z_{m\to n},k)} (1-\epsilon)^{(1-I(z_{m\to n},k))} \pi_{n,k}. \end{aligned}$$

We let the variational distribution of the group membership index z be

$$q(z_{n\to m} = k) \triangleq \phi_{n\to m,k}. \tag{26}$$

From [42], we have

$$\begin{aligned} &\phi_{n\to m,k} \mid \mathbf{D}(n,m) < d\\ &\quad \propto \exp\{\mathbb{E}_{q(\beta_k, z_{m\to n}, \pi_{n,k})}[\log(\beta_k^{I(z_{m\to n},k)} \epsilon^{(1-I(z_{m\to n},k))} \pi_{n,k})]\}\\ &\quad = \exp\{\phi_{m\to n,k}\mathbb{E}_{q(\beta_k)}[\log(\beta_k)] + (1-\phi_{m\to n,k}) \log \epsilon\\ &\qquad + \mathbb{E}_{q(\pi_{n,k})}[\log(\pi_{n,k})]\}. \end{aligned} \tag{27}$$

Equation (27) holds because of the mean field assumption and $\mathbb{E}[I(z_{m\to n},k)] = \phi_{m\to n,k}$. Similarly, we have

$$\begin{aligned} &\phi_{n\to m,k} \mid \mathbf{D}(n,m) \geq d\\ &\quad \propto \exp\{\mathbb{E}_{q(\beta_k, z_{m\to n}, \pi_{n,k})}[\log((1-\beta_k)^{I(z_{m\to n},k)}\\ &\qquad (1-\epsilon)^{(1-I(z_{m\to n},k))} \pi_{n,k})]\}\\ &\quad = \exp\{\phi_{m\to n,k}\mathbb{E}_{q(\beta_k)}[\log(1-\beta_k)]\\ &\qquad + (1-\phi_{m\to n,k}) \log(1-\epsilon)\\ &\qquad + \mathbb{E}_{q(\pi_{n,k})}[\log(\pi_{n,k})]\}. \end{aligned} \tag{28}$$

C. Source Indices $\mathbf{s}$

The posterior distribution of $s_n(\omega)$ is

$$\begin{aligned} &p(s_n(\omega) = k \mid \pi_n, y_n(\omega), \mathbf{l}_k)\\ &\quad \propto p(y_n(\omega) \mid s_n(\omega) = k, \mathbf{l}_k) p(s_n(\omega) = k \mid \pi_n)\\ &\quad = \mathcal{N}\big(f(\mathbf{l}_k, \mathbf{l}'_n),\ \sigma^2\big)\pi_{n,k}. \end{aligned}$$

Here $y_n(\omega) = \text{nan}$ if array n does not observe frequency ω. $s_n(\omega)$ is a discrete variable and follows a categorical distribution and thus we obtain

$$\begin{aligned} &p(s_n(\omega) \mid \pi_n, y_n(\omega), \mathbf{l}_k)\\ &\quad = \mathrm{Cat}\left(\left(\frac{\mathcal{N}\Big(f(\mathbf{l}_k, \mathbf{l}'_n),\ \sigma^2\Big)\pi_{n,k}}{\sum_{k=1}^K \mathcal{N}\Big(f(\mathbf{l}_k, \mathbf{l}'_n),\ \sigma^2\Big)\pi_{n,k}}\right)_{k=1}^{K}\right). \end{aligned}$$

We let the variational distribution of community index s be $q(s_n(\omega) = k) = \psi_{n,k}(\omega)$. Then, we have

