Semi-Supervised Multiple Source Localization Using Relative Harmonic Coefficients Under Noisy and Reverberant Environments

Yonggang Hu, Prasanga N. Samarasinghe, *Member, IEEE,*
Sharon Gannot, *Senior Member, IEEE,* Thushara D. Abhayapala, *Senior Member, IEEE*

***Abstract*—This paper develops a semi-supervised algorithm to address the challenging multi-source localization problem in a noisy and reverberant environment, using a spherical harmonics domain source feature of the *relative harmonic coefficients*. We present a comprehensive research of this source feature, including (i) an illustration confirming its sole dependence on the source position, (ii) a feature estimator in the presence of noise, (iii) a feature selector exploiting its inherent directivity over space. Source features at varied spherical harmonic modes, representing unique characterization of the soundfield, are fused by the Multi-Mode Gaussian Process modeling. Based on the unifying model, we then formulate the mapping function revealing the underlying relationship between the source feature(s) and position(s) using a Bayesian inference approach. Another issue of the overlapped components is addressed by a pre-processing technique performing overlapped frame detection, which in turn reduces this challenging problem to a single source localization. It is highlighted that this data-driven method has a strong potential to be implemented in practice because only a limited number of labeled measurements is required. We evaluate this proposed algorithm using simulated recordings between multiple speakers in diverse environments, and extensive results confirm improved performance in comparison with the state-of-art methods. Additional assessments using real-life recordings further prove the effectiveness of the method, even at unfavorable circumstances with severe source overlapping.**

***Index Terms*—Semi-supervised multiple source localization, relative harmonic coefficients, source feature estimator, Gaussian Process regression, Multi-Mode Gaussian Process.**

I. Introduction

KNOWLEDGE of the positions of sound sources within a given area is a fundamental requirement by many spatial acoustic techniques and applications, including teleconferencing systems, source dereverberation [1], speech separation [2], automatic speech recognition [3] and automated camera steering [4]. As a consequence, the topic of sound source localization and tracking [5] has attracted a significant attention in the research community as well as in the industry.

A. Literature Review

Most early localization methods, such as the generalized cross-correlation phase transform (GCC-PHAT) [6], require the time difference of arrival (TDOA) between microphone pairs [7]. Another type of solutions to source localization is the steered response power (SRP) [8] and SRP-phase transform (SRP-PHAT) [9] based algorithms, which explore all possible directions over the two-dimensional space to search the areas with higher response power. Aforementioned methods provide accurate direction of arrival (DOA) estimates in scenes where only a single sound source is active. The task becomes more challenging when multiple sound sources are present in the environment [10]. Subspace methods, utilizing the spatial covariance/correlation matrix of the recordings, are more suitable under such circumstances. Typical subspace methods comprise the adaptive eigenvalue decomposition [11], multiple signal classification (MUSIC) [12], [13] and estimation of signal parameters via rotational invariance (ESPRIT) [14], [15]. However, their localization accuracy degrades severely in a complex acoustic environment where the original recordings are contaminated by the multi-path reverberation resulting from strong reflections from objects in the enclosure as well as the noise with low signal-to-noise ratios.

In recent years, data-driven source localization algorithms have been widely investigated to address the degraded accuracy in the complex environments. In [16], a single-source DOA estimation was realized using a multi-layer neural network, which takes the generalized cross-correlation as the inputs. A deep neural network (DNN) based phase difference enhancement for multi-source DOA estimation was presented in [17]. Another approach in [10] used a convolutional recurrent neural network (CRNN) to estimate the DOAs of multiple sound sources using a first-order Ambisonics source feature. Chakrabarty et al. used the phase component of the short-time Fourier transform (STFT) coefficients of the received microphone signals as the input feature to a convolutional neural network (CNN) for supervised multi-speaker DOA estimations [18]. More recently, Fahim et al. also achieved multi-source DOA using a CNN to learn the modal coherence patterns of an incident soundfield through the measured spherical harmonic coefficients [19]. Aforementioned approaches use different types of source features, which are vital to the algorithm's accuracy as they contain relevant characteristics of the sound source to be localized. Intuitively, a source feature suitable

Yonggang Hu, P. N. Samarasinghe and T. D. Abhayapala are with the Research School of Engineering, Australian National University, Canberra, A.C.T. 2601, Australia. (e-mail: yonggang.hu@anu.edu.au; prasanga.samarasinghe@anu.edu.au; thushara.abhayapala@anu.edu.au; *Corresponding author: Yonggang Hu.*).

Sharon Gannot is with the Faculty of Engineering, Bar-Ilan University, Ramat-Gan 5290002, Israel (e-mail: sharon.gannot@biu.ac.il).

This project has received funding from ARC grant No. DP180102375 and European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245. Yonggang Hu is sponsored by CSC agency for funding.

for source localization shall better convey/represent the source position, and be less dependent on the time-varying source signal.

In recent decade, the relative transfer function (RTF) [20]–[22] has been proved to be another promising source feature for source localization as it serves as an acoustic fingerprint related to the source's location. Early use of RTF for source localization in [7] was to extract the TDOA of source signal in the first stage, which was then used for single source localization in the second stage. An investigation in [23] reveals that the RTF is intrinsically embedded in a low-dimensional manifold which is solely governed by its source position. Assuming a static reverberant environment, the source position is the only varying degree-of-freedom of the set of RTFs in the enclosure, thus it is capable of recovering the unknown source position. Using pairs of microphones, Laufer-Goldshtein et al. have exploited the RTF for both semi-supervised single source localization [24]–[27] and source tracking [28], respectively. With a binaural microphone setup, Li et al. achieved a supervised multiple source localization using the direct-path RTF where only a single source is active [29]. Thereafter, an online scheme using the RTF to track multiple moving speakers in a reverberant environment was presented in [30]. More recently, Brendel et al. exploited the RTF to propose an expectation-maximization (EM) based algorithm, achieving a joint speaker number counting and localization in adverse acoustic conditions [31]. Opochinsky et al. then fed the RTF into a deep-learning network for a weakly-supervised ranking-based source localization [32]. Motivated by the wide applications of RTF relating two individual microphones, a preliminary research in [33] studied a source feature called as *relative harmonic coefficients* in the spherical harmonics domain [34] (i.e., modal domain in [35]). This feature is generally more appropriate for higher-order microphone arrays, such as spherical and circular microphone arrays. The microphone arrays are capable of recording and analyzing the soundfield over a large spatial area, thus have been widely used in recently proposed localization methods [36]–[41].

B. Contribution by This Paper

This paper aims to address the multiple source localization in a noisy and reverberant environment by proposing a data-driven approach using the *relative harmonic coefficients*. Aforementioned deep learning based algorithms, such as [17]–[19], [32], accomplish source localization by classifying the desired source DOA into one of the candidate directions over the two dimensional space. Our proposed approach adopts a regression scheme, i.e., a Bayesian inference approach of Gaussian Process Regression (GPR) [42], because it suits more to localize the continuous variable of the source positions (i.e., x, y, z coordinates). Traditional GPR requires a single Gaussian Process modeling, while this paper adapts the Multi Gaussian Process modeling [27] to the spherical harmonics domain (called as Multi-Mode Gaussian Process (MMGP)), in order to fuse the relative harmonic coefficients over the varied spherical harmonic modes. Data-driven source localization is often criticized as a cumbersome task because it requires a large training set. To overcome the drawback, we are adopting the semi-supervised paradigm, previously used in [25]–[27], where only a small number of labeled samples is required. However, [25]–[27] only addressed the single-source scenario. Multiple source localization becomes much more challenging because the overlapped components, especially significantly overlapped recordings, hinder an accurate localization of the original sources. Recent studies [29], [31] addressed this issue using a pre-processing tool to detect and isolate the overlapped components. Motivated by this strategy, our paper simplifies the challenging multi-source localization into a single source localization problem by developing a new overlapped frame detector using the relative harmonic coefficients.

Some preliminary research in [33], [43] investigated sound source localization using the relative harmonic coefficients. However, [33] only addressed a single sound source and [43] addressed the multi-source localization while its accuracy degraded severely in noisy and reverberant environments. In comparison with [33], [43], additional contributions by this paper are briefly summarized as follows: (i) we study a semi-supervised multi-source localization approach, only using a small number of labeled training samples, (ii) we present a theoretical proof confirming the defined source feature only depends on its source position, (iii) we propose a new source feature estimator under noisy conditions, (iv) we develop a metric selecting the spherical harmonic modes that suits for source localization in a given area, (v) we provide a data-driven overlapped frame detection, (vi) we add more in-depth evaluations and analysis. The remaining part of the paper is structured as follows. We first formulate the problem addressed by this paper and then introduce the relative harmonic coefficients in Section II. Section III presents the source feature selector exploiting its inherent directivity. Section IV derives the mapping function that fuses the selected source features. Section V summarizes the block-diagram of the algorithm and explains the data-driven overlapped frame detection. Thereafter, extensive experimental results are reported in Section VI. Finally, conclusions are drawn and future directions are discussed in Section VII.

II. System Model

This section first briefly describes the problem to be addressed by this paper. Then, we introduce the spherical harmonics domain source feature called relative harmonic coefficients from several aspects, which will be used by the proposed source localization approach.

A. Problem Formulation

Let there be Q active sound sources inside the reverberant room (e.g., see Figure 1), whose Cartesian coordinates are $\boldsymbol{p}_q = [x_q, y_q, z_q]^T$ $(q = 1, \cdots, Q)$ with respect to the room origin of $O = [0, 0, 0]^T$. Consider a higher-order microphone array with M microphones that are located at $\boldsymbol{x}_j$ $(j = 1, \cdots, M)$ with respect to the array origin O_r. The sound pressure, measured by the j-th microphone of the array at the

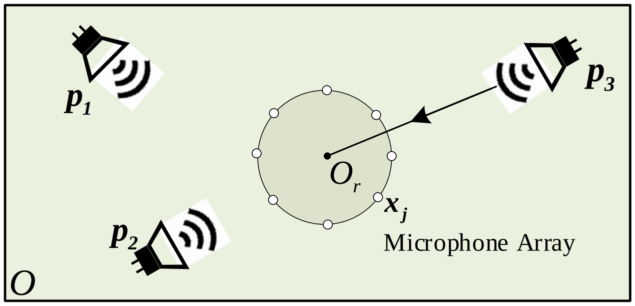

Fig. 1. Multiple source localization using a higher-order microphone array in a noisy and reverberant environment (top view).

k-th frequency bin, is represented by:

$$\begin{aligned}\bar{P}(\boldsymbol{x}_j,k) &= P(\boldsymbol{x}_j,k)+V(\boldsymbol{x}_j,k), \quad j=1,\cdots,M\\ &= \sum_{q=1}^{Q} S_q(k)A_q(\boldsymbol{x}_j,k)+V(\boldsymbol{x}_j,k)\end{aligned} \tag{1}$$

where $k=2\pi f/c$ is the wavenumber, f is the frequency bin, c is the speed of sound, $S_q(k)$ is the q-th source signal, $A_q(\boldsymbol{x}_j,k)$ denotes the acoustic transfer function (ATF) from the q-th sound source to the j-th microphone, $P(\boldsymbol{x}_j,k)$ and $\bar{P}(\boldsymbol{x}_j,k)$ denote the clean and noisy sound pressure and $V(\boldsymbol{x}_j,k)$ represents the additive noise signal at the j-th microphone. Given the multi-source recordings of $\bar{P}(\boldsymbol{x}_j,k)$, this paper aims to accurately recover the positions of the sound sources, i.e., $\boldsymbol{p}_q$ where $q=1,\cdots,Q$. In addition, we have $\mathcal{N}_D=\mathcal{N}_L+\mathcal{N}_U$ measurements in advance within a predefined source area of interest, consisting of $\mathcal{N}_L$ labeled samples whose known positions are $\mathbf{p}=\{\boldsymbol{p}_1,\cdots,\boldsymbol{p}_{\mathcal{N}_L}\}$, and $\mathcal{N}_U$ unlabeled samples randomly located at unknown positions. Note that the additive noise in (1) is assumed to be non-directional, otherwise, the directional noise could be treated as additional sources to be localized.

