Multilevel B-Splines based Learning Approach for Sound Source Localization

Linh Nguyen, *Member, IEEE,* Jaime Valls Miro, *Member, IEEE,* and Xiaojun Qiu

***Abstract*—In this paper, a new learning approach for sound source localization is presented using ad hoc either synchronous or asynchronous distributed microphone networks based on time differences of arrival (TDOA) estimation. It is first to propose a new concept in which coordinates of a sound source location are defined as functions of TDOAs, computing for each pair of microphone signals in the network. Then, given a set of pre-recorded sound measurements and their corresponding source locations, the multilevel B-Splines based learning model is proposed to be trained by the input of the known TDOAs and the output of the known coordinates of the sound source locations. For a new acoustic source, if its sound signals are recorded, the correspondingly computed TDOAs can be fed into the learned model to predict the location of the new source. Superiorities of the proposed method are to incorporate acoustic characteristics of a targeted environment and even remaining uncertainty of TDOA estimations into the learning model before conducting its prediction and to be applicable for both synchronous or asynchronous distributed microphone sensor networks. Effectiveness of the proposed algorithm in terms of localization accuracy and computational cost in comparisons with state-of-the-art methods was extensively validated on both synthetic simulation experiments as well as in three real-life environments.**

***Index Terms*—Microphone array, sound source localization, multilevel B-splines, learning approach.**

I. Introduction

In various audio/acoustic based applications, localizing a sound source is a fundamental but still challenging problem. Apparently, the sound source localization (SSL) happens in many nowadays-concerned research topics comprising automatically steering a camera to the direction of a speaker in a teleconferencing room [1]–[4], separating multiple speaker speeches [5], detecting a source in an environment where it requires privacy preserve or has poor lighting conditions and occlusions [6], [7], search and rescure [8], [9] and mapping a 3D source in autonomous robotic systems [10], [11]. Though there is a variety of methods proposed for the SSL in the past decades, accurately and robustly localizing a sound source in an adverse environment with the gloom of noises and reflections/refractions is still not comprehensively understood.

Fundamentally, there are two categories of conventional methods proposed to find a source given its sound signals captured by microphone sensors. They both are mainly based on computing the generalized cross-correlation (GCC) [12], [13] of a microphone signal pair. The first type is a direct approach [14]–[16] that aims to maximize the steered response power (SRP) of the output of a delay and sum beamformer. That is, to locate a sound source, the algorithm has to exhaustively search the whole SRP space to find global maxima, which leads to its computational burden. Furthermore, the SRP technique is mostly limited to centralized and synchronous microphone network scenarios since it requires all synchronized sound signals available at its processing centre. That is, both synchronization and bandwidth requirements prevent the method from an asynchronous distributed network [17]. In contrast, localizing a sound source in the indirect method requires two separate steps [18]–[20]. Time difference of arrival (TDOA) is firstly estimated from the GCC peaks [21]. Then the source location, given the correspondingly estimated TDOAs, can be ascertained by optimally addressing an optimality criteria [22] such as the hypercone fitting problem [23]. Superiorities of the TDOA based approach is that it can be utilized in both synchronous and asynchronous distributed microphone networks as what required to be transmitted among sensor nodes are TDOA values not raw sound signal data. Nevertheless, results obtained the indirect method are quite sensitive to the presence of noises and reflections/refractions [24].

Up to now, most of the conventional approaches in the SSL context are proposed to employ merely measurements recorded by microphones at an instant time to localize the corresponding sources. Nonetheless, there also have recently some supervised and semi-supervised learning methods that utilize both prior information and current microphone recordings for the purpose of SSL. For instance, Deleforge *et al.* [25], [26] employed the manifold concept to develop a learning model for localizing both azimuth and elevation in a binaural system. The binaural manifold model is firstly learned from pre-recorded audio measurements by estimating its parameters using the closed form expectation maximization algorithm. Then when new observations are recorded the bearings of a sound source can be inferred in a fashion of the probabilistic Bayesian perspective. Similarly, by introducing definition of the relative transfer function, Laufer-Goldshtein *et al.* [27], [28] presented a new semi-supervised SSL method based on the manifold regularization that aims to retrieve the bearing azimuth of a sound source given its corresponding samples. In terms of SSL using a distributed ad-hoc microphone network, where coordinates of a source can be computed, authors in [29] proposed to use features as a function of relative transfer function samples. A Gaussian process model is utilized to present those features, where its parameters are estimated from

This work was supported by Sydney Water Operationalisation of the ACAPFP funded by Sydney Water Corporation and the Blue Sky Grant funded by University of Technology Sydney, Australia.

Linh Nguyen and Jaime Valls Miro are with Centre for Autonomous Systems, University of Technology Sydney, Australia (e-mail: vanlinh.nguyen@uts.edu.au, jaime.vallsmiro@uts.edu.au).

Xiaojun Qiu is with Centre for Audio, Acoustics and Vibration, University of Technology Sydney, Australia (e-mail: xiaojun.qiu@uts.edu.au).

the pre-recorded acoustic training data set by the use of the maximum likelihood algorithm. Given new audio observations, the learned Gaussian process model can now predict the source location. It is noted that a particular covariance function of the Gaussian process model may appropriately work for a particular environment. In [30], Li *et al.* employed a machine learning technique to estimate sound intensity in order to localize a sound source in scenarios using a small-sized microphone array. Furthermore, Wang *et al.* in [31] formulated sound source localization as a sparse signal recovery and parametric dictionary learning problem, which can be solved by the variational Bayesian expectation maximization method.

Acoustically, given geometrical configuration, each environment or space has its own acoustic characteristics, which consist of noisy levels, reverberations, reflections and some unknown features. Therefore, if one is more aware of attributes of an acoustic environment, they are more capable of accurately localizing a sound source positioned in that environment. Furthermore, we acknowledge that in most of popular environments such as offices, meeting rooms and conference rooms in commercial buildings and homes, the acoustic features are approximately unchanged over time (or at some period of time). It is clear that, in such scenarios, if there exist some pre-recorded acoustic measurements, the acoustic attributes of the environment can be learned a priori before being efficiently utilized to predict the location of any new sound source given its corresponding acoustic signals. In this paper, we propose a new learning model based on the multilevel B-Splines for this purpose. If the Gaussian process model is significantly dependent on its covariance function that must be intelligently selected, the multilevel B-Splines approximation consists of the predefined functions that enable our proposed model to be generically applied for any acoustic environments.

Contrasted with the supervised and semi-supervised learning methods aforesaid, the proposed approach relies on features of TDOAs. In equivalent words, it is assumed that TDOAs can be first estimated for each pair of microphone signals in a known training data set. The pre-recorded training measurements can be easily collected in advance in a given room by using a speaker (sound source) moving randomly around the room. At each position, the speaker's location is recorded and its sound signals are also observed by a microphone sensor network. We then present a new concept in which coordinates of a sound source location are defined as functions of TDOAs. By employing the multilevel B-Splines, we introduce a learning paradigm with the defined coordinate functions where a TDOA grid is hierarchically estimated, given prior information of both the coordinates of the source locations and the TDOAs in the training data set, which is ultimately utilized to interpolate the location of a source when new TDOAs are computationally observed.

It is apparent that TDOA estimation given adverse conditions of noisy and reverberant environments is highly uncertain as can be seen in any the indirect methods. Nevertheless, in the proposed approach, though TDOAs are still required to be smartly selected from their spurious counterparts, the remaining uncertainty in their estimation can be adapted by the learning model. More importantly, our proposed algorithm is independent from configuration of microphone array; that is, it can be employed in both the synchronous and asynchronous distributed sensor networks. Eventually, the proposed approach has been extensively validated in the synthetic simulation experiments as well as in the three real environments including a typical office, a large workstation room and a laboratory. The results obtained by our algorithm are highly promising when their accuracy of the source location estimation outperforms those ascertained by renowned state-of-the-art methods.