$$\begin{aligned} &\psi_{n,k}(\omega)\\ &\propto \exp\{\mathbb{E}_{q(\mathbf{l}_k, \boldsymbol{\pi}_n)}[\log(\mathcal{N}\Big(f(\mathbf{l}_k, \mathbf{l}'_n),\ \sigma^2\Big)\pi_{n,k})]\}\\ &= \exp\Big(\mathbb{E}_{q(\mathbf{l}_k)}[\log(\mathcal{N}\Big(f(\mathbf{l}_k, \mathbf{l}'_n),\ \sigma^2\Big))] + \mathbb{E}_{q(\boldsymbol{\pi}_n)}[\log(\pi_{n,k})]\Big)\\ &= \exp\left(\mathbb{E}_{q(\mathbf{l}_k)}\left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_n(\omega) - f(\mathbf{l}_k, \mathbf{l}'_n))^2}{2\sigma^2}\right)\right)\right]\right.\\ &\qquad \left. + \mathbb{E}_{q(\boldsymbol{\pi}_n)}\left[\log(\pi_{n,k})\right]\right)\\ &\propto \exp\left(\mathbb{E}_{q(\mathbf{l}_k)}\left[-\frac{(y_n(\omega) - f(\mathbf{l}_k, \mathbf{l}'_n))^2}{2\sigma^2}\right] + \mathbb{E}_{q(\boldsymbol{\pi}_n)}[\log(\pi_{n,k})]\right)\\ &\propto \exp\left(\mathbb{E}_{q(\mathbf{l}_k)}\left[-\frac{(f(\mathbf{l}_k, \mathbf{l}'_n) - M'_{n,k} + M'_{n,k} - y_n(\omega))^2}{2\sigma^2}\right]\right.\\ &\qquad \left. + \mathbb{E}_{q(\boldsymbol{\pi}_n)}[\log(\pi_{n,k})]\right)\\ &= \exp\left(-\frac{1}{2\sigma^2}[V'_{n,k} + (M'_{n,k} - y_n(\omega))^2] + \mathbb{E}_{q(\boldsymbol{\pi}_n)}[\log(\pi_{n,k})]\right) \end{aligned} \tag{29}$$

where $\mathbb{E}_{q(\boldsymbol{\pi}_n)}[\log(\pi_{n,k})]$ is computed using (20), $M'_{n,k} \triangleq \mathbb{E}_{q(\mathbf{l}_k)}[f(\mathbf{l}_k, \mathbf{l}'_n)]$, and $V'_{n,k} \triangleq \mathbb{E}_{q(\mathbf{l}_k)}[(f(\mathbf{l}_k, \mathbf{l}'_n) - M'_{n,k})^2]$. Both $M'_{n,k}$ and $V'_{n,k}$ are computed with the Monte Carlo method and the samples of $\mathbf{l}_k$ are drawn from (21).

D. Source Weights $\boldsymbol{\pi}$

The posterior distribution of $\boldsymbol{\pi}_n$ is

$$\begin{aligned} &p(\boldsymbol{\pi}_n \mid \{s_n(\omega)\}_{\omega=1}^{\Omega}, \mathbf{z}, \widetilde{\boldsymbol{\pi}})\\ &= \prod_{\omega=1}^{\Omega} p(s_n(\omega) \mid \boldsymbol{\pi}_n) \prod_{m} p(z_{n\to m} \mid \boldsymbol{\pi}_n) p(\boldsymbol{\pi}_n \mid \widetilde{\boldsymbol{\pi}})\\ &= \mathrm{Dir}\left(\left\{\widetilde{\pi}_k + \sum_{m=1, m\neq n}^{N} I(z_{n\to m}, k) + \sum_{\omega=1}^{\Omega} I(s_n(\omega), k)\right\}_{k=1}^{K}\right), \end{aligned}$$

where $\sum_{m=1,m\neq n}^{N} I(z_{n\to m}, k)$ is corresponding to the clustering of the distance relation network and $\sum_{\omega=1}^{\Omega} I(s_n(\omega), k)$ is corresponding to the clustering of the observations. We let the variational distribution of the mixture weight $\boldsymbol{\pi}_n$ be $q(\boldsymbol{\pi}_n) \triangleq \mathrm{Dir}(\boldsymbol{\gamma}_n)$, where γ_n is a K dimensional vector, K is the maximum number of communities, then we obtain (19).

E. Position of Sources

Assume the prior distributions of sources' locations are uniform distributions. The posterior distribution of $\mathbf{l}_k$ is then proportional to its corresponding likelihood. We obtain the posterior distribution of $\mathbf{l}_k$:

$$\begin{aligned} & p(\mathbf{l}_k \mid \{y_{n,\omega}\}_{n,\omega}, \{s_n(\omega)\}_{n,\omega}) \\ & \propto \prod_{n=1}^{N}\prod_{\omega=1}^{\Omega} p(y_n(\omega) \mid s_n(\omega), \mathbf{l}_k)^{I(s_n(\omega),k)} \\ & = \prod_{n=1}^{N}\prod_{\omega=1}^{\Omega} \mathcal{N}\Big(y_n(\omega); f(\mathbf{l}_k, \mathbf{l}'_n),\ \sigma^2\Big)^{I(s_n(\omega),k)}. \end{aligned}$$