B. Relative Harmonic Coefficients (RHC)

The sound pressure at an arbitrary point microphone $\boldsymbol{x}_j=(r,\theta_j,\phi_j)$, $j=1,\cdots,M$ within the recording area can be represented in the spherical harmonics domain [44],

$$P(\boldsymbol{x}_j,k)=\sum_{n=0}^{N}\sum_{m=-n}^{n}\alpha_{nm}(k)\,b_n(kr)\,Y_{nm}(\theta_j,\phi_j) \tag{2}$$

where $n(\geqslant 0)$ and m are integers, $N=\lceil kr\rceil$ is the truncated order of the soundfield [34],

$$Y_{nm}(\theta,\phi)=\sqrt{\frac{(2n+1)}{4\pi}\frac{(n-m)!}{(n+m)!}}P_{nm}(\cos\theta)e^{im\phi} \tag{3}$$

is the spherical harmonic function, $P_{nm}(\cdot)$ represents the associated Legendre function, and $b_n(\cdot)$ is the function based on the configuration of the spherical microphone array,

$$b_n(kr)=\begin{cases} j_n(kr), & \text{for an open array}\\ j_n(kr)-\frac{j_n^{'}(kR)}{h_n^{'}(kR)}h_n(kr), & \text{for a rigid array}\end{cases} \tag{4}$$

where R denotes the radius of the spherical microphone array, $j_n^{'}(\cdot)$ and $h_n^{'}(\cdot)$ denote the partial derivatives of spherical Bessel and Hankel functions, respectively.

The $\alpha_{nm}(k)$ in (2) denotes the spherical harmonic coefficients which characterize/describe the measured soundfield in the spherical harmonics domain. Assume the soundfield decomposed by (2) originates from a single sound source. Preliminary research in [33], [43], [45]–[47] define the relative harmonic coefficients (RHC) of order n and mode m, as the ratio between $\alpha_{nm}(k)$ and $\alpha_{00}(k)$,

$$\beta_{nm}(k)=\frac{\alpha_{nm}(k)}{\alpha_{00}(k)}. \tag{5}$$

Let the frequency band of interest be $[k_{\text{min}},k_{\text{max}}]$. Then, we propose a $F\times 1$ feature vector for each (n,m) mode,

$$\boldsymbol{\beta}_{nm}=\left[\beta_{nm}(k_1),\beta_{nm}(k_2),\cdots,\beta_{nm}(k_F)\right]^T \tag{6}$$

where $k_{\text{min}}\leqslant k_1,\cdots,k_F\leqslant k_{\text{max}}$. We combine feature vectors of all the spherical harmonic modes to obtain $F\times(N+1)^2$ matrix of relative harmonic coefficients as,

$$\boldsymbol{B}=\left[\boldsymbol{\beta}_{00},\boldsymbol{\beta}_{1,-1},\cdots,\boldsymbol{\beta}_{NN}\right] \tag{7}$$

where the $F\times 1$ feature vector $\boldsymbol{\beta}_{00}=[1,\cdots,1]^T$ because the $\beta_{00}(k)=1$ by the definition in (5). Note that, we mainly use abbreviation of relative harmonic coefficients i.e., RHC, when referring to the defined source feature in the following.

C. Illustration of the Source Feature

This subsection illustrates the composition of relative harmonic coefficients by deriving its theoretical expression in both free and reverberant environments, which confirm to be only dependent on its source position.

1) Free-field: Assume the q-th sound source, located at $\boldsymbol{p}_q=[x_q,y_q,z_q]^T$ with respect to the room origin O in Figure 1, has the polar coordinate of (r_q,θ_q,ϕ_q) with respect to origin of the microphone array O_r. Its spherical harmonic coefficients due to the incoming direct-path recordings are given by [35],

$$\alpha_{nm}^{\text{dir}}(k)=S_q(k)ikh_n(kr_q)Y_{nm}^{*}(\theta_q,\phi_q) \tag{8}$$

where $h_n(\cdot)$ is the spherical Hankel function of the first kind and $*$ denotes the conjugate operator. Following the definition in (5), we derive its RHC of order n and mode m:

$$\beta_{nm}^{\text{dir}}(k)=\frac{2\sqrt{\pi}h_n(kr_q)Y_{nm}^{*}(\theta_q,\phi_q)}{h_0(kr_q)} \tag{9}$$

which only depends on the source position (r_q,θ_q,ϕ_q).

2) Reverberant-field: Assuming the case of a reverberant soundfield produced by the q-th sound source, its spherical harmonic coefficient over the recording area is represented as,

$$\alpha_{nm}^{\text{rev}}(k)=\alpha_{nm}^{\text{dir}}(k)+\underbrace{\sum_{v=0}^{N}\sum_{u=-v}^{v}\hat{\alpha}_{nm}^{vu}(k)S_q(k)ikj_v(kr_q)Y_{vu}^{*}(\theta_q,\phi_q)}_{\text{Reverberant-path}} \tag{10}$$

where $\hat{\alpha}_{nm}^{vu}(k)$ is the coupling coefficients that is independent of the time-varying source signal [48]. Note that (10) considers an arbitrary acoustic environment so that the coupling coefficients have no explicit expression. Following the definition in

(5), we have the corresponding RHC,

$$\beta^{\text{rev}}_{nm}(k)=\frac{h_n(kr_q)Y^*_{nm}(\theta_q,\phi_q)+\sum\limits_{v=0}^{N}\sum\limits_{u=-v}^{v}\hat{\alpha}^{vu}_{nm}(k)j_v(kr_q)Y^*_{vu}(\theta_q,\phi_q)}{h_0(kr_q)Y^*_{00}(\theta_q,\phi_q)+\sum\limits_{v=0}^{N}\sum\limits_{u=-v}^{v}\hat{\alpha}^{vu}_{00}(k)j_v(kr_q)Y^*_{vu}(\theta_q,\phi_q)} \tag{11}$$

which also only depends on the source position in a static acoustic environment where the settings of the environment and microphone array are assumed to remain fixed/unchanged.

D. Biased Source Feature Estimator

This subsection proposes a method to estimate the source feature in the presence of noise. We focus on the estimation at a single frequency bin as estimations over a wide frequency band follow a similar process. It is of common technique to measure the spherical harmonic coefficients for a spherical microphone array [34] (also measurable using multiple circular microphone arrays [49], and a planar microphone array [50]),

$$\bar{\alpha}_{nm}(k)=\frac{1}{b_n(kr)}\sum_{j=1}^{M}a_j\bar{P}(\boldsymbol{x}_j,k)\,Y^*_{nm}(\theta_j,\phi_j) \tag{12}$$

in which a_j works as the weight of each microphone (known in advance) to ensure the error between the measured and theoretical estimations is as small as possible. The measured $\bar{\alpha}_{nm}(k)$ contains non-negligible noise components,

$$\bar{\alpha}_{nm}(k)=\alpha_{nm}(k)+\gamma_{nm}(k) \tag{13}$$

where $\alpha_{nm}(k)$ and $\gamma_{nm}(k)$ denote the spherical harmonic coefficients of the source and noise signal, respectively. Assuming the recordings originate from a single sound source (e.g., the q-th source in Figure 1), we can rewrite $\bar{\alpha}_{nm}(k)$ as,

$$\bar{\alpha}_{nm}(k)=\beta_{nm}(k)\alpha_{00}(k)+\gamma_{nm}(k) \tag{14}$$

where $\beta_{nm}(k)$ denotes the defined relative harmonic coefficients that relate to the $\alpha_{00}(k)$. However, we only have the noisy $\bar{\alpha}_{00}(k)$ in practice,

$$\bar{\alpha}_{00}(k)=\alpha_{00}(k)+\gamma_{00}(k). \tag{15}$$

Note that $\beta_{nm}(k)$ is independent of the source signal, thus it is constant over the time-varying signal. Inspired by the estimator of aforementioned RTF based source feature [27], we exploit the power spectral density (PSD) and cross PSD (CPSD) of the measured signals to alleviate the negative effects caused by the noise when calculating the $\beta_{nm}(k)$,

$$\frac{S_{\bar{\alpha}_{nm}\bar{\alpha}_{00}}(k)}{S_{\bar{\alpha}_{00}\bar{\alpha}_{00}}(k)-S_{\gamma_{00}\gamma_{00}}(k)}=\frac{\beta_{nm}(k)S_{\alpha_{00}\alpha_{00}}(k)}{S_{\alpha_{00}\alpha_{00}}(k)}=\beta_{nm}(k) \tag{16}$$

where

$$\begin{aligned}S_{\bar{\alpha}_{nm}\bar{\alpha}_{00}}(k)&=\mathbb{E}\left\{\bar{\alpha}_{nm}(k)\bar{\alpha}^*_{00}(k)\right\}\\ S_{\bar{\alpha}_{00}\bar{\alpha}_{00}}(k)&=\mathbb{E}\left\{\bar{\alpha}_{00}(k)\bar{\alpha}^*_{00}(k)\right\}\\ S_{\gamma_{00}\gamma_{00}}(k)&=\mathbb{E}\left\{\gamma_{00}(k)\gamma^*_{00}(k)\right\}\\ S_{\alpha_{00}\alpha_{00}}(k)&=\mathbb{E}\left\{\alpha_{00}(k)\alpha^*_{00}(k)\right\}\end{aligned} \tag{17}$$

where $\mathbb{E}\left\{\cdot\right\}$ denotes the statistical expectation over the time-varying signal. Note that (16) exploits the fact that the spherical harmonic coefficients of source signal and noise signal are uncorrelated because their corresponding sound pressure are assumed to be uncorrelated. However, the noise PSD of $S_{\gamma_{00}\gamma_{00}}(k)$ at the denominator of (16) is still unknown. Some state-of-art power spectral density techniques [51] are available to update the $S_{\gamma_{00}\gamma_{00}}(k)$. For simplicity, we adopt a biased feature estimator by neglecting the noise PSD in (16). Thus, the source feature is estimated using

$$\beta_{nm}(k)\approx\frac{S_{\bar{\alpha}_{nm}\bar{\alpha}_{00}}(k)}{S_{\bar{\alpha}_{00}\bar{\alpha}_{00}}(k)}. \tag{18}$$

III. Source Feature Selector

This section first shows that the proposed spherical harmonic domain feature has a unique directivity pattern over space. Thus, we then develop a quantitative metric to select a subset of the spherical harmonic modes that are suitable for source localization within a limited-size source area of interest.