The remaining of the manuscript is organized as follows. Section II introduces how to compute TDOAs for pairs of microphone signals recorded in both the synchronous and asynchronous distributed microphone sensor networks. Note that procedures of selecting the best TDOA from its specious counterparts is also delineated in this section. In Section III, we interpret the multilevel B-Splines based learning strategy of localizing a sound source in a step-by-step fashion. The computational complexity of our algorithm is also given in this section. Section IV represents how the synthetic and real experiments were carried out, and the accuracy of the resulting localization as well as the operational cost of the proposed algorithm are compared with those ascertained by well-known state-of-the-art methods. Conclusions of the work are summarized in Section V.

II. TDOA Computation

As a first step of the proposed method, here presents how TDOAs are computed from the microphone measurements.

A. Signal Model

Consider a network of M microphone sensors that are deployed arbitrarily in a reverberation environment. A signal acquired by the m^{th} microphone ($m = 1, \cdots, M$) at time t can be presented by a reverberation model [32] as follows,

$$d_m(t) = h_m(t) \odot s(t) + \epsilon_m(t), \tag{1}$$

where $\odot$ denotes the linear convolution operator, $h_m(t)$ is the complete room impulse response from the sound source to the m^{th} microphone sensor, $s(t)$ is the sound source signal and $\epsilon(t)$ is the additive noise. Normally, $\epsilon(t)$ is assumed to be uncorrelated with $s(t)$ and a noise at another sensor.

In this work, all sound signals collected by the microphone sensors are processed in a frame to frame basis. Hence, samples of a frame with a length of L_f at the m^{th} microphone and a discrete time k can be specified by

$$\mathbf{d}_m(k) = [d_m(kL_f), d_m(kL_f + 1), \cdots, d_m(kL_f + L_f - 1)].$$

B. TDOA Interpretation

Let τ_{mn} define the time difference of arrival (TDOA) between the signals at the two any microphones m and n. By using the reverberation model introduced in the subsection II-A, the GCC for a pair of sound signals $d_m(t)$ and $d_n(t)$ in the frequency domain can be given as

$$\mathcal{R}_{mn}(\tau_{mn}) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Phi_{mn}(\omega) D_m(\omega) D_n^*(\omega) e^{j\omega\tau_{mn}} d\omega, \tag{2}$$

where * denotes the complex conjugation operator, $D_m(\omega)$ and $D_n(\omega)$ are the Fourier Transforms of the signals $d_m(t)$ and $d_n(t)$, respectively, and $\Phi_{mn}(\omega)$ is the combined frequency dependent weighting factor. This weighting factor is frequently defined by the well-known phase transform (PHAT) [12] for realistic applications as follows,

$$\Phi_{mn}(\omega) = \frac{1}{|D_m(\omega)D_n^*(\omega)|},$$

where $|\cdot|$ denotes an absolute operator. Eventually, the TDOA of each pair of the microphone signals can straightforwardly resolved by globally maximizing the GCC-PHAT in (2) as given below,

$$\widehat{\tau}_{mn} = \underset{\tau_{mn} \in [-\tau_{mn}^{max}, \tau_{mn}^{max}]}{\textbf{argmax}} \mathcal{R}_{mn}(\tau_{mn}), \qquad (3)$$

where $\tau_{mn}^{max} = \frac{\|\mathbf{l}_m - \mathbf{l}_n\|_2}{c}$, $\mathbf{l}_m$ is the location of the microphone m, c is the speed of sound propagation and $\|\cdot\|_2$ denotes the $l^2 - norm$ operator.

Nonetheless, due to ambient noise and reverberation conditions in the environment, which cause severe deteriorations in the received signals, the accuracy of the TDOA estimation is substantially influenced. Two methods in the following will be introduced to reduce effect of the disruptive noise and reflections on the TDOA results in both the synchronous and asynchronous distributed microphone networks.

1) Synchronous Networks: In a synchronized microphone network, our approach proposes to employ only three sensors to localize a sound source. Thus, it is computationally practical to better tune the TDOA for each microphone signal pair by using geometrical interpretation [23]. In particular, in the first step, it is proposed to employ the zero-sum condition to disambiguate the TDOAs from the spurious ones. P time delays ($P = 10$ in our experiments) corresponding to the P largest local maxima of the GCC of each signal pair are selected. Then the best combination of the TDOAs in the network of three microphones must theoretically hold the condition

$$\tau_{12} + \tau_{23} + \tau_{31} = 0. \qquad (4)$$

Nevertheless, due to erroneous and noisy TDOAs, the condition (4) can be relaxed to $|\tau_{12} + \tau_{23} + \tau_{31}| < \zeta$, where ζ is a defined minute positive number, which results in a possible set of the time delays for each pair of the microphones.

In the second step, three quality metrics including average of normalized GCC peaks, average between GCC maxima and product of all ratios between first and seconds peaks of all the GCCs are used to form a quality score function, where each metric is factored at a proper weight. By maximizing the quality score function, the corresponding solution time delays, which are highly associated with the direct paths from the sound source, are the three best TDOAs for the three synchronous microphone network.

2) Asynchronous Distributed Networks: For an asynchronous distributed network, it is proposed to utilize only four microphones into two unsynchronized nodes, where each node has a pair of synchronized sensors, meaning there are only two TDOAs obtained from the network at a particular time. Since the microphones on the whole network are unsynchronized, the zero-sum condition is not applicable. Here it is proposed to employ the local window search [33] on the GCC to find the TDOA in each sensor node.

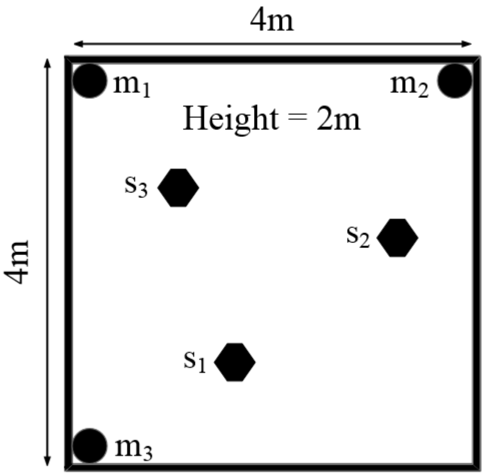


Fig. 1: A reverberant room with a set-up of a synchronous microphone network and a sound source randomly moving around, including: microphones $\mathrm{m_i}$ and sound source footprints when it stops $\mathrm{s_i}$.

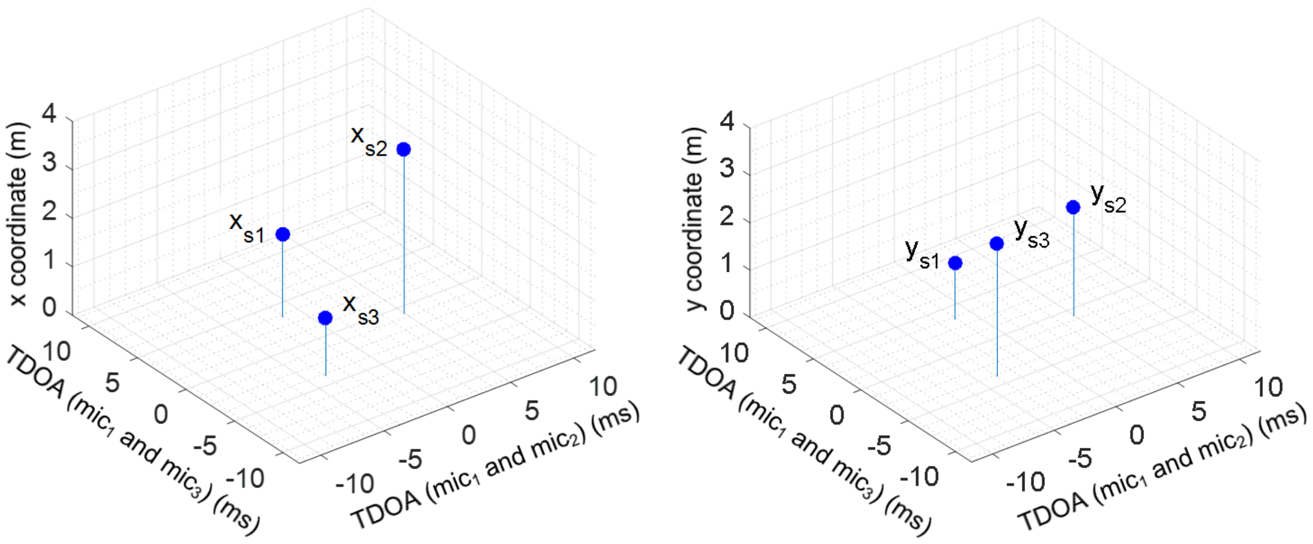
Similar to the fist step of the method for the synchronous network, one also selects P largest peaks of the GCC at each microphone node. For each element in the selected set, one computes an energy ratio between sum of that peak element and its P_n neighbour samples on the GCC and sum of the remaining samples on the GCC. As shown in [33] the energy ratio is reliable for discriminating the true peaks from the specious ones. The time delay corresponding to the GCC local peak whose energy ratio is maximum is the TDOA for the sensor node.