We can see that $p(\mathbf{l}_k \mid \{y_{n,\omega}\}_{n,\omega}, \{s_n(\omega)\}_{n,\omega})$ is not an exponential family distribution. To find $q(\mathbf{l}_k)$ to minimize (12), we take the functional derivative of the objective function (12) with respect to $q(\mathbf{l}_k)$ and set it to zero, namely $\frac{\partial \mathcal{L}(q)}{\partial q(\mathbf{l}_k)} = 0$, we obtain the maximizer as

$$\begin{aligned} q(\mathbf{l}_k) & \propto \exp(\mathbb{E}_{q(\mathbf{s})}[\log p(\mathbf{l}_k \mid \{y_{n,\omega}\}_{n,\omega}, \{s_n(\omega)\}_{n,\omega})]) \\ & \propto \exp\Bigg(\mathbb{E}_{q(\mathbf{s})}\Bigg[\sum_{n=1}^{N}\sum_{\omega=1}^{\Omega}\log\Big[\mathcal{N}\Big(y_n(\omega); f(\mathbf{l}_k, \mathbf{l}'_n),\ \sigma^2\Big)\Big] \\ & \qquad I(s_n(\omega), k)\Bigg]\Bigg) \\ & = \prod_{n=1}^{N}\prod_{\omega=1}^{\Omega}\mathcal{N}\Big(y_n(\omega); f(\mathbf{l}_k, \mathbf{l}'_n),\ \sigma^2\Big)^{\psi_{n,k}(\omega)} \end{aligned} \quad (30)$$

Equation (30) is difficult to analyze and thus we use Laplace approximation method proposed in [53] to find a Gaussian approximation, given as (21) and (22).

F. Prior Distribution Parameter $\widetilde{\boldsymbol{\pi}}$

The posterior distribution of $\widetilde{\boldsymbol{\pi}} = \{\boldsymbol{\pi}_k\}_k$ is

$$\begin{aligned} p(\widetilde{\boldsymbol{\pi}} \mid \boldsymbol{\pi}) & \propto \prod_{n=1}^{N} p(\boldsymbol{\pi}_n \mid \widetilde{\boldsymbol{\pi}}) p(\widetilde{\boldsymbol{\pi}}) \\ & = \prod_{n=1}^{N} \mathrm{Dir}\,(\boldsymbol{\pi}_n; \widetilde{\boldsymbol{\pi}})\,\mathrm{LogNormal}\,(\mathbf{M},\ \mathbf{V}) \end{aligned} \quad (31)$$

Similar to Section III-E, we take the functional derivative of the objective function (12) with respect to $q(\widetilde{\boldsymbol{\pi}})$ and set it to zero, namely $\frac{\partial \mathcal{L}(q)}{\partial q(\widetilde{\boldsymbol{\pi}})} = 0$ and obtain

$$\begin{aligned} q(\widetilde{\boldsymbol{\pi}}) & \propto \exp\left(\sum_{n=1}^{N}\mathbb{E}_{q(\boldsymbol{\pi}_n)}[\log p(\boldsymbol{\pi}_n \mid \widetilde{\boldsymbol{\pi}})] + \log p(\widetilde{\boldsymbol{\pi}})\right) \\ & \propto \exp\Bigg(\sum_{n=1}^{N}\mathbb{E}_{q(\boldsymbol{\pi}_n)}\Bigg[\log\Gamma\Bigg(\sum_{k=1}^{K}\widetilde{\pi}_k\Bigg) - \sum_{k=1}^{K}\log\Gamma(\widetilde{\pi}_k) \\ & \qquad + \sum_{k=1}^{K}(\widetilde{\pi}_k - 1)\mathbb{E}_q[\log(\pi_{n,k})]\Bigg] \\ & \qquad - \frac{K}{2}\log(2\pi) - \frac{1}{2}\log(\det(\mathbf{V})) - \sum_{k=1}^{K}\log(\widetilde{\pi}_k) \\ & \qquad - \frac{1}{2}(\log(\widetilde{\boldsymbol{\pi}}) - \mathbf{M})^T\mathbf{V}^{-1}(\log(\widetilde{\boldsymbol{\pi}}) - \mathbf{M})\Bigg). \end{aligned} \quad (32)$$

The last formula holds due to (7) and (8). We approximate $q(\widetilde{\boldsymbol{\pi}})$ using a normal distribution, given as (23) and (24).