A. Directivity Pattern Analysis

The studied RHC based source feature has a unique directivity pattern over the three dimensional space because of its direct relation with the spherical harmonic function (e.g., see the theoretical expressions in (9) and (11)). Figure 2 exhibits some examples at the spherical harmonic modes of $(1,-1)$ and $(2,-1)$, respectively, each representing a distinct characterization/description of the soundfield. The RHCs over the remained modes are not exhibited due to the space limit. A unique directivity pattern generally assists in distinguishing the sound sources located at the area (i.e., active area) where the source features have a large difference. By contrast, the source features at any given harmonic mode also have an inactive area where they have little differences (i.e., the directivity pattern is not significant within some areas). The following are some examples: (i) $(n,m)=(0,0)$: the source features equal to 1 wherever the source locates. (ii) $(n,m)=(1,-1)$: the source features are close to zero for the sources located around the plane where $y=0$ (Figure 2 (a) and (c)). (iii) $(n,m)=(2,-1)$: when the sources are located on the horizontal plane where $z=0$, their features are inactive (Figure 2 (b) and (d)).

In practice, a data-driven source localization often implements within a limited region predefined in advance. Given the estimated source features up to the N-th order, we expect to select a subset of the spherical harmonic modes whose active area covers the given source area. As explained in the next subsection, this paper achieves the spherical harmonic modes selection by proposing a statistical metric based on the training feature set over this area.

B. Spherical Harmonic Modes Selector Using the Training Set

Assume the coordinates of the predefined sound source area for localization are, e.g.,

$$\boldsymbol{\Phi}=\left\{(x,y,z):x_1<x<x_2,y_1<y<y_2,z_1<z<z_2\right\} \tag{19}$$

in which x_1,x_2,y_1,y_2,z_1,z_2 are some constants. Consider $\mathcal{N}_D$ training samples distributed within $\boldsymbol{\Phi}$ have been measured

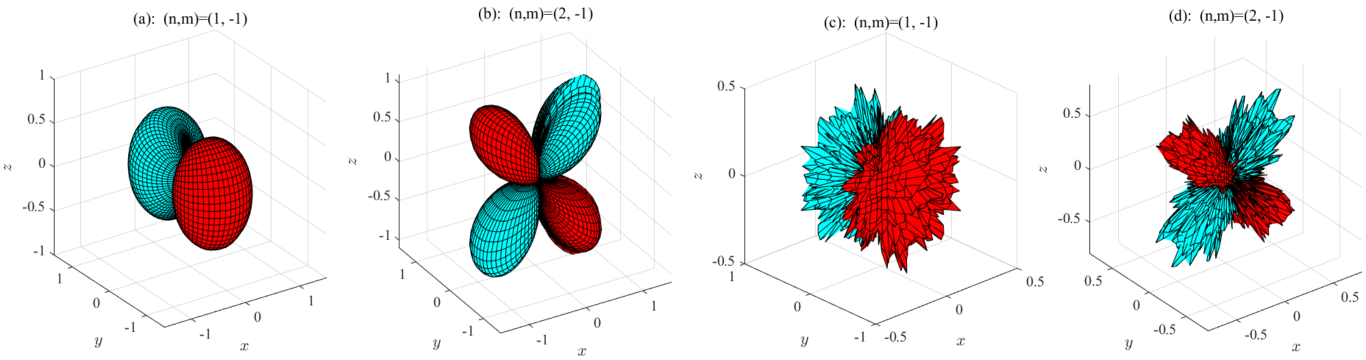

Fig. 2. Real part of the source features at the spherical harmonic modes of $(1, -1)$ and $(2, -1)$, respectively. The (a)-(b) denote the source features using direct-path recordings without any room reverberations. By contrast, (c)-(d) denote the reverberant features whose $T_{60} = 500$ ms with a room reflection order of ten. The red and cyan portions represent regions where the values of the features are positive and negative, respectively. The distance of the surface from the origin indicates the absolute value of the features in angular direction over space. We generate the source features using the estimator given the simulated recordings in a $6 \times 4 \times 3$ m room in the presence of noise at 25 dB. We simulate a set of sound sources located on a spherical shell with respect to a spherical microphone array at the origin of the shell, i.e., $\mathbf{\Phi}^2 = \{(r, \vartheta_s, \varphi_s) : r = 1, 0 < \vartheta_s \leq \pi, 0 < \varphi_s \leq 2\pi\}$. Twenty frequency bins approximately ranging from 1500 Hz to 2500 Hz are used, which records the soundfield up to the 2nd order. Note that the presented figures denote the mean values over this wide frequency band. We notice the source features in a reverberant environment appear to be less smoothly distributed over space, caused by the random interfering signals of the room reverberation.

and the corresponding training source features are estimated. We then construct a vector by collecting the relative harmonics coefficients at the mode of (n, m) for all the samples,

$$\left[\beta_{nm}^{1}(k), \beta_{nm}^{2}(k), \cdots, \beta_{nm}^{\mathcal{N}_D}(k)\right]^T. \tag{20}$$

As analyzed, the sound sources within the active area appear with more different values, i.e., the source features distribute more decentralized. For a quantitative measurement, we exploit the index of dispersion (i.e., Variance to Mean Ratio) [52] with respect to the vector of (20),

$$\lambda_{nm}(k) = \left|\frac{\sigma_{nm}^2(k)}{\mu_{nm}(k)}\right| \tag{21}$$

where

$$\begin{aligned} \mu_{nm}(k) &= \frac{1}{\mathcal{N}_D} \sum_{n_\ell=1}^{\mathcal{N}_D} \beta_{nm}^{n_\ell}(k) \\ \sigma_{nm}^2(k) &= \frac{1}{\mathcal{N}_D - 1} \sum_{n_\ell=1}^{\mathcal{N}_D} |\beta_{nm}^{n_\ell}(k) - \mu_{nm}(k)|^2 \end{aligned} \tag{22}$$

denote the mean and variance of the elements in (20) (note that $\mu_{nm}(k) \neq 0$), respectively, and $1 \leq n_\ell \leq \mathcal{N}_D$ denotes the index number. Note that above calculation only uses the source feature at the k-th frequency bin. In the case of a wide frequency band (e.g., F frequency bins), we then compute the mean number as,

$$\bar{\lambda}_{nm} = \frac{1}{F} \sum_{i=1}^{F} \lambda_{nm}(k_i). \tag{23}$$

The measure of dispersion is successively applied for all the $(N+1)^2$ spherical harmonic modes to produce a vector as,

$$\left[\bar{\lambda}_{00}, \bar{\lambda}_{1,-1}, \cdots, \bar{\lambda}_{NN}\right]^T. \tag{24}$$

Intuitively, we select the spherical harmonic mode exhibiting a larger index of dispersion, i.e., the source features have larger differences when the sources are located differently,

$$\bar{\lambda}_{nm} > \zeta \tag{25}$$

where ζ is a positive threshold empirically specified as long as it performs with sufficient localization accuracy. For example, $\bar{\lambda}_{00} = 0$, thus source features at this spherical harmonic mode $(n, m) = (0, 0)$ are discarded.

Up to now, we have estimated the training source features and selected a subset of the spherical harmonic modes that well suit to localize the sources within the given area. As explained in the next section, we show how to use the training features to formulate a mapping function that recovers the testing source's unknown position.

IV. Mapping Function Formulation

This section aims to formulate the mapping function revealing the underlying relation between the source feature(s) and source position(s). We first use the Multi-Mode Gaussian Process (MMGP) to model the variable of source position, fusing/merging the features at the selected spherical harmonic modes. Then, we use the MMGP based Gaussian Process Regression (GPR) to recover the unknown source position. Note that the proposed GPR based source localization approach localizes the source x, y, z-coordinate separately because the Gaussian Process modeling mainly applies into scalar variable [42]. Hence, the source position variable p used in this section denotes a scalar of p_x, p_y or p_z. Finally, we claim in advance that the underlying theory discussed in this section is a direct inspiration and adaptation of a recently proposed method in [27]. The original method exploits the RTFs for the mapping function formulation while this section uses the RHCs defined in the spherical harmonic domain.

A. Multi-Mode Gaussian Process (MMGP)

Assume an arbitrary sound source whose feature matrix is $\boldsymbol{B} \in \mathbb{C}^{F \times V}$ where $V \leq (N+1)^2$ denotes the number of the selected spherical harmonic modes. Using a single feature vector at the v-th mode, we model the variable of its source position by a zero mean Gaussian Process,

$$p^{(v)}(\boldsymbol{\beta}) \sim \mathcal{GP}(0, \mathcal{K}) \tag{26}$$

where $p^{(v)}$ denotes the source position variable at the v-th mode, $\boldsymbol{\beta} \in \mathbb{C}^{F\times 1}$ denotes the feature vector containing all the F frequency bins at this mode, $\mathcal{K}$ denotes the kernel or covariance function that specifies the Gaussian Process. We adopt the manifold-based covariance function [27], where the relation between two sources is not only a function of the current two samples, but also exploits the information of the entire training set,

$$\text{cov}(p^{(v)}_{n_i}, p^{(v)}_{n_j}) \equiv \sum_{n_\ell=1}^{\mathcal{N}_D} \mathcal{K}(\boldsymbol{\beta}_{n_i}, \boldsymbol{\beta}_{n_\ell})\mathcal{K}(\boldsymbol{\beta}_{n_j}, \boldsymbol{\beta}_{n_\ell}) \tag{27}$$

where subscript of n_i and n_j denotes the index of two arbitrary sources, n_ℓ is the index of the training sources and $\mathcal{K}(\cdot)$ is the kernel function between any pair of features. Theoretically, a series of kernel functions is applicable as long as its covariance matrix is positive semi-definite and symmetric [42]. We use the squared exponential (SE) covariance function as,

$$\mathcal{K}(\boldsymbol{\beta}_{n_i}, \boldsymbol{\beta}_{n_\ell}) = \exp\left(-\frac{\|\boldsymbol{\beta}_{n_i} - \boldsymbol{\beta}_{n_\ell}\|^2}{2\sigma_y^2}\right), 1 \leqslant n_i, n_\ell \leqslant \mathcal{N}_D \tag{28}$$

where $\|\cdot\|$ represents the Euclidean ℓ_2 norm, and σ_y denotes the characteristic length-scale hyperparameter that is initialized with a random value and then optimized using the marginal likelihood [42].