III. Multilevel B-Splines based Learning Sound Source Localization

This section introduces a novel model that first learns acoustic characteristics of a given environment from the pre-recorded measurement then predicts a location of a source when new TDOAs are computed from the observed acoustic measurements.

A. Source Coordinates against TDOAs

Let us consider the sound source localization in a noisy and reverberant room with a three synchronous microphone network and a sound source set up, demonstrated in Fig. 1. Three microphone are deployed on the walls of a shoe-box shaped room of dimensions 4 m × 4 m × 3 m. It is noticed that there is only one sound source in the environment, and it is assumed to randomly move around. At each stop, the sound source is assumed to emit sound signals and the microphones record them to compute TDOAs. In the example, three randomly chosen locations of the sound source when it stops are located at $s_1 = [1.7, 1.2, 2]$, $s_2 = [3.4, 2.3, 2]$ and $s_3 = [1.2, 2.8, 2]$. TDOAs are limited by the maximum



(a)

(b)

Fig. 2: Sound source coordinates against TDOAs: (a) x coordinate and (b) y coordinate.

possible TDOA at each pair of the sensors, which can be straightforwardly obtained when the microphone locations are known. Given a acoustic source, three TDOAs can be ideally obtained by the microphone network. Nonetheless, only any two TDOAs are needed in solving the sound source localization problem, assuming τ_{12} and τ_{13} herein.

It can be seen that at a particular scenario, when a sound source location changes, τ_{12} and τ_{13} vary accordingly. Equivalently, on the opposite way, when the microphones produce the different τ_{12} and τ_{13}, the sound source has a new location. Fig. 2 demonstrates the coordinates of the three source locations s_1, s_2 and s_3 depending on the TDOAs. Therefore, we propose to define coordinates of the sound source as functions of the TDOAs.

$$f_s = f(\tau_{12}, \tau_{13}), \tag{5}$$

where f_s is a coordinate of the source. Therefore, the sound source location can be found, given τ_{12} and τ_{13}, if the function $f(\tau_{12}, \tau_{13})$ is known. In other words, a model can be learned from TDOAs as inputs and the sound source locations as outputs.

For the sake of simplicity, let τ_1 and τ_2 denote any two TDOAs obtained by the microphone network. We define $\mathcal{Q} = \{(\tau_1, \tau_2) | -\tau_{max} \leq \tau_1, \tau_2 \leq \tau_{max}\}$ as a domain of the TDOAs. In addition, let $x(\tau_1, \tau_2)$ and $y(\tau_1, \tau_2)$ denote the x and y coordinate functions of the sound source location. To formulate the functions $x(\tau_1, \tau_2)$ and $y(\tau_1, \tau_2)$, let us discretize the TDOA domain $\mathcal{Q}$ into a $n_1 \times n_2$ TDOA grid.

Given the uniform cubic B-spline basis functions [34] defined as

$$\begin{aligned}
F_1(u) &= \frac{(1-u)^3}{6}, \\
F_2(u) &= \frac{3u^3 - 6u^2 + 4}{6}, \\
F_3(u) &= \frac{-3u^3 + 3u^2 + 3u + 1}{6}, \\
F_4(u) &= \frac{u^3}{6},
\end{aligned}$$

where $0 \leq u < 1$, the functions $x(\tau_1, \tau_2)$ and $y(\tau_1, \tau_2)$ can be specified in the following form,

$$f_s(\tau_1, \tau_2) = \Sigma_{p=1}^4 \Sigma_{q=1}^4 F_p(u) F_q(v) W(i+p, j+q), \tag{6}$$

where

$$\begin{aligned}
u &= |\tau_1 - \lfloor \tau_1 \rfloor|, \\
v &= |\tau_2 - \lfloor \tau_2 \rfloor|, \\
i &= \begin{cases} \lfloor \tau_1 \rfloor - 1, & \tau_1 \geq 0 \\ \lfloor \tau_1 \rfloor + 1, & \tau_1 < 0 \end{cases} \\
j &= \begin{cases} \lfloor \tau_2 \rfloor - 1, & \tau_2 \geq 0 \\ \lfloor \tau_2 \rfloor + 1, & \tau_2 < 0 \end{cases}
\end{aligned}$$

and $W(i+p, j+q)$ is a weight at the TDOA grid cell $(i+p, j+q)$. Note that $i \in \{-\lfloor \frac{n_1}{2} \rfloor - 2, -\lfloor \frac{n_1}{2} \rfloor - 1, \cdots, \lfloor \frac{n_1}{2} \rfloor + 1\}$, $j \in \{-\lfloor \frac{n_2}{2} \rfloor - 2, -\lfloor \frac{n_2}{2} \rfloor - 1, \cdots, \lfloor \frac{n_2}{2} \rfloor + 1\}$ and $\lfloor \cdot \rfloor$ denotes the floor operator.

The sound source can be localized by simply substituting τ_1 and τ_2 into (6) if the weights at the TDOA grid cells are known. The following section will introduce how to compute those parameters on the grid mesh.

B. Sound Source Location (SSL) Inference

Let us consider a known source at the location $[x_s, y_s, 2]$ in the room in Fig. 1. Without loss of generality, it is supposed that the two any TDOAs obtained by the array of the microphones are τ_{1s} and τ_{2s}. We also define $W^x(p,q)$ and $W^y(p,q)$ as the grid cell weights of the TDOA domains for x and y coordinates, respectively. Then the TDOA grid cells are weighted [34] as follows,

$$W^x(p,q) = \frac{F_p(u_s)F_q(v_s)x_s}{\Sigma_{p=1}^4\Sigma_{q=1}^4(F_p(u_s)F_q(v_s))^2}, \tag{7}$$

$$W^y(p,q) = \frac{F_p(u_s)F_q(v_s)y_s}{\Sigma_{p=1}^4\Sigma_{q=1}^4(F_p(u_s)F_q(v_s))^2}, \tag{8}$$

where $u_s = |\tau_{1s} - \lfloor\tau_{1s}\rfloor|$ and $v_s = |\tau_{2s} - \lfloor\tau_{2s}\rfloor|$.