ACKNOWLEDGMENT

The authors would like to thank Dr. Hongbin Wang for providing useful advice.

REFERENCES

[1] M. Brandstein, "A framework for speech source localization using sensor arrays," Ph.D. thesis, Brown University, Providence, USA, 1995.

[2] J. H. Dibiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. thesis, Brown University, Providence, USA, 2000.

[3] Y. Huang, J. Benesty, and G. W. Elko, "Passive acoustic source localization for video camera steering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, Jun. 5–9, 2000, pp. II909–II912.

[4] R. D. D. Zotkin and L. S. Davis, "Multimodal 3-D tracking and event detection via the particle filter," in *Proc. IEEE Workshop Detection Recognit. Events Video*, vol. 2, 2001, pp. 20–27.

[5] M. Brandstein and D. Ward, *Microphone Arrays. Signal Process. Techniques and Applications*. Berlin, Germany: Springer-Verlag, 2001.

[6] F. Talantzis, A. Pnevmatikakis, and A. Constantinides, "Audio-visual active speaker tracking in cluttered indoors environments," *IEEE Trans. Syst., Man, Cybern., B: Cybern.*, vol. 39, no. 1, pp. 7–15, Feb. 2009.

[7] H. Schau and A. Robinson, "Passive source localization employing intersecting spherical surfaces from time-of-arrival differences," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 35, no. 8, pp. 1223–1225, Aug. 1987.

[8] M. Brandstein and H. Silverman, "A practical methodology for speech source localization with microphone arrays," *Comput., Speech, Lang.*, vol. 11, pp. 91–126, Apr. 1997.

[9] Y. Huang, J. Benesty, G. Elko, and R. Mersereati, "Real-time passive source localization: a practical linear-correction least-squares approach," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 943–956, Nov. 2001.

[10] S. Doclo and M. Moonen, "Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments," *EURASIP J. Appl. Signal Process.*, vol. 2003, no. 11, pp. 1110–1124, 2003.

[11] D. Ward, E. Lehmann, and R. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 826–836, Nov. 2003.

[12] T. G. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Signal Process.*, vol. 85, no. 1, pp. 177–204, Jan. 2005.

[13] U. Klee, Tobias, and J. McDonough, "Kalman filters for time delay of arrival-based source localization," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 1–15, 2006.

[14] J. Chen, J. Benesty, and Y. A. Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 1–19, 2006.

[15] E. A. Lehmann, "Particle filtering approach to adaptive time-delay estimation," in *Proc. Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, May 2006, pp. 1129–1132.

[16] S. Gannot and T. G. Dvorkind, "Microphone array speaker localizers using spatial-temporal information," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 1–17, 2006.

[17] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[18] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Amer.*, vol. 107, no. 1, pp. 384–391, Jan. 2000.

[19] X. Zhong and J. Hopgood, "Nonconcurrent multiple speakers tracking based on extended kalman particle filter," in *Proc. Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2008, pp. 293–296.

[20] X. Zhong and J. R. Hopgood, "Particle filtering for time-delay of arrival based room acoustic source tracking: Mutiple nonconcurrent speakers," *Signal Process.*, vol. 96, pp. 382–394, 2014.

[21] M. Brandstein, J. Adcock, and H. Silverman, "A closed-form location estimator for use with room environment microphone arrays," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 1, pp. 45–50, Jan. 1997.

[22] S. J. Spencer, "The two-dimensional source location problem for time differences of arrival at minimal element monitoring arrays," *J. Acoust. Soc. Amer.*, vol. 121, no. 6, pp. 3579–3594, 2007.

[23] J. Velasco, D. Pizarro, J. Macias-Guarasa, and A. Asaei, "Tdoa matrices: Algebraic properties and their application to robust denoising with missing data," *IEEE Trans. Signal Process.*, vol. 64, no. 20, pp. 5242–5254, 2016.

[24] M. Compagnoni *et al.*, "A geometrical–statistical approach to outlier removal for TDOA measurements," *IEEE Trans. Signal Process.*, vol. 65, no. 15, pp. 3960–3975, Aug. 2017.

[25] X. Alameda-Pineda and R. Horaud, "A geometric approach to sound source localization from time-delay estimates," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 6, pp. 1082–1095, Jun. 2014.