Note that the Gaussian Process modeling above only uses a single feature vector at the v-th spherical harmonic mode. By contrast, the MMGP fuses all the source features by modeling source position p as the mean of the Gaussian Processes between all the V spherical harmonic modes, i.e., the n_i-th source final position p_{n_i} equals to the average value of all the estimations,

$$p_{n_i} = \frac{1}{V}\left(p^{(1)}_{n_i} + p^{(2)}_{n_i} + \cdots + p^{(V)}_{n_i}\right). \tag{29}$$

We emphasize the difference between the Multi-Node Gaussian Process in [27] that fused recordings from the distributed microphone pairs and our proposed method in which we fuse the features of different spherical harmonic modes given by a higher-order microphone array[1]. Due to the assumption that the processes are jointly Gaussian, p also follows a zero-mean Gaussian Process, whose covariance between two arbitrary source positions is computed as,

$$\begin{aligned}\text{cov}(p_{n_i}, p_{n_j}) &= \bar{\mathcal{K}}(\boldsymbol{B}_{n_i}, \boldsymbol{B}_{n_j})\\ &= \frac{1}{V^2}\text{cov}\left(\sum_{z=1}^{V} p^{(z)}_{n_i}, \sum_{w=1}^{V} p^{(w)}_{n_j}\right)\\ &= \frac{1}{V^2}\sum_{z,w=1}^{V}\text{cov}(p^{(z)}_{n_i}, p^{(w)}_{n_j})\end{aligned} \tag{30}$$

in which $\bar{\mathcal{K}}(\cdot)$ denotes the kernel function of the MMGP, $\boldsymbol{B}_{n_i}$ and $\boldsymbol{B}_{n_j}$ are the feature matrix containing all the V modes, and z and w are the index of spherical harmonic mode. This paper defines the covariance of variables between two different modes as,

$$\text{cov}(p^{(z)}_{n_i}, p^{(w)}_{n_j}) \equiv \text{cov}(p^{(z)}_{n_i}, p^{(z)}_{n_j})\text{cov}(p^{(w)}_{n_i}, p^{(w)}_{n_j}) \tag{31}$$

where $\text{cov}(p^{(v)}_{n_i}, p^{(v)}_{n_j})$ denotes the covariance function in (27) using all the training samples at the v-th mode where $v = z, w$. Substituting (31) into (30), the final calculations of the covariance between the variables p_{n_i} and p_{n_j} are,

$$\begin{aligned}\text{cov}(p_{n_i}, p_{n_j}) &= \bar{\mathcal{K}}(\boldsymbol{B}_{n_i}, \boldsymbol{B}_{n_j})\\ &= \frac{1}{V^2}\sum_{z,w=1}^{V}\text{cov}(p^{(z)}_{n_i}, p^{(z)}_{n_j})\text{cov}(p^{(w)}_{n_i}, p^{(w)}_{n_j})\end{aligned} \tag{32}$$

Note that above calculations of the covariance between the positional variables only use the source features (i.e., source positional information is not required). Hence, both the labeled and unlabeled training samples are exploited. In the next subsection, we show how to estimate the unknown testing source position using a GPR tool.

B. Estimate Unknown Source Position Using GPR

Based on the MMGP, localization of a single sound source, located at an unknown source position, can be reviewed as a regression problem,

$$\begin{aligned}\bar{p}_{n_\ell} &= p_{n_\ell} + \varepsilon_{n_\ell}\\ &= f(\boldsymbol{B}_{n_\ell}) + \varepsilon_{n_\ell}, \quad n_\ell = 1, \cdots, \mathcal{N}_L\end{aligned} \tag{33}$$

where n_ℓ is the index of labeled training sources, $\bar{p}_{n_\ell}$ and p_{n_ℓ} denote the measured and desired source position, respectively, $f(\boldsymbol{B}_{n_\ell})$ is the mapping function between the n_ℓ-th source feature $\boldsymbol{B}_{n_\ell}$ and source position p_{n_ℓ} and $\varepsilon_{n_\ell} \sim \mathcal{N}(0, \sigma^2)$ denotes a zero-mean Gaussian noise (i.e., the calibration inaccuracies originating from inevitable errors such as imprecise positional measurements). Given the feature matrix $\boldsymbol{B}^\star$ of a testing source, feature set of the labeled training samples, i.e., $\hat{\boldsymbol{B}} = [\boldsymbol{B}_{n_1}, \cdots, \boldsymbol{B}_{n_{\mathcal{N}_L}}]$, and their positional information $\mathbf{p} = [p_{n_1}, \cdots, p_{n_{\mathcal{N}_L}}]^T$, we recover the unknown testing source position using a standard Bayesian approach,

$$\begin{aligned}Pr(f^\star|\boldsymbol{B}^\star, \hat{\boldsymbol{B}}) &= \int Pr(f^\star, f|\boldsymbol{B}^\star, \hat{\boldsymbol{B}})df\\ &= \int Pr(f^\star|f, \boldsymbol{B}^\star, \hat{\boldsymbol{B}})Pr(f|\boldsymbol{B}^\star, \hat{\boldsymbol{B}})df\end{aligned} \tag{34}$$

in which f and $f^\star$ denote source position of $f(\hat{\boldsymbol{B}})$ and $f(\boldsymbol{B}^\star)$, respectively. For the sake of clarity, we directly present the probability distribution of $Pr(f^\star|\boldsymbol{B}^\star, \hat{\boldsymbol{B}})$, which follows a Gaussian distribution [27],

$$\mathcal{N}\left(\mathbf{K}^*(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{p}, \mathbf{K}^{**} - \mathbf{K}^*(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{K}^{*^{T}}\right) \tag{35}$$

where $\mathbf{I}$ denotes an Identity matrix, σ^2 is the noise variance in (33), $\mathbf{K}^* \in \mathbb{R}^{N_T \times N_L}$ is the covariance matrix containing the covariance of two arbitrary positional variables between

[1] Please note that the Multi-Mode Gaussian Process refers to the proposed method by this paper while the Multi-Node Gaussian Process refers to the method in [27].

the training and testing sources, $\mathbf{K} \in \mathbb{R}^{N_L \times N_L}$ and $\mathbf{K}^{**} \in \mathbb{R}^{N_T \times N_T}$ represent the covariance matrix for the training and testing sources, respectively. Note that N_T above denotes the total number of tastings based on the recordings from a single source. Hence, the unknown positions of the testing source $\mathbf{p}^* = [p_1^*, \cdots, p_{N_T}^*]^T$ is given by the mean value of the Gaussian distribution in (35) as the probability reaches its global maximum,

$$\mathbf{p}^* = \mathbf{K}^*(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{p} \tag{36}$$

which can be interpreted as linear combination of the source positions in the labeled training set, i.e., $\mathbf{p}^* = \mathbf{w}^T\mathbf{p}$ where $\mathbf{w}^T = \mathbf{K}^*(\mathbf{K} + \sigma^2\mathbf{I})^{-1}$ are the linear weights. Alternatively, the estimator of (36) can also be reviewed as a linear combination $\mathbf{p}^* = \mathbf{K}^*\mathbf{u}$, whose weights are $\mathbf{u} = (\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{p}$.

Some necessary comments are given with respect to the mapping function above:

- The mapping function is semi-supervised as it requires no positional information of the unlabeled samples. Although the unlabeled samples do not appear explicitly in (36), they play a part in the calculations of the covariance between positional variables (i.e., they appear in (27) which is used in (32)). Usage of the unlabeled samples enables a more precise measurement of the covariance for the MMGP modeling. Additionally, they exert a negligible influence on the practicality of the algorithm as we can easily obtain the unlabeled samples by randomly sampling the source area.
- There still remain some unknown parameters in the estimated position of (36), i.e., parameters of the covariance function and noise covariance. This paper uses the marginal likelihood [42] to specify those parameters. However, its non-convexity easily leads to local optimality with non-negligible errors from the global optimal results. To tackle this issue, we adopt the empirical method of cross-validation [42] to split the training set into two disjoint sets, one of which is used for training, and the other set, i.e., the validation or reference set, is used to monitor performance.
- The number of parameters in our mapping function depends on the total number of spherical harmonic modes. A larger number of parameters make it more difficult to optimize them simultaneously, as well as reducing the practicality of this approach. From this point of view, the spherical harmonic modes selector not only increases validity of the source features, but also reduces the algorithm complexity.

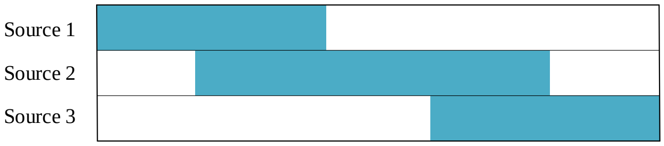

Fig. 3. An example of overlapped multi-source recordings by 3 sound sources. The cyan color denotes the periods where a sound source is active.

V. Proposed Multiple Source Localization

A. Framework of the Algorithm

Multi-source localization in this paper mainly considers the overlapped recordings as they are very common in practice, such as conversational recordings between several speakers [31]. Figure 3 exhibits the overlapped recordings with a 40% overlapped ratio (i.e., the percentage of the overlapped periods among the recording). Figure 4 presents a compact block diagram of the proposed multi-source localization algorithm, which mainly consists of two disjoint stages, i.e., a training stage and testing stage.

Training stage:

(i) Select $\mathcal{N}_L$ labeled and $\mathcal{N}_U$ unlabeled training samples within the defined reverberant sound source area of interest (e.g., Φ). (ii) Measure the recordings due to each training source separately using a higher-order microphone array and then collect the training feature set by estimating the features using the estimator given in (18). Note that, since the feature is independent of the source signal, we can use any given source signal (e.g., speech sentences or random signal) to drive the loudspeakers placed at different positions within the source area. (iii) Implement the defined metric of (25) to select a proper subset of spherical harmonic modes. (iv) Formulate the mapping function using the MMGP, optimize and specify the parameters required by the test stage.

Testing stage:

(i) Record the overlapped recordings from multiple sources (e.g., $Q > 1$ sources) within the source area of Φ, divide them into source frames in time domain (e.g., T frames in total and each lasting 0.5 s), and then obtain their source features using the feature estimator of (18). (ii) Use the overlapped frame detection, as explained in the next subsection, to detect and isolate the components overlapped by multiple sources. (iii) Only preserve the source features at the single source frames (e.g., N_T single source frames where $1 \leqslant N_T \leqslant T$), and estimate their positions using the mapping function of (36) obtained during the training stage. (iv) collect all the estimated positions of the single-source frames and use a clustering tool (e.g., K-means [53]) for the final estimates. The final estimated positions correspond to the centered location of each cluster.

B. Overlapped Frame Detection Using the Training Set

This subsection explains the last step of the algorithm in Figure 4, i.e., the overlapped frame detection that simplifies the multi-source localization into a single source localization.