It can be clearly seen that (7) and (8) can only handle the weights at the grid cells that are neighbours of the point (τ_{1s}, τ_{2s}). As a result, we propose to employ a set of training data of pre-recordings to learn the weights on the whole TDOA grid. In equivalent words, it is assumed that there are multiple sound sources, which are randomly positioned in the room but their locations are known (we can utilize one speaker to move around the room and multiple recordings at various locations are gathered in sequences). For each set of pre-recorded measurements from a location-known sound source, the TDOAs are computed. To capture the characteristics of an acoustic environment, all the TDOAs in the training data are obtained from the sound signals by using the method introduced in Section II, not by using geometries of the microphone network and the source locations. (7) and (8) are then used to compute the weights at the grid cells given their corresponding TDOAs neighbours. The more sound sources are known, the more coverage of computationally weighted grid cells is. In the worst case, if a grid cell is faraway from the TDOAs associated with all the location-known sound sources, it cannot be mathematically weighted. In that case, we define the grid cell weight as zero. On the other hand, there are also many grid cells assigned multiple weights from their known TDOA neighbours. Averaged weights at those shared grid cells can be computed by

$$W^x(i,j) = \frac{\Sigma_n(F_p(u_{sn})F_q(v_{sn}))^2 W^x(p,q)}{\Sigma_n(F_p(u_{sn})F_q(v_{sn}))^2}, \tag{9}$$

$$W^y(i,j) = \frac{\Sigma_n(F_p(u_{sn})F_q(v_{sn}))^2 W^y(p,q)}{\Sigma_n(F_p(u_{sn})F_q(v_{sn}))^2}, \tag{10}$$

where $p = i + 1 - \lfloor\tau_{1sn}\rfloor$, $q = j + 1 - \lfloor\tau_{2sn}\rfloor$, $u_{sn} = |\tau_{1sn} - \lfloor\tau_{1sn}\rfloor|$, $v_{sn} = |\tau_{2sn} - \lfloor\tau_{2sn}\rfloor|$, and τ_{1sn} and τ_{2sn} are all the TDOA neighbours of the grid cell (i,j).

After the TDOA grid on the $\mathcal{Q}$ domain is learned, a location of any unknown sound source can be estimated. That is, computing the TDOAs from microphone recordings and substituting them into (6), coordinates of the sound source location can be straightforwardly obtained. Nonetheless, uncertainty of the source location estimation is significantly dependent on the size of the TDOA grid cell. For instance, if the grid is too coarse, training data could be mingled together. On the other hand, if the grid is pretty fine, the grid cell weight is restricted to a small vicinity. Both scenarios leads to erroneous estimation of the source location. Consequently, we herein propose to employ hierarchical architecture of the TDOA coarse-to-fine grids, where the coordinates of the sound source location can be summed up in sequential steps.

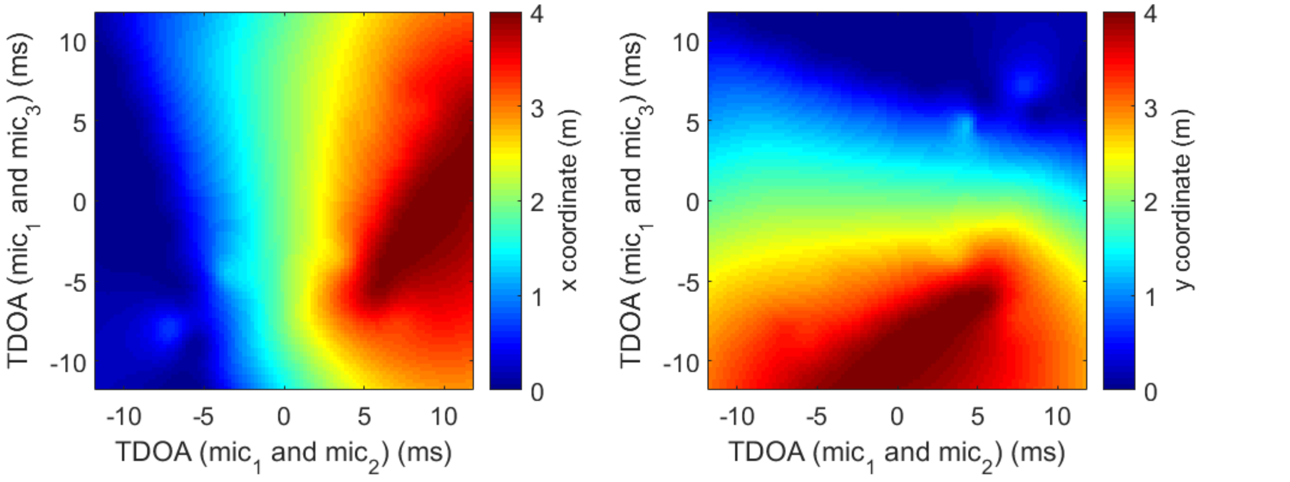
In the first step, a very coarse grid is created on the TDOA domain $\mathcal{Q}$. By employing equations (6)-(10), we recompute the coordinates of the known source locations, which are then utilized to calculate uncertainties of the estimations. For instance, errors of the coordinates of the known source s and their estimations are

$$\delta^1_{x_s} = x_s - f^1_{x_s}(\tau_{1s}, \tau_{2s}), \tag{11}$$

$$\delta^1_{y_s} = y_s - f^1_{y_s}(\tau_{1s}, \tau_{2s}), \tag{12}$$

where $f^1_{x_s}(\tau_{1s}, \tau_{2s})$ and $f^1_{y_s}(\tau_{1s}, \tau_{2s})$ are the estimations of x_s and y_s, respectively.

In the second step, a less coarser grid is created on the domain $\mathcal{Q}$, where sizes of the grid cell is halved as compared to those in the grid in the previous step. Both x_s and y_s are replaced by $\delta^1_{x_s}$ and $\delta^1_{y_s}$; and then we find the new



(a)

(b)

Fig. 3: Sound source coordinates maps (a) x coordinate and (b) y coordinate.

estimations and deviations $f^2_{x_s}(\tau_{1s}, \tau_{2s})$, $f^2_{y_s}(\tau_{1s}, \tau_{2s})$, $\delta^2_{x_s}$ and $\delta^2_{y_s}$, respectively. The algorithm repeatedly runs till the current grid is fine enough or a maximum deviation is lower than a defined threshold. Eventually, coordinates of an unknown sound source location s' can be interpolated, for instance,

$$x_{s'} = f_{x_{s'}}(\tau_{1s'}, \tau_{2s'}) = \Sigma_g f^g_{x_{s'}}(\tau_{1s'}, \tau_{2s'}), \tag{13}$$

$$y_{s'} = f_{y_{s'}}(\tau_{1s'}, \tau_{2s'}) = \Sigma_g f^g_{y_{s'}}(\tau_{1s'}, \tau_{2s'}), \tag{14}$$

where $\tau_{1s'}$ and $\tau_{2s'}$ are the corresponding TDOAs obtained by the microphone network and g is the number of steps where the TDOA grid is recreated. The proposed approach is summarized in Algorithm 1.

Algorithm 1 Multilevel B-Spline based learning algorithm for the sound source localization

Input:

1) Locations of microphones
2) Measurements recorded by microphones
3) Number of TDOA grids being created g
4) Sizes of the first TDOA grid $n_1 \times n_2$

Output:

1) Estimated locations of sound sources

```
 1: for each pair of microphones do
 2:     Compute GCC
 3:     Find all possible TDOAs
 4:     Disambiguate a real TDOA from spurious ones
 5: end for
 6: Select only two TDOAs for each network
 7: For learning
 8: Deploy the sound sources in an environment as many as possible
 9: Record their locations and sound signals
10: Compute TDOAs corresponding each sound source location in steps 2, 3 and 4
11: for $k$ = 1 to $g$ do
12:     Create a $n_1 \times n_2$ TDOA grid
13:     Compute the grid cells $W^x(i, j)$ and $W^y(i, j)$ in (9) and (10)
14:     for Each known source location $s$ do
15:         Compute estimated coordinates $f^k_{x_s}(\tau_{1s}, \tau_{2s})$ and $f^k_{y_s}(\tau_{1s}, \tau_{2s})$ by using (6)
16:         Compute estimation errors $\delta^k_{x_s}$ and $\delta^k_{y_s}$ in (11) and (12)
17:         $x_s \leftarrow \delta^k_{x_s}$
18:         $y_s \leftarrow \delta^k_{y_s}$
19:     end for
20:     $n_1 \leftarrow 2n_1$
21:     $n_2 \leftarrow 2n_2$
22: end for
23: For prediction
24: for Each unknown source location $s'$ do
25:     Given the weights of the TDOA grid cells learned, compute the estimated coordinates of the unknown source location $x_{s'}$ and $y_{s'}$ in (13) and (14)
26: end for
```

Let us take the room in Fig. 1 as an example, where 50 sets of sound signals are recorded correspondingly to 50 location-known sound sources deployed randomly in the room, using for the training purpose. After learning from the training data, the proposed algorithm can predict coordinates of any sound source locations in the room, given their corresponding TDOAs obtained by the microphone network. Fig. 3 illustrates the source coordinates x and y of all possible sound source positions in the room as the TDOAs are covered in a full range of maximum time delays. Note that in this example, the room has reverberation time of $T_{60} = 0.2s$ and a signal-to-noise ratio (SNR) of 30 dB.