[26] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[27] S. Makino, T. W. Lee, and H. Sawada, Eds., *Blind Speech Separation*. Berlin, Germany: Springer, 2007.

[28] M. I. Mandel and D. P. W. Ellis, "EM localization and separation using interaural level and phase cues," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.,* Oct. 2007, pp. 275–278.

[29] A. Brutti, M. Omologo, and P. Svaizer, "Localization of multiple speakers based on a two step acoustic map analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2008, pp. 4349–4352.

[30] M. S. Brandstein, "Time-delay estimation of reverberated speech exploiting harmonic structure," *J. Acoust. Soc. Amer.*, vol. 105, no. 5, pp. 2914–2919, 1999.

[31] H. Christensen, N. Ma, S. Wrigley, and J. Barker, "A speech fragment approach to localising multiple speakers in reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2009, pp. 4593–4596.

[32] M. Mandel, R. Weiss, and D. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 382–394, Feb. 2010.

[33] A. Lombard, Y. Zheng, H. Buchner, and W. Kellermann, "Tdoa estimation for multiple sound sources in noisy and reverberant environments using broadband independent component analysis," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 19, no. 6, pp. 1490–1503, Aug. 2011.

[34] C. Liu, B. C. Wheeler, W. D. O'Brien, Jr, R. C. Bilger, C. R. Lansing, and A. S. Feng, "Localization of multiple sound sources with two microphones," *J. Acoust. Soc. Amer.*, vol. 108, no. 4, pp. 1888–1905, 2000.

[35] J. Woodruff and D. Wang, "Binaural detection, localization, and segregation in reverberant environments based on joint pitch and azimuth cues," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 21, no. 4, pp. 806–815, Apr. 2013.

[36] A. Griffin, A. Alexandridis, D. Pavlidi, Y. Mastorakis, and A. Mouchtaris, "Localizing multiple audio sources in a wireless acoustic sensor network," *Signal Process.*, vol. 107, pp. 54–67, 2015.

[37] A. Griffin, A. Alexandridis, D. Pavlidi, and A. Mouchtaris, "Real-time localization of multiple audio sources in a wireless acoustic sensor network," in *Proc. Eur. Signal Process. Conf.*, 2014, pp. 306–310.

[38] H. Buchner, R. Aichner, and W. Kellermann, "TRINICON: A versatile framework for multichannel blind signal processing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2004, vol. 3, doi: 10.1109/ICASSP.2004.1326688.

[39] M. Swartling, B. Sällberg, and N. Grbić, "Source localization for multiple speech sources using low complexity non-parametric source separation and clustering," *Signal Process.*, vol. 91, no. 8, pp. 1781–1788, 2011.

[40] C. Evers and P. A. Naylor, "Acoustic slam," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 9, pp. 1484–1498, Sep. 2018.

[41] D. Marković, F. Antonacci, A. Sarti, and S. Tubaro, "Multiview soundfield imaging in the projective ray space," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 6, pp. 1054–1067, Jun. 2015.

[42] A. Alexandridis and A. Mouchtaris, "Multiple sound source location estimation in wireless acoustic sensor networks using DOA estimates: The data-association problem," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 26, no. 2, pp. 342–356, Feb. 2018.

[43] X. Dang, Q. Cheng, and H. Zhu, "Indoor multiple sound source localization via multi-dimensional assignment data association," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 1944–1956, Dec. 2019.

[44] B.-N. Vo, S. Singh, and W. K. Ma, "Tracking multiple speakers using random sets," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.,*, vol. 2, May 2004, pp. 357–360.

[45] B.-N. Vo, W.-K. Ma, and S. Singh, "Localizing an unknown time-varying number of speakers: A bayesian random finite set approach," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, Mar. 18–23, 2005, pp. 1073–1076.

[46] W.-K. Ma, B.-N. Vo, S. S. Singh, and A. Baddeley, "Tracking an unknown time-varying number of speakers using TDOA measurements: A random finite set approach," *IEEE Trans. Signal Process.*, vol. 54, no. 9, pp. 3291–3304, Sep. 2006.

[47] N. T. Pham, W. Huang, and S. H. Ong, "Tracking multiple speakers using cphd filter," in *Proc. 15th Int. Conf. Multimedia*, pp. 529–532, 2007.

[48] N. Roman and D. Wang, "Binaural tracking of multiple moving sources," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 728–739, May 2008.