Let us assume that the t-th frame originates from a single source located at $\boldsymbol{p}_q$. Due to the direct relation between source feature(s) and position(s), its feature $\boldsymbol{B}_t^\star$ has a strong similarity to features of the training samples located close to $\boldsymbol{p}_q$. By contrast, if the given source frame is overlapped by multiple sources, the similarity is much weaker since the feature now is constructed from a time-varying combination of acoustic features. From this discussion, the single-source frames have a stronger similarity with a subset of the training features, while the overlapped frames, on the contrary, have a weaker similarity. Hence, we can separate the overlapped and single source frames by introducing a proper metric measuring the

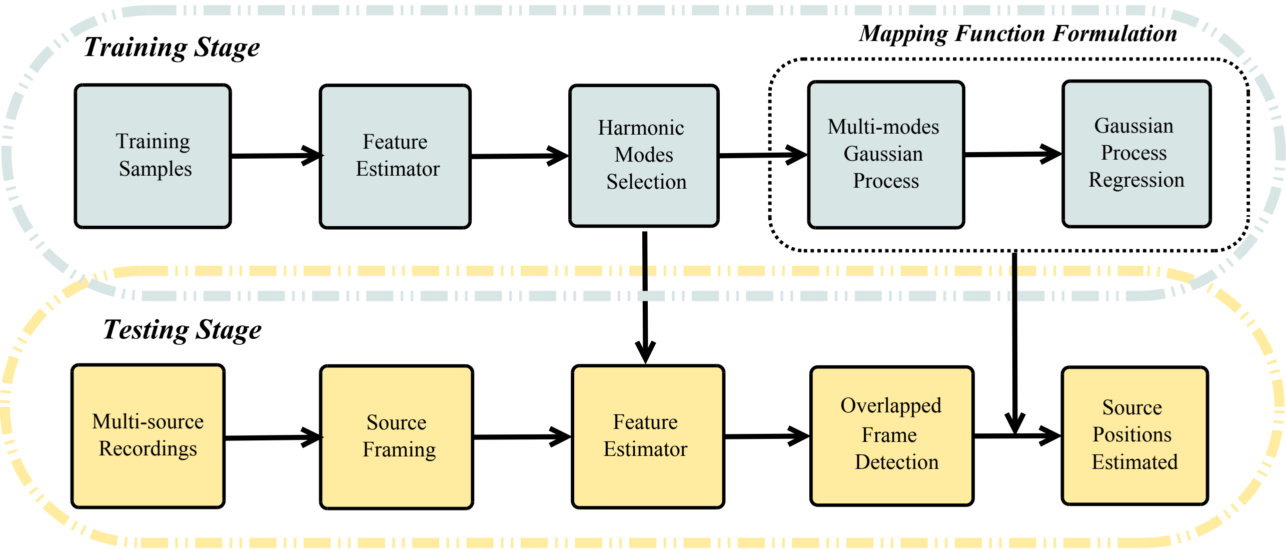

Fig. 4. Block diagram of the proposed multiple source localization approach, which mainly comprises of a training and test stage respectively.

similarity. For that, we use a distance function $\mathcal{T}(\cdot)$ to measure the similarity between the source features,

$$d(t, n_\ell) = \mathcal{T}(\boldsymbol{B}_t^\star, \boldsymbol{B}_{n_\ell}) \tag{37}$$

where $1 \leq t \leq T$ denotes the index of the segmented source frames in the time domain, $\boldsymbol{B}_{n_\ell}$ denotes the n_ℓ-th training feature matrix where $1 \leq n_\ell \leq \mathcal{N}_D$. Note that above-mentioned SE kernel function in (28), with unknown parameters, cannot be used as the distance function in (37). Several theoretical distance metrics can be used, such as the normalized Euclidean distance in (43) used in the experimental study. Intuitively, a smaller distance denotes the inputs have a stronger similarity. Then, we use a repetitive calculation over all the training samples to generate a vector,

$$\boldsymbol{d}(t) = \left[d(t,1), d(t,2), \cdots, d(t, \mathcal{N}_D)\right]^T \tag{38}$$

where both the labeled and unlabeled training samples are used because positional information is not required. A small subset of elements in $\boldsymbol{d}(t)$ is used to compute the distance,

$$d(t) = \frac{1}{I}\sum_{i=1}^{I} \boldsymbol{d}_i^{\mathrm{s}}(t) \tag{39}$$

where $\boldsymbol{d}^{\mathrm{s}}(t)$ denotes the ascending sorted vector of $\boldsymbol{d}(t)$. This measure is successively applied for all T frames to produce,

$$\boldsymbol{d} = \left[d(1), d(2), \cdots, d(T)\right]^T. \tag{40}$$

Intuitively, given the vector of $\boldsymbol{d}$, we choose the overlapped source frames, to be isolated from source localization, that satisfy the following inequality,

$$d(t) > \eta, \quad t = 1, \cdots, T \tag{41}$$

where η denotes a user defined threshold that is empirically specified. We emphasize the difference between a recently proposed detector in [43], as the current one in this paper requires a training feature set. Note that the detection here directly uses the source features, not requiring the source position information, so that both the labeled and unlabeled training samples are exploited.

VI. Experiments

A. *Experimental Methodology*

This section presents experimental results for multi-source localization in noisy and reverberant environments using both the simulated and real-life source recordings. The experiments are implemented following the procedures presented in Figure 4. As said, the source localization approaches localize the source x, y, z-coordinates separately. For simplicity, the following localization scheme focuses on x-coordinate of the sources as localization of other coordinates follows a similar procedure. Performance of our localization system is evaluated using a quantitative metric of mean absolute estimated error (MAEE) over the source's x-coordinate,

$$\frac{1}{Q}\sum_{q=1}^{Q} |x_{\text{ori}}(q) - x_{\text{est}}(q)| \tag{42}$$

where Q denotes the number of the sound sources presented in the environment, $x_{\text{ori}}(q)$ and $x_{\text{est}}(q)$ represents the original and estimated x-coordinate of the q-th sound source with respect to the origin of the room (not the microphone array). Note that the distance function of $\mathcal{T}(\cdot)$ in (37), required by the overlapped frame detection, has not been specified yet. Here, we choose to use the normalized Euclidean distance function,

$$\mathcal{T}(\boldsymbol{B}_t^\star, \boldsymbol{B}_i) = \frac{||\boldsymbol{B}_t^\star - \boldsymbol{B}_i||_2}{||\boldsymbol{B}_t^\star||_2 ||\boldsymbol{B}_i||_2} \tag{43}$$

in which $||\cdot||_2$ represents a ℓ_2 norm of the input feature matrix. Note that other distance metrics can be equally used for (43).

The experiment adopts two additional source localization approaches for comparison. (i) The distance function of (43), measuring the similarity between the source features of the testing and labeled training sources, is used. For this method, the estimated position corresponds to the labeled training source which locates closest to the testing source. (ii) The other is the state-of-art Multi-Node Gaussian Process based source localization approach using the RTF based source feature. The original algorithm recently proposed in [27] aims at single source localization and uses RTF between all pairs of microphones. For a fair comparison, we adjust and estimate the RTFs between all the pressure on the surface of the array and the one at the origin of the array, and then apply it to the multi-source localization assisted by the overlapped frame detector. Note that, some structured spherical arrays, such as the rigid spherical arrays, only have microphones on the array surface. For such a case, we approximate the pressure at the array origin as the addition of the ones on the surface for the RTF based localization method.

B. *Simulated Recordings*

The size of the simulated reverberant room is 6 × 4 × 3 m for the length, width and height, respectively. We set the left-front-bottom corner of the room as the reference origin for the source coordinates, i.e., $(0, 0, 0)$. We simulate a open-sphere spherical microphone array (32 channels and radius 4.2 cm), and place it at an unknown position in the room. Note that, although a spherical microphone array is used, the theory developed by this paper is equally applicable for other microphone arrays as well, such as a planar microphone array and circular microphone array. The time-domain room impulse response from the sound sources to the microphone array is generated using an available toolbox[2] that implements the image source method [54]. Speech signal randomly selected from the TIMIT database at the sampling frequency of 8 KHz is used as the input source signal. We use a convolution operation between the simulated room impulse response and speech signals to generate the measured recordings. After that, Gaussian white noise is added into the time domain recordings. Then, the measured noisy recordings are segmented into 0.5s frames with a 50% overlapping. The segmented time domain recordings are first transferred into the STFT domain and then decomposed into the spherical harmonics domain. Finally, the proposed estimator in (18) is used to compute the RHCs for all the segmented frames. Thirty frequency bins approximately ranging from 1500 Hz to 2500 Hz are exploited, which records the soundfield up to the 2nd order as $N = \lceil kr \rceil$ (i.e., 9 spherical harmonic modes). By contrast, lower frequency bins reduce the uniqueness of the RHC vector whose dimension is reduced to 4 (i.e., 4 spherical harmonic modes). The other drawback at low frequencies is the "Bessel zero problem", causing erroneous estimations of the desired spherical harmonics coefficients because the noise signal can be easily amplified [55]. Higher frequency bins contain less valid speech components.

TABLE I
MAEE OF SINGLE SOURCE LOCALIZATION USING DIFFERENT NUMBERS OF LABELED TRAINING SAMPLES.

Number	20	33	49	66	86
MAEE/m	0.380	0.314	0.248	0.238	0.223

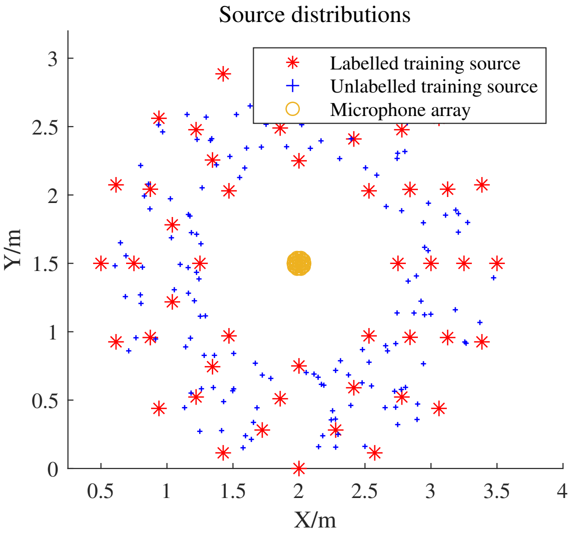

Fig. 5. Top view of the simulated source distribution. The labeled and unlabeled samples are represented by the red and blue points respectively.

1) Sound Source Area for Localization: In the experiments, we apply the proposed method to common scenes of group conversations between multiple speakers. Consider a specific scenario to localize the speakers in a conference room. Hence, the sitting area around the conference table is taken as the source area for localization. Our first task is to select a number of labeled and unlabeled training samples over the defined source area. We address the problem using two separate steps as follows: (i) labeled samples selection: Intuitively, the number of labeled training samples involves a trade-off: increasing the number generally leads to higher localization accuracy, while in return it increases the complexity of the system. This algorithm's overall practicality is considered by this paper. Hence, we select a relatively small number of labeled samples, while still achieving acceptable localization accuracy. Table I reports the accuracy of single source localization using an increasing number of labeled samples. From the results, we set the labeled training number to 49 because the accuracy starts to degrade severely when using a smaller number. (ii) Unlabeled samples selection: As explained, the unlabeled training samples are much easier to acquire, for simplicity, we directly select 250 unlabeled samples randomly distributed within the defined sound source area.

TABLE II
ACCURACY OF OVERLAPPED FRAME DETECTOR UNDER VARIOUS REVERBERATION TIMES, WHERE THE SNR LEVEL IS 25 DB.

T_{60}/ ms	300	400	500	600	700
Accuracy/%	75.0	73.3	71.7	68.3	65.0

TABLE III
ACCURACY OF OVERLAPPED FRAME DETECTOR UNDER VARIOUS SNR LEVELS, WHERE THE REVERBERATION TIME IS 700 MS.

SNR/dB	5	10	15	20	25
Accuracy/%	46.7	55.0	58.3	61.7	66.7

Figure 5 exhibits the sound source area filled by the selected training samples, which encircles the conference table whose radius is 0.75 m. The microphone array, placed at the center of the table, records the incoming soundfield in the rever-

[2] https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator

berant room. In the training stage, we measure the simulated soundfield due to each training source separately, and estimate the respective source feature using the feature estimator in (18). After that, we implement the spherical harmonic modes selection using the metric in (25), and for this particular example, we preserve four spherical harmonic modes in total, whose indexes of (n, m) are $(1, -1), (1,\ 1), (2, -2), (2,\ 2)$, respectively.