C. Computational Complexity

Computational cost is one of significant factors to consider feasibility of a sound source localization method in real-time applications. Here, we consider computing time of our proposed approach for both the synchronous and asynchronous distributed networks.

Let N_t define the number of training data, which is a set of the pre-recorded measurements from the N_t location-known sources. The complexity of computing and discriminating TDOAs from the spurious peaks of the GCC time delays are specified by

$$\text{TDOA}_{cost} \approx N_t \left[\frac{408}{9} L_f \log_2(2L_f) + 14L_f \right] \tag{15}$$

for a synchronous microphone network [23], and

$$\text{TDOA}_{cost} \approx N_t \left[\frac{598}{9} L_f \log_2(2L_f) + 38L_f \right] \tag{16}$$

for an asynchronous distributed microphone network [33].

In addition, as presented in Section III-B, to interpolate a sound source, given its corresponding estimated TDOAs, the TDOA grid is repeatedly designed at every step in which its cell weights are also recomputed. If we define n_{1f} and n_{2f} as the sizes of the finest TDOA grid, then operational cost to compute all these steps until the grid is fine enough is as follows [34],

$$SSL - Inference_{cost} \approx cN_t + \frac{4}{3} n_{1f} n_{2f}. \tag{17}$$

Therefore, the overall computational cost of our algorithm to localize a sound source is summed up by (15) and (17) for a synchronous network and (16) and (17) for an asynchronous distributed network, respectively.

IV. Experimental Results

Effectiveness of the proposed approach for the sound source localization in a noisy and reverberant condition was evaluated in both synthetic simulation and real-life environments where physical phenomena were encountered. Moreover, the performances of the proposed method are compared with those of the conventional well-known techniques including SRP-PHAT and SRC.

A. A Synchronous Network of Microphones

1) Simulations: We extensively carried out different experiments on a simulated environment by using the room impulse response (RIR) generator [35], an implementation of the acoustic image method [36]. The simulated room was set up as demonstrated in Fig. 1. Two different reverberation times including $T_{60} = 0.2s$ and $T_{60} = 0.6s$, which are normally recommended for typical rooms, were set for the simulations. In all the simulations, the sampling rate and the sound propagation speed were set to 16 kHz and 343 m/s, respectively. At first, room impulse response from the source recorded by a microphone in the RIR generator was convolved with a source signal, which is a female speech utterance of length 8 s. The convolution result was then contaminated by the white Gaussian noise, generating a noisy and reverberant synthetic recording at the corresponding microphone sensor. Note that variance of the white Gaussian noise can be set to different values, which define various SNRs in diverse background noise environments.

As delineated in Section III-B, 50 sound sources were randomly deployed in the room and their corresponding known locations and sound signals were recorded for the training purposes. To validate the proposed algorithm, another 50 sound sources were also generated in the room with an assumption that their locations were unknown, whose emitting sound signals were also recorded by three synchronized microphones. In the signal processing procedure, each microphone recording was split into frames with a length of 2048 samples and overlap of 50%, and for each source position, we computed two TDOAs τ_{12} and τ_{13}. By learning the TDOA grid from the training data set, our approach can hence work out coordinates of a location of an unknown sound source given its corresponding TDOAs obtained by the microphone network. In this example, the starting number of the grid cells on the first TDOA grid is 1 and the algorithm stopped when the number of the grid cells on the current grid reached 128×128.

Due to random deployments, we repeated the simulation experiment at each scenario, given a reverberation time and a SNR, 1000 Monte-Carlo trials. The averaged results of root mean square error (RMSE) for each scenario were calculated and are illustrated in Fig. 4. The proposed approach apparently outperforms the well-known state-of-the-art methods in both the examined scenarios of noisy and reverberant environments, where as expected, the lower reflection and the higher SNR conditions are, the better the sound source location is estimated. Note that since the SPR-PHAT method is based on searching the source location coordinates at vertices of the discrete spatial grids of the source location space, we created those spatial grids at two different resolutions of 0.05 m and 0.1 m, respectively.

2) Real experiments: To study the performances of the proposed method in real-life scenarios, we conducted two real experiments in the realistic noisy and reverberant conditions in the campus of University of Technology Sydney, Australia. Note that the experimented rooms are daily working environments where staff and students were walking, talking, discussing and doing their own works. Therefore, there have noises from many sources such as human activities, door opening or closing, ventilation fans, air conditioners and research/study equipments. There also exist polluted noises from the traffic roads nearby. All these presented noise elements were naturally captured by the microphones during the recordings, which leads to the very noisy measurements in our experiments. The reverberation times and SNRs in both realistic environments were unknown at the time of the experiments.

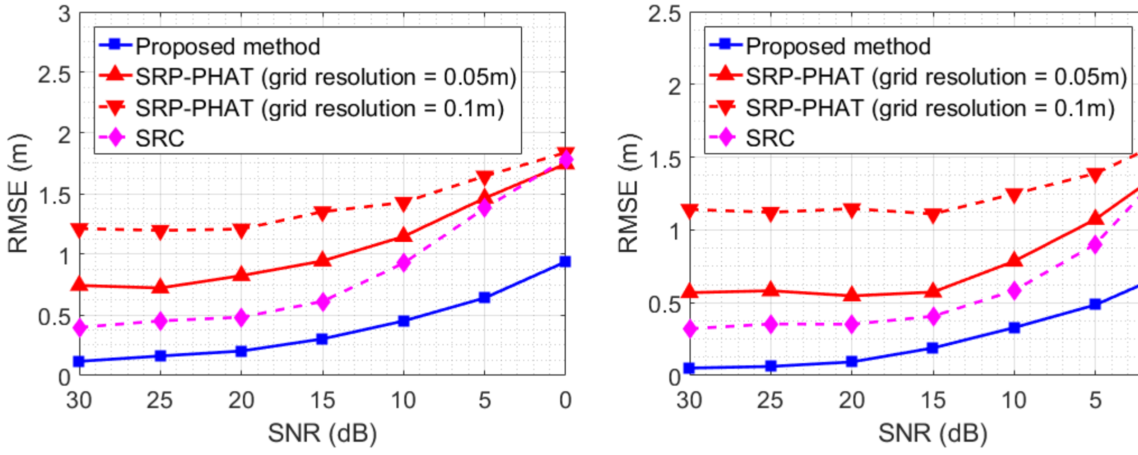
The experiment equipment comprised three the G.R.A.S. type 40PH free-field microphones and two National Instruments modules of the compact data acquisition cDAQ-9171 and the ADCs NI 9234. The sound source was a mobile phone playing an audio recording of a eight-second speech utterance by a female speaker. The sound signals were then recorded by the NI LabVIEW 2014®. All the measurements were sampled at frequency of 16kHz with a resolution of 24-bits. The procedures of the signal processing were similar to those in the simulation experiments.