[49] D. Ayllón, R. Gil-Pita, M. Rosa-Zurera, and H. Krim, "Real-time multiple doa estimation of speech sources in wireless acoustic sensor networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 2709–2713.

[50] A. Masnadi-Shirazi and B. Rao, "An ICA-SCT-PHD filter appraoch for tracking and separation of unknown time-varying number of sources," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 21, no. 4, pp. 828–841, Apr. 2013.

[51] J. Grimmer, "An introduction to Bayesian inference via variational approximations," *Political Anal.*, vol. 19, no. 1, pp. 32–47, 2011.

[52] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *J. Amer. Stat. Assoc.*, vol. 112, no. 518, pp. 859–877, 2017.

[53] C. Wang and D. M. Blei, "Variational inference in nonconjugate models," *J. Mach. Learn. Res.*, vol. 14, Apr. pp. 1005–1031, 2013.

Jielong Yang received the B.Eng. and M.Sc. degrees from Xi'an Jiaotong University, Xi'an, China, in 2012 and 2014, respectively, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2020. He is currently an Assistant Professor with the School of Artificial Intelligence, Jilin University. His research interests include statistical machine learning and signal processing over networks.

Xionghu Zhong received the B.Eng. and M.Sc. degrees from Northwestern Polytechnical University, China, in 2003 and 2006, respectively, and the Ph.D. degree from the Institute for Digital Communications, The University of Edinburgh, U.K., in 2010. He was a Research Fellow with the School of Computer Engineering and a Senior Research Fellow with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. He was with Xylem Inc., as a Data Scientist from 2017 to 2018. He is currently a Professor with the College of Computer Science and Electronic Engineering, Hunan University, China. His research interests include statistical signal processing, target localization and tracking, and machine learning methods, and their applications to distant speech enhancement and recognition, V2X communications, and water distribution network monitoring.

Weiguang Chen received the B.Eng. degree in 2018 from Hunan University, Hunan, China, where he is currently working toward the Ph.D degree. His research interests include nonparametric Bayesian modeling and statistical signal processing.

Wenwu Wang received the B.Sc., M.E., and Ph.D. degrees from the College of Automation, Harbin Engineering University, Harbin, China, in 1997, 2000, and 2002, respectively. He then worked in King's College London (2002–2003), Cardiff University (2004–2005), Tao Group Ltd. (now Antix Labs Ltd.) (2005–2006), Creative Labs (2006–2007), before joining the University of Surrey in May 2007, where he is currently a Professor in Signal Processing and Machine Learning, and a Co-Director of the Machine Audition Lab within the Centre for Vision Speech and Signal Processing. He is also a Guest Professor with Qingdao University of Science and Technology, China. He was a Visiting Scholar with Ohio State University, USA, in 2008. He has (co)-authored over 250 publications in these areas. His current research interests include blind signal processing, sparse signal processing, audio-visual signal processing, machine learning and perception, artificial intelligence, machine audition (listening), and statistical anomaly detection. He and his team have won the Judge's Award on DCASE 2020, Reproducible System Award on DCASE 2019 and DCASE 2020, Best Student Paper Award on LVA/ICA 2018, the Best Oral Presentation on FSDM 2016, the Top-Quality Paper Award in IEEE ICME 2015, Best Student Paper Award finalists on ICASSP 2019 and LVA/ICA 2010. He and his team have achieved the first place in 2020 DCASE challenge on "Urban Sound Tagging with Spatio-Temporal Context", and the first place in the 2017 DCASE Challenge on "Large-scale Weakly Supervised Sound Event Detection for Smart Cars", the TVB Europe Award for Best Achievement in Sound in 2016, the finalist for GooglePlay Best VR Experience in 2017, and the Best Solution Award on the Dstl Challenge "Under-sampled Signal Recognition" in 2012. He has been a Senior Area Editor (2019-) and an Associate Editor (2014–2018) for the IEEE TRANSACTIONS ON SIGNAL PROCESSING. He is an Associate Editor (2020-) IEEE/ACM TRANSACTIONS ON AUDIO SPEECH AND LANGUAGE PROCESSING and an Associate Editor (2019-) for EURASIP Journal on Audio Speech and Music Processing. He was a Publication Co-Chair for ICASSP 2019, Brighton, UK. He is a member of the International Steering Committee of Latent Variable Analysis and Signal Separation (2019).