*2) **Accuracy of Overlapped Frame Detection**:* The localization scheme proposed in this paper exploits a pre-processing step of the overlapped frame detector. Hence, accuracy of the detection has a direct influence on the eventual localization performance. Prior to source localization, let us evaluate the effectiveness of the detector. We measure conversational recordings due to three speakers within the defined source area. The recordings, lasting 30 seconds in total, are measured in a reverberant room where $T_{60} = 700$ ms, and are then contaminated by Gaussian white noise with a SNR of 25 dB. The overlapped ratio by the mixed recordings in the time domain is approximately 30%. Note that the overlapped frame detector in (39) has a parameter I. The exact number I depends on the total trainings samples used by the detector. Throughout the simulations, we set at I by around 2% of all the training samples. Hence, $I = 6$ when we use around 300 training samples in simulations. Figure 6 exhibits the conversational recordings. The 4-th sub-figure presents the calculated distance of the source frames to the training set. The 5-th sub-figure at the bottom displays the detected overlapped periods. The results confirm the detector has successfully discovered most of the overlapped components. In the meanwhile, we notice that the detector occasionally detects the frames where the speech is weak or silent, i.e., absent or inactive speech. This is because the source feature is not accurately estimated there, thus has a larger distance to the training set. The capability to detect and remove the weak/inactive speech frames is beneficial for source localization because it ensures the selected frames contain valid speech signal.

We then examine the proposed detector using conversational recordings in diverse environments. We generate the multi-source recordings in different acoustic environments, involving simultaneously three speaker positions, with a 30% overlap ratio. Table II and III reports the performance of the detector at different reverberation and SNR levels, respectively. Note that, for each tested room reverberation time, we re-simulated all the training samples and re-calculate all the training feature set. For consistent results, we implement the evaluations up to five times. For each case, the three speakers originate from randomly selected source positions and use randomly selected speech sentences. Hence, each number in both Table II and Table III denotes the mean detection accuracy of the five groups of evaluations. The results demonstrate that the accuracy gradually degrades in a more complex environment. Under most scenarios, it is capable to recognize more than 50% of all the overlapped frames.

Finally, we confirm the direct influence of the overlapped frame detection on the localization accuracy. Five repetitive examinations are conducted in the $T_{60} = 700$ ms reverberant room where the SNR level is set at 25 dB. We still adopt three speakers whose overlapped ratio is 30%. We then segment the mixed recordings into 0.5 s frames, and then apply the overlapped frame detection to recognize the isolate the overlapped frames. Finally, we apply the proposed semi-supervised localization method to estimate the unknown speakers' positions. The average MAEE over the five groups of evaluations using the overlapped frame detection is 0.205 m. by contrast, the MAEE without the overlapped frame detection is degraded to 0.255 m.

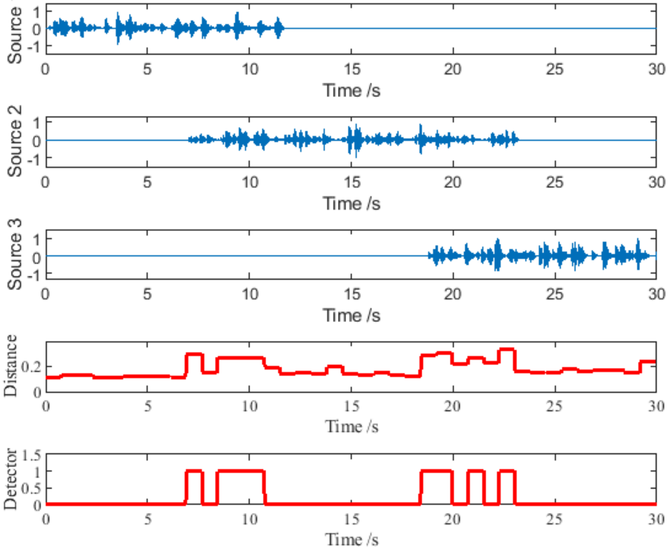

Fig. 6. Conversation between three speakers (30s long), and the performance of the overlapped frame detector. The distance, calculated by (43), denotes the similarity between the features of the testing frame and training set. A larger distance implies this frame is more likely to be an overlapped one.

TABLE IV
MAEE OF MULTIPLE SOURCE LOCALIZATION UNDER VARIOUS REVERBERATIONS, WHERE THE SNR LEVEL IS 15 DB.

MAEE/m Methods	Reverberation time (/ms)				
	300	400	500	600	700
RTF	0.183	0.214	0.253	0.240	0.265
Euclidean	0.301	0.288	0.259	0.296	0.298
All modes	0.207	0.237	0.229	0.259	0.285
Proposed	**0.179**	**0.166**	**0.186**	**0.194**	**0.228**

TABLE V
MAEE OF MULTIPLE SOURCE LOCALIZATION UNDER VARIOUS SNR LEVELS, WHERE THE REVERBERATION TIME IS 700 MS.

MAEE/m Methods	SNR levels (/dB)				
	5	10	15	20	25
RTF	0.333	0.301	0.279	0.273	0.244
Euclidean	0.311	0.327	0.315	0.289	0.250
All modes	0.336	0.282	0.289	0.267	0.260
Proposed	**0.246**	**0.221**	**0.232**	**0.192**	**0.204**

*3) **Performance of Multi-source Localization**:* Let us now evaluate the proposed localization method in comparison with the baseline methods. As introduced at the beginning of this section, one baseline is the RTF based method using Multi-Node Gaussian Process modeling in [27]. The other baseline directly uses a distance metric in (43). In addition, we also examine the proposed method without the spherical harmonic modes selection in order to analyze the proposed feature selector's influence on the localization accuracy. Therefore, four localization approaches are implemented, whose abbreviations used below for convenience are denoted by 'RTF', 'Euclidean', 'All modes' and 'Proposed', respectively. To increase the reliability of the results, under each acoustic environment (i.e., SNR and room reverberation time), ten successive examinations are implemented. And, each case uses three speakers with randomly selected source positions within the source area and randomly selected speech sentences. Hence, the values presented below denote the mean number over the ten successive evaluations.

Diverse acoustic environments are simulated. We first analyze the impacts of reverberation on the localization algorithms. Table IV displays the performance in different reverberation time ranging from 300 ms to 700 ms. In each varied reverberation, we re-simulated the training samples, optimized the parameters, and then applied the settings to the test stage. As expected, we observe that a longer reverberation time has negative impact on the localization accuracy. A longer reverberation time implies an increased complexity of the acoustic path from the sound sources to the recording area, increasing the difficulty to accurately model the relation between the source features and source positions. We then evaluate the algorithms under various noisy conditions (SNR level ranging from 5 dB to 25 dB). Table V depicts the results. We recognize slightly degraded localization accuracy when the SNR level decreases. The strong robustness to noise is a result of the proposed biased feature estimator in (18), which has already alleviated some noise components. Since the estimator has not fully cancelled the noise, the algorithms have non-negligible errors when the SNR level becomes very low. These results confirm the superiority of the proposed algorithm over the baseline methods. The improved accuracy when using selected harmonic modes, compared with that using all spherical harmonic modes, validates the effectiveness of the spherical harmonic modes selection.

TABLE VI
TIME COST BY TEN REPETITIVE EXECUTIONS AT THE TEST STAGE.

Methods	Number of views	Time
RTF	32	239.5s
All modes	9	69.3s
Proposed	4	30.4s

*4) **Algorithm Complexity Analysis**:* In addition to the localization accuracy, it is of necessity to evaluate the data-driven localization algorithm's computational complexity. Several factors determine the proposed algorithm's complexity, such as the number of labeled and unlabeled training samples, microphone channels in the array, and the soundfield order. Both our proposed method and the baseline using RTF adopt multi Gaussian Process modeling so that they generally follow similar procedures. Note that the RHC and RTF based methods use the Multi-Mode Gaussian Process and Multi-Node Gaussian Process, respectively. Both methods can be interpreted as an effort to capture or describe the acoustic event using multiple views. Thus we refer to them with the common terminology "Multi View Gaussian Process".

However, the numbers of views for the RHC and RTF based methods differ a lot, causing major consequences on the algorithm complexity. For validations, we evaluate the computational complexity of the algorithms by directly measuring their average time cost, using a Matlab implementation on a standard desktop (CPU Intel Core i7-4790 Quad 3.6 GHz, RAM 16 GB). Table VI presents the speed of the algorithms as well as their numbers of view. The proposed method is much faster than the baseline as it only has 4 views in total. Intuitively, a smaller number of views implies that less parameters should be adjusted. A comparison between the method using either selected number of modes or all the modes confirms the advantage of selecting modes on the computational complexity.

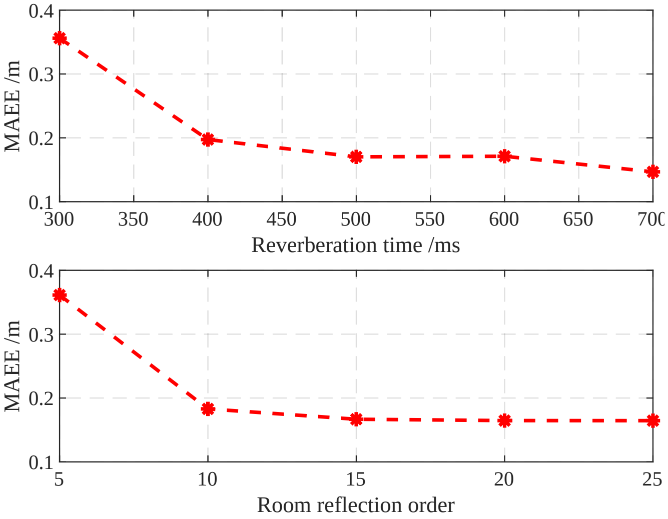

Fig. 7. MAEE of multiple source localization when room reverberation level is changed during the test stage (SNR is 25 dB). The different room reflection orders are with $T_{60} = 700$ ms.