In the first real experiment, we set up a test in a *typical office room* with approximate dimensions of 2.6 m × 3.8 m × 3.0 m. The room is formed by two partition walls and two glass walls with a glass door on one wall. There were office furnitures and of course two desktops and some other equipment presenting at the experimental time. Three microphones were located at a height of 1.04 m. The microphone numbered 1 was positioned at one corner of the room and the other two against the two walls perpendicular at that corner. Distances from the microphones numbered 2 and 3 to 1 were 0.97 m and 1.07 m, respectively. The experimental setup is shown in Fig. 5. In the experiment, the speaker was manually deployed at 20 different locations in the room, and 20 corresponding sets of measurements were recorded. Of which 10 sets of the recordings were used to learn the TDOA grid cells' weights in the training step, and the others were utilized to validate the estimated results of the sound source locations. The measurements of the acoustic signals emitted from each validating sound source were employed to input into the learned models. The outputs of those models are coordinates of a predicted location of the validating sound source, which were then utilized to compare with those of the realistic location. Errors between the predicted and real locations of the 10 validating sound sources were averaged in one reporting parameter as the RMSE. The resulting RMSEs are compared and shown in Table I, which confirms the outperformance of our algorithm.

TABLE I: RMSEs FOR THE TYPICAL OFFICE ROOM EXPERIMENT

Proposed	SRC	SRP-PHAT (grid resolution = 0.05 m)	SRP-PHAT (grid resolution = 0.1 m)
0.30 m	0.83 m	0.76 m	1.00 m

In the second realistic experiment, a similar setup of the measurement system in the office room was replicated in a much larger *workstation room* of the Centre for Autonomous Systems, which is a shared space for many staff and postgraduate students. The room is not a shoe-box but a nearly "L"



(a)

(b)

Fig. 4: Synchronized microphone synthetic experiments: RMSEs against SNRs where (a) $T_{60} = 0.6s$ and (b) $T_{60} = 0.2s$.

shape. Its length, maximum width and height are about 21.6 m, 10.1 m and 3.0 m, respectively. Two sides of the room are surrounded by private offices, and the glass doors of those offices and the partitions are intermingled on its walls. The other four walls consist of two partitions and two glass ones. Many items appeared inside the room when the experiments were conducted includes staff's and students' workstations, lockers, printers, furnitures and other study equipment. Due to restrictions of the working and studying environment, we only conducted the experiments in one of corners of the room with an area of 4 m × 4 m. 20 sets of the sound measurements were also collected and processed accordingly, as presented in the first real experiment. The results are summarised in Table II, which demonstrates robustness of the proposed approach against different adverse conditions while the performance of SRP-PHAT is better as compared with itself in the office experiment since reflections come at late stages of the sound signals in a large room.

TABLE II: RMSEs FOR THE WORKSTATION ROOM EXPERIMENT

Proposed	SRC	SRP-PHAT (grid resolution = 0.05 m)	SRP-PHAT (grid resolution = 0.1 m)
0.31 m	1.11 m	0.61 m	0.59 m

More importantly, in the proposed scheme, the more training data used to learn the TDOA grid is, the higher accuracy of the sound source localization can be archived. In particular, in both the real-life experiments, we varied the number of the training data sets from 10 to 15 and utilized the rest for the validations in each scenario. The results in terms of RMSEs are illustrated in Fig. 6. It can be clearly seen that the uncertainty of the source localization gradually goes down when the number of the training measurements is increased. In other words, if one is more aware of characteristics of an acoustic environment, they are more capable of accurately localizing a sound source positioned in that environment.

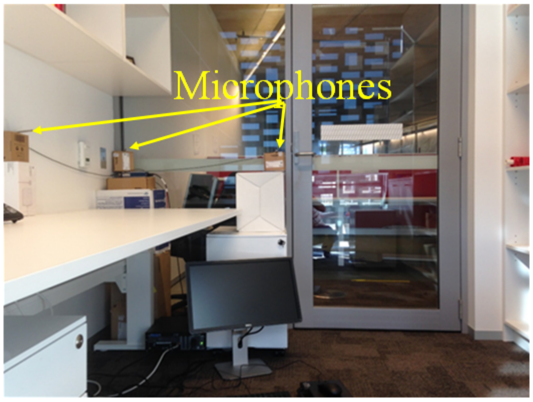


Fig. 5: Synchronized microphone real experiments: The office room layout.

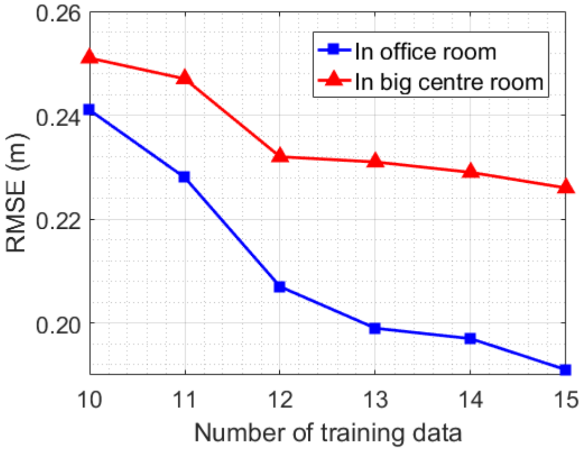


Fig. 6: Synchronized microphone real experiments: RMSEs against the number of the training data sets.

B. An Asynchronous Distributed Network of Microphones

1) Simulation: This section presents the results of the simulation experiments that verify the performance of our proposed algorithm employed in an asynchronous distributed

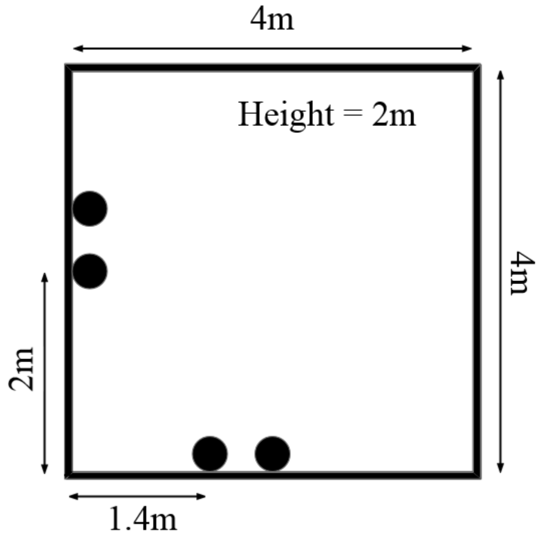
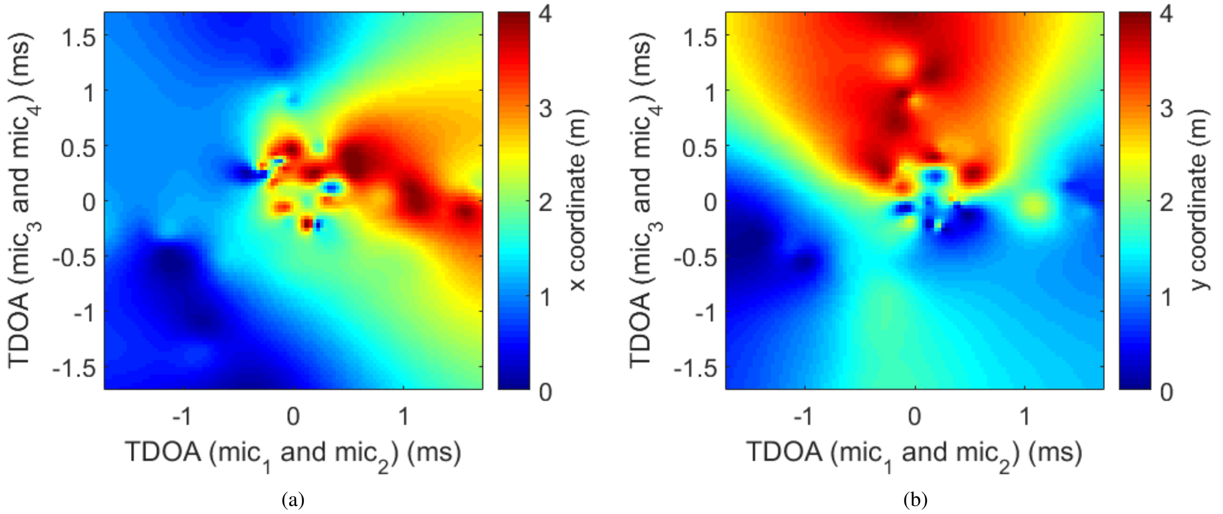


Fig. 7: Asynchronous distributed microphone synthetic experiments: The room layout.

microphone network. In this illustrated example, we utilized the simulation tool similar to that used in the experiments of the synchronized microphone network. The experimental settings and signal processing were also obeyed the strategy delineated in Section IV-A1. Nonetheless, as mentioned in Section II-B2, our algorithm proposes to use only 4 microphones that are grouped into two pairs of synchronized nodes but the nodes are asynchronous. The nodes were placed on the two walls of the room measuring 4 m × 4 m × 3 m, as demonstrated in Fig. 7. We conducted the experiments by first randomly deploying 50 sound sources in the room and their positions were known to the algorithm for the purposes of learning the TDOA grid cell weights. It is noticed that only one TDOA for each node was computed in this type of the proposed asynchronous distributed network. If one assumed that the room was given a reverberation time of $T_{60} = 0.6s$ and SNR = 0 dB, then the x and y possible coordinates of all potential sound source locations in the room were mapped to their corresponding TDOAs and are now shown in Fig. 8a and 8b, respectively. It can be seen that under a more adverse condition, the maps of the estimated coordinates of the source locations are rougher than their counterparts obtained in a less adverse environment shown in Fig. 3.

Furthermore, we considered another 50 assumingly unknown location sound sources by recording their corresponding sound signals. For each source position, we computed the error between the estimated location and its true peer. Every experiment with a given condition of the reverberation time and SNR combination was repeated 1000 trials. The summarized results in terms of averaged RMSEs are plotted in Fig. 9 for two cases of which the reverberation times are $T_{60} = 0.2s$ and $T_{60} = 0.6s$, respectively.

2) Real experiments: To validate the proposed method in a realistic scenario given an asynchronous distributed microphone network, we deployed four the G.R.A.S. microphones in a laboratory in a distributed manner. In other words, two the microphones were connected to a personal computer (PC) through the data acquisition cDAQ-9171, while the two others were connected to another PC. And two PCs were able to communicate through a local area network. Dimension of the laboratory is about 5.7 m × 12.1 m × 3.0 m. The laboratory experimental setup is illustrated in Fig. 10. It can be seen that two couples of the acoustic sensors were against the two opposite walls, where each couple recorded sound signals and saved them in a separate PC. Distance between the microphones in each couple was about 0.70 m, and they were located, from floor, at 1.74 m height in the right wall and 1.52 m height in the left wall, respectively. The sound source, in this example, presented by a laptop was manually moved around and randomly stopped at 30 locations. At each random location, the speaker played a male speech of 7 s length, and the acoustic signals were recorded. In 30 sets of the acoustic measurements, 15 of which were employed to train the TDOA



(a)

(b)

Fig. 8: Asynchronous distributed microphone synthetic experiments: Sound source coordinate maps (a) x coordinate and (b) y coordinate.

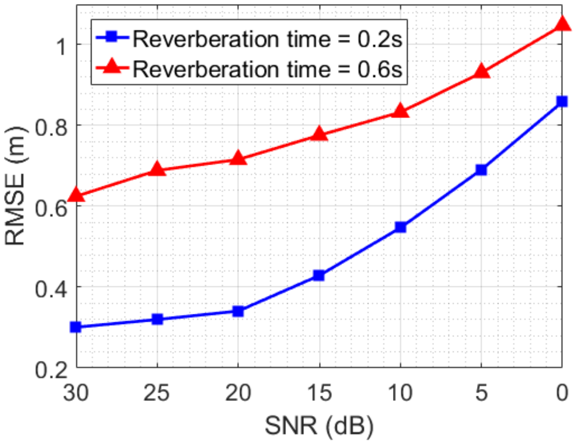


Fig. 9: Asynchronous distributed microphone synthetic experiments: RMSEs against SNRs.

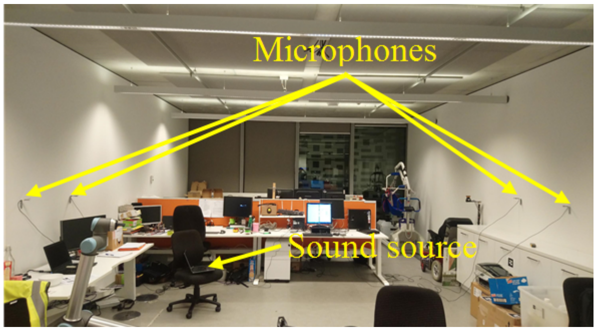


Fig. 10: Asynchronous distributed microphone realistic experiments: The laboratory layout.

grid cell weights as introduced in the previous sections, while the rest was utilized for the purpose of validation. Given the measurements recorded from a particular validating sound source, the learned models estimated a corresponding location, which was then compared with its realistic peer. The RMSE in the 15 validations cases shows that the estimated location of the sound source is about 0.47 m away from the ground truth. It is noted that the reverberation times and SNRs in this illustrated example were comprehensively unknown and the laboratory environment observed was very noisy.

C. Computing Time

To discuss efficiency of our proposed approach in terms of computation, this section demonstrates the total computing time of finding a sound source given its corresponding sound signals captured by the G.R.A.S microphones in the two real experiments in the synchronous networks. Note that all the processing procedures were carried out on the platform of Matlab 2016® on a 64-bit PC with computational capability of Intel(R) Core(TM) i5-6500 CPU @ 3.20 GHz and memory of 8.00 GB. The running time of our algorithm is compared with those of SRC and SRP-PHAT with discrete grid resolutions of 0.05 m and 0.1 m, respectively. The results are shown in Table III.

As can be seen in Table III, give the number of training data, our method does not depend on sizes of an environment while SRP-PHAT does. More importantly, its computing time requirement is significantly less than those of other renowned state-of-the-art techniques, which is highly potential for various real-time audio applications.

TABLE III: COMPARISONS OF COMPUTING TIME IN REAL EXPERIMENTS

Room	Proposed	SRC	SRP-PHAT (grid resolution = 0.05 m)	SRP-PHAT (grid resolution = 0.1 m)
Typical office	3.16 s	16.80 s	37.06 s	19.43 s
Workstation room	2.93 s	16.20 s	523.19 s	140.91 s

V. Conclusions

The paper has addressed the problem of localizing a sound source using a distributed microphone network in both synchronous and asynchronous architectures. A novel learning paradigm has been proposed, which relies on the TDOA features. In other words, coordinates of a sound source location are formulated as functions of TDOAs, which are employed in the multilevel B-Splines based learning model. It is assumed that a set of pre-recorded sound signals can be gathered a priori in a targeted environment, accompanied by corresponding locations of the emitting sound sources. This training data set is utilized to hierarchically estimate TDOA grids for the coordinate functions, which allows the proposed learning model to effectively predict a location of a new acoustic source when its corresponding TDOAs are computationally obtained from the microphone measurements. The new approach were substantially tested in the simulated experiments as well as the realistic environments of the typical office, workstation room and laboratory scenarios in a university campus during business hours, given both synchronization and asynchronization of the distributed microphones. The resulting localization accuracy obtained by our proposed algorithm and its computational cost outperform those obtained by the well-known state-of-the-art techniques including SRP-PHAT and SRC. The proposed approach will be advanced to localize multiple sound sources in our future works.

References

[1] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proc. EEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, August 1997, pp. 187–190.

[2] Y. Huang, J. Benesty, and G. W. Elko, "Passive acoustic source localization for video camera steering," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, June 2000, pp. 909–912.

[3] F. Antonacci, M. Matteucci, D. Migliore, D. Riva, A. Sarti, M. Tagliasacchi, and S. Tubaro, "Tracking multiple acoustic sources in reverberant environments using regularized particle filter," in *Proc. IEEE International Conference on Digital Signal Processing,*, Cardiff, UK, July 2007, pp. 99–102.