*5) **Robustness to Test Environment Changes**:* As assumed, the source feature solely depends on the source position in a static room environment. Hence, aforementioned assessments assumed that the acoustic environment did not change between the train and test stages. However, this assumption hardly holds in practice. It occasionally happens that the setup of the room changes during the test stage. For example, the doors and windows may be opened or closed, or someone may walk around in the room. To meet practical requirements, our localization method should be robust to changes in the room characteristics. Hence, let us examine our method's robustness. Figure 7 reports the localization errors for room environments that are different between test and training stages. We simulate the changes in the test environment by using different room reverberation time as well as varied room reflection orders when $T_{60} = 700$ ms. In the training stage, we generate the training samples at the reverberation $T_{60} = 700$ ms, using a full reflection order. The examination results, presented in Figure 7, demonstrate slightly degraded accuracy when the test environment is not significantly different from that in the training stage. Hence, the localization method, learning the cues for localization in the training stage, is still applicable in the different/changed test environments. Additional evaluations at different reflection orders confirmed the improved localization accuracy at a higher reflection order. The reason is the testing source feature at a higher reflection order match more to the training features that captured a full reflection pattern. However, the Figure 7 implies the performance

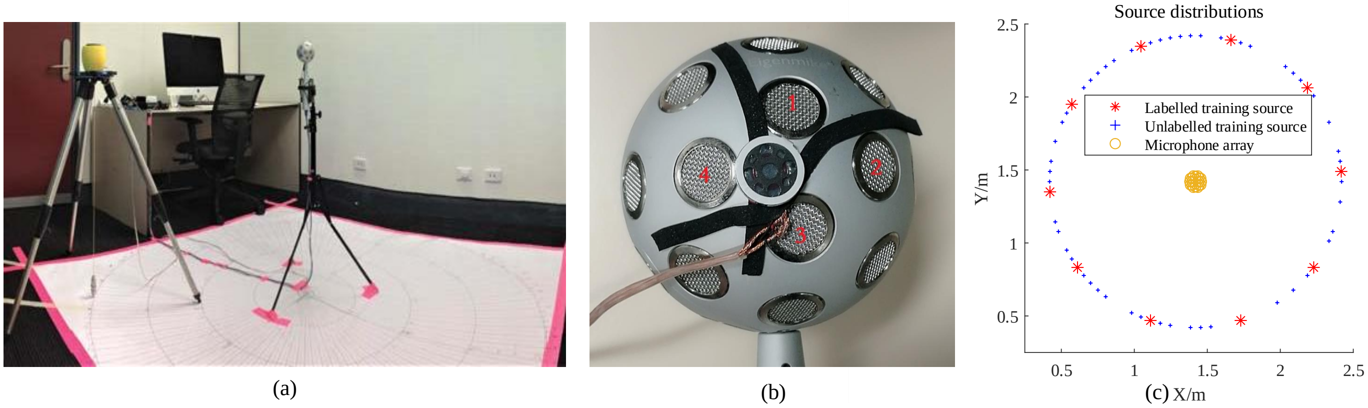

Fig. 8. (a): The setup for practical acoustic measurements used by our source localization approach in a reverberant room. (b): The commercial EigenMike and the mini-loudspeaker. (c): Top view of the defined source area in experiments, i.e., a 1m circle.

degrades more if the testing environment has more different characteristics in comparison with the training environment. And, it is recognized with dramatically reduced localization accuracy when the testing room environments change a lot (e.g., more than 0.35 m error when $T_{60} = 300$ ms or with room reflection order 5).

Additionally, the testing environment's temperature or air humidity also occasionally changes, which could be simulated by changing the speed of sound value by a few percent. Hence, we now change the speed of sound in the testing stage and examine the performance of the algorithm for both the training and testing stages the room reverberation time is $T_{60} = 700$ ms and the SNR level is 25 dB. Table VII presents the proposed method's MAEE with various sound speeds ranging from 336 m/s to 350 m/s. Note that the reference temperature, in the training stage, is 20 °C and the corresponding speed is 343 m/s. We observe that with varying values of speed (caused by changes in room temperature), the localization accuracy sometimes degrade. However with common indoor temperatures, the degradation is minimal.

TABLE VII
LOCALIZATION PERFORMANCE USING DIFFERENT SOUND SPEEDS IN THE TEST STAGE.

Speed (m/s)	336	339	343	346	350
Temperature (/°C)	8	14	20	25	30
MAEE/m	0.207	0.184	0.157	0.174	0.192

C. *Real Recordings*

This subsection validates the availability of the proposed algorithm under real-life scenarios, using practical recordings measured in the acoustic lab of Australian National University.

*1) **Experimental Setup**:* Figure 8 presents the setup for the practical measurements, a spherical microphone array called EigenMike and a circular source area, respectively. The EigenMike is a rigid 32-microphone array with a similar size as the above simulated open-sphere array. An advantage using a rigid array is avoiding the division by very small values in (12) at low frequencies, alleviating the aforementioned "Bessel zero problem". The defined source area only comprises of 10 labeled training samples along with 80 unlabeled training samples. The EigenMike, placed at the center of the source area, measures the incoming soundfield. The experiment room dimensions are $[3.54, 4.06, 2.70]$ for the length, width and height, respectively, with the reverberation time around $T_{60} =$ 330 ms. The same frequency band used by the simulated recordings, ranging from 1500 Hz to 2500Hz, is exploited for the real recordings.

Note that we obtain the real recordings using a convolution operation between the measured room impulse response (RIR) and the source signal. Hence, it is of great necessity to ensure high-quality RIR measurements. During practical recordings, the system time delay, caused by the hardware for example, is unavoidable. It degrades the spatial measurements accuracy if the unknown delay is large. Here, we provide a calibration technique to measure the delay by attaching a mini-loudspeaker (Manufacturer: VISATON, External Diameter: 16mm) close to the EigenMike (see Figure 8 (b)). Specifically, when driving the desired loudspeaker, we simultaneously drive the mini-loudspeaker using a known labeled signal. Since the two speakers are driven synchronously, the delay can be detected by location of the labeled signal within the measured recordings. Note that we just measure the system delay once as it generally keeps constant. When the delay is known, we then extract the source recordings right after the delay time where contains valid source signal.

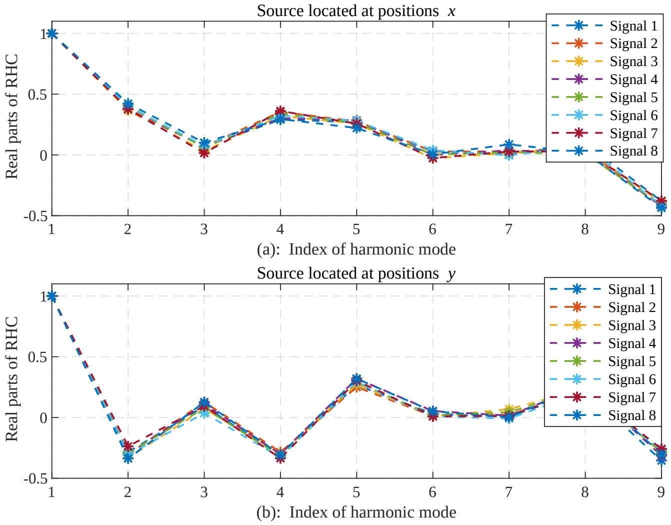

Fig. 9. Real parts of the features for sources located at different positions. Note that, for convenience, the presented values denote the average over the wide frequency band.

*2) **Validation of the Illustration in the Section II-C:*** Before presenting the localization accuracy, we first use real-life recordings to validate the illustration that the RHCs are independent of the particular source signal. We first compare the source features generated by the same sound source while using different source signal. For generality, ten pieces of random signal lasting around 0.5 second are used. For each signal, we calculate the mean values of the RHCs over a wide frequency band ranging from 1500 Hz to 2500 Hz. Figure 9 (a) depicts the real part of source features, using a sound source whose polar coordinates are $(r_1, \theta_1, \phi_1) = (1, 1.57, 3.63)$ with respect to the EigenMike's origin. For this sound source, its source feature is repetitively estimated using ten random signals. The observed consistency of the features using different random signal confirms its independence from the specific signal. Note that the curves presented in Figure 9 (a) also contain a slight inconsistency. One possible reason is the feature estimator in (18) uses a short frame windowing, which cannot cover the full reverberated test signal and therefore causes slight inconsistency on the estimated RHC.

TABLE VIII
AVERAGE MAEE OF MULTI-SOURCE LOCALIZATION USING 10 GROUPS.

Methods	RTF	Euclidean	All modes	Proposed
MAEE/m	0.159	0.205	0.181	0.120

TABLE IX
MAEE OF MULTIPLE SOURCE LOCALIZATION USING STRONG OVERLAPPED RECORDINGS.

MAEE/m	Overlapped ratio (%)				
Methods	50	60	70	80	90
RTF	0.192	0.187	0.193	0.206	0.214
Euclidean	0.217	0.223	0.244	0.214	0.209
All modes	0.191	0.195	0.202	0.211	0.205
Proposed	**0.141**	**0.143**	**0.146**	**0.161**	**0.175**

Then, we expect to see whether source feature significantly changes if placing the sound source at a different source position. Figure 9 (b) depicts the real part of source features, due to the sound source located at a new position, i.e., $(r_2, \theta_2, \phi_2) = (1, 1.57, 0.56)$ with respect to the array origin. We use the same setting to estimate the source features as the case in Figure 9 (a). We observe much greater differences between the source features in sub-figure (a) and (b), representing the sources located at different positions have different source features. Above analysis confirms, in a real-life reverberant room, the defined feature is mostly source-independent and changes significantly when the source position changes.

Finally, we add a quantitative study on how the RHC changes when the source moves to different positions. We first pick one reference position located at $(1, 1.57, 3.21)$ within the source area in Figure 8. Then, we move the source to the different positions with respect to the reference position and examine how the feature changes. For simplicity, the movement is carried along the azimuth axis only while elevation and distance are fixed. Note that we drive the source using a randomly generated signal and then use the proposed estimator to calculate the corresponding RHC. For quantitative evaluations, we use the normalized Euclidean distance function in (43) to measure the features' change. A larger distance value denotes the feature changes more significantly. Figure 10 denotes the changes of RHC against increasing value of the source azimuth change. It is observed that the RHC changes proportionally to the deviation of source azimuth. Aforementioned analysis using real recordings verifies the arguments that the defined feature is source-independent and mainly depends on the source position. Thus, we conclude that the RHC contains relevant cues to localize the source position.

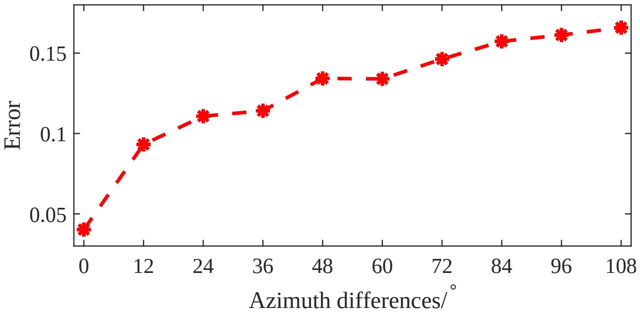

Fig. 10. The changes of the source feature with an increasing change of the source azimuths.

*3) **Localization Using Conversational Recordings:*** We exactly follow the steps summarized in Figure 4 to complete both the training and test stages. We use ten measurement groups, each containing three sound sources at randomly selected positions within the circular area. Each source uses a unique speech sentence lasting around 20 s, and the mixed multi-source recordings measured by the array have an overlapped ratio of about 30%. Table VIII presents the performance using all the algorithms. Each number denotes the mean MAEE over the ten measurements. Improved localization accuracy over the baselines confirms the relevance of the proposed multi-source localization approach under real-life scenarios.