[4] Q. Yan, J. Chen, G. Ottoy, and L. D. Strycker, "Robust AOA based acoustic source localization method with unreliable measurements," *Signal Processing*, vol. 152, pp. 13–21, 2018.

[5] M. I. Mandel, R. J. Weiss, and D. P. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18(2), p. 382394, 2010.

[6] L. Parker, B. Birch, and C. Reardon, "Indoor target intercept using an acoustic sensor network and dual wavefront path planning," in *Proc. IEEE/RSJ International Conferece on Intelligent Robots and Systems*, Las Vegas, NV, USA, December 2003, pp. 278–283.

[7] K. Na, Y. Kim, and H. Cha, *Acoustic sensor network-based parking lot surveillance system*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, p. 247262.

[8] T. Latif, E. Whitmire, T. Novak, and A. Bozkurt, "Sound localization sensors for search and rescue biobots," *IEEE Sensors Journal*, vol. 16(10), pp. 3444–3453, 2016.

[9] L. Wang and A. Cavallaro, "Acoustic sensing from a multi-rotor drone," *IEEE Sensors Journal*, vol. 18(11), pp. 4570–4582, 2018.

[10] D. Su, K. Nakamura, K. Nakadai, and J. V. Miro, "Robust sound source mapping using three-layered selective audio rays for mobile robots," in *Proc. IEEE/RSJ International Conferece on Intelligent Robots and Systems*, Daejeon, Korea, October 2016, pp. 2771–2777.

[11] D. Su, T. V. Calleja, and J. V. Miro, "Towards real-time 3D sound sources mapping with linear microphone arrays," in *Proc. IEEE International Conferece on Robotics and Automation*, Singapore, Singapore, May 2017, pp. 1662–1668.

[12] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24(4), pp. 320–327, 1976.

[13] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, *Robust Localization in Reverberant Rooms*. Springer, Berlin, Heidelberg, 2001, ch. 8, pp. 157–180.

[14] P. Aarabi, "The fusion of distributed microphone arrays for sound localization," *EURASIP Journal on Advances in Signal Processing*, vol. 2003(4), p. 338347, 2003.

[15] H. F. Silverman, Y. Yu, J. M. Sachar, and W. R. P. III, "Performance of real-time source-location estimators for a large-aperture microphone array," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 13(4), p. 593606, 2005.

[16] M. Cobos, A. Marti, and J. J. Lopez, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling," *IEEE Signal Processing Letters*, vol. 18(1), pp. 71–74, 2011.

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

[18] Y. Huang, J. Benesty, G. Elko, and R. Mersereati, "Real-time passive source localization: A practical linear-correction least-squares approach," *IEEE Transactions on Speech and Audio Processing*, vol. 9(8), pp. 943–956, 2001.

[19] M. Compagnoni, P. Bestagini, F. Antonacci, A. Sarti, and S. Tubaro, "Localization of acoustic sources through the fitting of propagation cones using multiple independent arrays," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 20(7), pp. 1964–1975, 2012.

[20] A. Canclini, F. Antonacci, A. Sarti, and S. Tubaro, "Acoustic source localization with distributed asynchronous microphone networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 21(2), p. 439443, 2013.

[21] T. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Signal Processing*, vol. 85(1), pp. 177–204, 2005.

[22] K. Ho, "Bias reduction for an explicit solution of source localization using tdoa," *IEEE Transactions on Signal Processing*, vol. 60(5), p. 21012114, 2012.

[23] A. Canclini, P. Bestagini, F. Antonacci, M. Compagnoni, A. Sarti, and S. Tubaro, "A robust and low-complexity source localization algorithm for asynchronous distributed microphone networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23(10), pp. 1563–1575, 2015.

[24] R. Parisi, A. Cirillo, M. Panella, and A. Uncini, "Source localization in reverberant environments by consistent peak selection," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, HI, USA, April 2007, pp. 37–40.

[25] A. Deleforge, F. Forbes, and R. Horaud, "Variational em for binaural sound-source separation and localization," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, BC, Canada, October 2013, pp. 76–80.

[26] ——, "Acoustic space learning for sound-source separation and localization on binaural manifolds," *International Journal of Neural Systems*, vol. 25(1), p. article number: 1440003, 2015.

[27] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Semi-supervised sound source localization based on manifold regularization," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24(8), pp. 1393–1407, 2016.

[28] ——, "Manifold-based Bayesian inference for semi-supervised source localization," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Shanghai, China, March 2016, pp. 6335–6339.

[29] ——, "Semi-supervised source localization on multiple manifolds with distributed microphones," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25(7), pp. 1477–1491, 2017.

[30] Y. Li and H. Chen, "Reverberation robust feature extraction for sound source localization using a small-sized microphone array," *IEEE Sensors Journal*, vol. 17(19), pp. 6331–6339, 2017.

[31] L. Wang, Y. Liu, L. Zhao, Q. Wang, X. Zeng, and K. Chen, "Acoustic source localization in strong reverberant environment by parametric bayesian dictionary learning," *Signal Processing*, vol. 143, pp. 232–240, 2018.

[32] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, pp. 1–19, 2006.

[33] Q. Zhang, Z. Chen, and F. Yin, "Distributed marginalized auxiliary particle filter for speaker tracking in distributed microphone networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24(11), pp. 1921–1934, 2016.

[34] S. Lee, G. Wolberg, and S. Y. Shin, "Scattered data interpretation with multilevel B-splines," *IEEE Transactions on Visualization and Compute Graphics*, vol. 3(3), pp. 228–244, 1997.

[35] E. A. P. Habets, "Room impulse response generator," 2010. [Online]. Available: https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator

[36] J. Allen and D. Berkley, "Image method for efficiently simulating small room acoustics," *Journal of Acoustical Society of Ameria*, vol. 65(4), pp. 943–950, 1979.

Linh Nguyen received his PhD degree in Robotics from the University of Technology Sydney (UTS), Australia in March 2015. He then worked at the same university as a Postdoctoral Research Associate until September 2015. From January to July 2016, he was as Research Fellow at the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. He rejoined the Centre for Autonomous Systems at UTS in August 2016, where he is currently a Research Fellow. His research interest includes robotics, internet of robotic things, artificial intelligence, machine learning, signal processing and non-destructive testing.

Jaime Valls Miro received his B.Eng. and M.Eng. in Computer Science (Systems Engineering) from the Valencia Polytechnic University (UPV, Spain), in 1990 and 1993 respectively. He received his Ph.D. in robotics and control systems from Middlesex University (UK) in 1998, and worked in the underwater robotics industry as a software and control systems analyst until 2003. In 2004 he joined the Centre for Autonomous Systems in UTS (Australia), where he is currently an Associate Professor. His areas of interest span across the field of robotics, most notably modelling sensor behaviours for perception and mapping, computational Intelligence in HRI - assistive robotics in particular, and robot navigation. In the last few years he has devoted this interest in pursuing a better understating of condition assessment sensing for critical water mains in close collaboration with the water industry.

Xiaojun Qiu received his Bachelor and Master degrees from Peking University in 1989 and 1992, and his PhD from Nanjing University in 1995, all majoring in Acoustics. He worked in the University of Adelaide as a Research Fellow in the field of active noise control from 1997 to 2002, worked in the Institute of Acoustics of Nanjing University as a professor on Acoustics and Signal processing from 2002 to 2013, and worked at RMIT University as a Professor of Design on Audio Engineering from 2013 to 2016. He joined University of Technology Sydney in 2016 as a professor in Audio, Acoustics and Vibration and founded the center for Audio, Acoustics and Vibration there. His main research areas include noise and vibration control, room acoustics, electro-acoustics, and audio signal processing, particularly, applications of active control technologies. He is a Fellow of Audio Engineering Society, and serves as an Associate Technical Editor for the Journal of Audio Engineering Society for many years.