*4) **Localization Using Significantly Overlapped Recordings:*** The aforementioned examinations of the algorithms are limited to conversational recordings, whose overlapped ratios are generally mild (e.g., overlapped ratio is 30% or less). In the remained content, we implement the proposed method at some unfavorable circumstances where the recordings have a severe overlapped ratio (e.g., higher than 50%). Figure 11 demonstrates significantly overlapping recordings. Then, we use the proposed detector to recognize the overlapped frames. The 4-th and 5-th sub-figure present the calculated distance to the training set and the detected overlapped periods, respectively. The results confirm that it successfully detects most of the overlapped components. We further evaluate the algorithm's localization accuracy using such severely overlapped recordings. We use ten measurement groups where each consists of three sound sources. The measured multi-source recordings have varied overlapped ratios ranging from 50% to 90%. Table IX reports the localization accuracy using all the algorithms. The results show slightly degraded localization accuracy when the overlapped ratio gradually increases. The reason is the overlapped frame detection accurately isolates most invalid frames (even when the overlapped ratio is up to 90%), thus all the approaches are then capable to localize the sources successfully. Being consistent with above evaluations, the proposed algorithm outperforms the baselines by achieving improved localization accuracy.

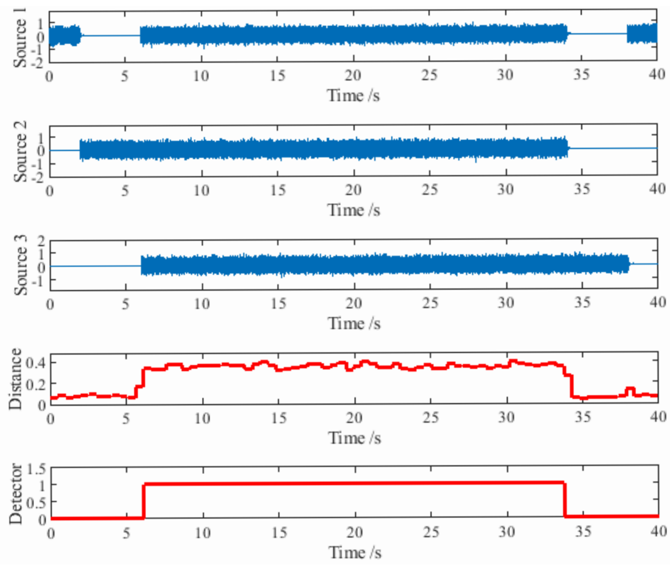

Fig. 11. Overlapped frame detector for significantly overlapped recordings. Around 70% of the recordings, in the middle, are overlapped by the three sources sending out random source signal.

VII. Conclusion

This paper has presented a semi-supervised multi-source localization algorithm in a noisy and reverberant environment, using a spherical harmonics domain source feature of the relative harmonic coefficients. Extensive simulations showed that the proposed algorithm achieved improved localization accuracy in comparison with the baseline methods tested in this study. Real-life evaluations confirmed the capability of this method even at unfavorable cases of severe source overlapping recordings. Several aspects of the proposed method are highlighted: (i) A comprehensive investigation of the relative harmonic coefficients: including a feature estimator in the noisy environment, a data-driven feature selector as well as an overlapped frame detector. (ii) The Multi-Mode Gaussian Process modeling (MMGP) nicely fuses the source features at the selected spherical harmonic modes, each representing a distinct/unique description of the soundfield. (iii) The unlabeled training samples not only enable a more precise measurement of the covariance for the MMGP modeling, but also play an active role in the source feature selection and overlapped frame detection, while exerting a negligible influence on the algorithm practicality. While the proposed method performs better than similar data based methods, some inherent limitations of it include: (i) the studied biased feature estimator with relatively short window frames may not fully cover a strong reverberation, which causes some inconsistency between the testing and training features in strong reverberant environments; (ii) current paper mainly considers the overlapped recordings so that is unusable for the simultaneous multi-source recordings, i.e., with an overlapped ratio of 100%. In the near future, we intend to propose a new feature estimator that better suits for strong noisy and reverberant environments and then achieve sufficient localization accuracy for simultaneous multiple source recordings in the complex environments.

Acknowledgment

The authors would like to thank Lachlan Birnie who assisted us to record the measurements in the acoustic lab at the Australian National University and also thank the anonymous reviewers for their valuable comments and suggestions that helped to improve the clarity and quality of this manuscript.

References

[1] N. Antonello, S. E. De, M. Moonen, P. A. Naylor, and W. T. Van, "Joint source localization and dereverberation by sound field interpolation using sparse regularization," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6892–6896.

[2] A. Fahim, P. N. Samarasinghe, and T. D. Abhayapala, "PSD estimation and source separation in a noisy reverberant environment using a spherical microphone array," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

[3] F. Asano, M. Goto, K. Itou, and H. Asoh, "Real-time sound source localization and separation system and its application to automatic speech recognition," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[4] Y. Hu, J. Benesty, and G. W. Elko, "Passive acoustic source localization for video camera steering," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 909–912.

[5] C. Evers, H. Loellmann, H. Mellmann, A. Schmidt, H. Barfuss, P. Naylor, and W. Kellermann, "The LOCATA challenge: Acoustic source localization and tracking," *arXiv preprint arXiv:1909.01008*, 2019.

[6] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, 1976.

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

[8] K. Yao, J. C. Chen, and R. E. Hudson, "Maximum-likelihood acoustic source localization: experimental results," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, pp. 2949–2952.

[9] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Springer Science, 2013.

[10] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "CRNN-based multiple DOA estimation using acoustic intensity features for ambisonics recordings," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 22–33, 2019.

[11] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *the Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 384–391, 2000.

[12] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[13] D. Khaykin and B. Rafaely, "Coherent signals direction-of-arrival estimation using a spherical microphone array: Frequency smoothing approach," in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 221–224.

[14] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 7, pp. 984–995, 1989.

[15] H. Sun, H. Teutsch, E. Mabande, and W. Kellermann, "Robust localization of multiple sources in reverberant environments using EB-ESPRIT with spherical microphone arrays," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 117–120.

[16] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 2814–2818.

[17] J. Pak and J. W. Shin, "Sound localization based on phase difference enhancement using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–1, 2019.

[18] S. Chakrabarty and E. A. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 8–21, 2019.

[19] A. Fahim, P. Samarasinghe, and T. D. Abhayapala, "Multi-source DOA estimation through pattern recognition of the modal coherence of a reverberant soundfield," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–1, 2019.

[20] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[21] S. Gannot and T. Dvorkind, "Microphone array speaker localizers using spatial-temporal information," *EURASIP journal on applied signal processing*, pp. 174–174, 2006.

[22] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1071–1086, 2009.

[23] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "A study on manifolds of acoustic responses," in *2016 International Conference on Latent Variable Analysis and Signal Separation*. Springer, pp. 203–210.

[24] ——, "Relative transfer function modeling for supervised source localization," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–4.

[25] ——, "Manifold-based Bayesian inference for semi-supervised source localization," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6335–6339.

[26] ——, "Semi-supervised sound source localization based on manifold regularization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 8, pp. 1393–1407, 2016.

[27] ——, "Semi-supervised source localization on multiple manifolds with distributed microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1477–1491, 2017.

[28] ——, "A hybrid approach for speaker tracking based on TDOA and data-driven models," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 26, no. 4, pp. 725–735, 2018.

[29] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of the direct-path relative transfer function for supervised sound-source localization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2171–2186, 2016.

[30] X. Li, Y. Ban, L. Girin, X. Alameda-Pineda, and R. Horaud, "Online localization and tracking of multiple moving speakers in reverberant environments," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 88–103, 2019.

[31] A. Brendel, B. Laufer-Goldshtein, S. Gannot, R. Talmon, and W. Kellermann, "Localization of an unknown number of speakers in adverse acoustic conditions using reliability information and diarization," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7898–7902.

[32] R. Opochinsky, B. Laufer-Goldshtein, S. Gannot, and G. Chechik, "Deep ranking-based sound source localization," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 283–287.

[33] Y. Hu, P. N. Samarasinghe, and T. D. Abhayapala, "Sound source localization using relative harmonic coefficients in modal domain," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 348–352.

[34] D. B. Ward and T. D. Abhayapala, "Reproduction of a plane-wave sound field using an array of loudspeakers," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 9, no. 6, pp. 697–707, 2001.

[35] H. Teutsch, *Modal array signal processing: principles and applications of acoustic wavefield decomposition*. Springer, 2007, vol. 348.

[36] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2193–2206, 2013.

[37] B. Jo and J. Choi, "Direction of arrival estimation using nonsingular spherical ESPRIT," *The Journal of the Acoustical Society of America*, vol. 143, no. 3, pp. 181–187, 2018.

[38] O. Nadiri and B. Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1494–1505, 2014.

[39] L. I. Birnie, T. D. Abhayapala, and P. N. Samarasinghe, "Reflection assisted sound source localization through a harmonic domain music framework," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 279–293, 2020.

[40] M. B. Çöteli and H. Hacıhabiboğlu, "Multiple sound source localization with rigid spherical microphone arrays via residual energy test," in *ICASSP 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 790–794.

[41] J. Daniel and S. Kitić, "Time domain velocity vector for retracing the multipath propagation," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 421–425.

[42] C. E. Rasmussen and C. K. Williams, *Gaussian process for machine learning*. MIT press, 2006.

[43] Y. Hu, P. N. Samarasinghe, T. D. Abhayapala, and S. Gannot, "Unsupervised multiple source localization using relative harmonic coefficients," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 571–575.

[44] E. G. Williams, *Fourier acoustics: sound radiation and nearfield acoustical holography*. Academic press, 1999.

[45] Y. Hu, T. D. Abhayapala, P. N. Samarasinghe, and S. Gannot, "Decoupled direction-of-arrival estimations using relative harmonic coefficients," in *2020 IEEE 28th European Signal Processing Conference (EUSIPCO)*, p. accepted.

[46] D. P. Jarrett, M. Taseska, E. A. Habets, and P. A. Naylor, "Noise reduction in the spherical harmonic domain using a tradeoff beamformer and narrowband doa estimates," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 5, pp. 967–978, 2014.

[47] S. Braun, D. P. Jarrett, J. Fischer, and E. A. P. Habets, "An informed spatial filter for dereverberation in the spherical harmonic domain," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 669–673.

[48] P. Samarasinghe, T. Abhayapala, M. Poletti, and T. Betlehem, "An efficient parameterization of the room transfer function," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 23, no. 12, pp. 2217–2227, 2015.

[49] T. D. Abhayapala and A. Gupta, "Spherical harmonic analysis of wavefields using multiple circular sensor arrays," *IEEE Transactions on Audio Speech and Language Processing*, vol. 18, no. 6, pp. 1655–1666, 2010.

[50] H. Chen, T. D. Abhayapala, and W. Zhang, "Theory and design of compact hybrid microphone arrays on two-dimensional planes for three-dimensional soundfield analysis," *Journal of the Acoustical Society of America*, vol. 138, no. 5, pp. 3081–3092, 2015.

[51] J. K. Nielsen, M. S. Kavalekalam, M. G. Christensen, and J. Boldt, "Model-based noise PSD estimation from speech in non-stationary noise," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5424–5428.

[52] I. Pzsit and Y. Yamane, "The variance-to-mean ratio in subcritical systems driven by a spallation source," *Annals of Nuclear Energy*, vol. 25, no. 9, pp. 667–676, 1998.

[53] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[54] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[55] P. Samarasinghe, T. Abhayapala, and M. Poletti, "Wavefield analysis over large areas using distributed higher order microphones," *IEEE-ACM Transactions on Audio Speech and Language Processing*, vol. 22, no. 3, pp. 647–658, 2014.