Design and In-Field Testing of Target Sound-Source Positioning with Insights from Indoor Acoustic Environment

Xiyu Song[1], Zhenghong Liu[1], Shiqi Wang[1], Fangzhi Yao[1], and Mei Wang[1,2(✉)]

[1] Ministry of Education Key Laboratory of Cognitive Radio and Information Processing, Guilin University of Electronic Technology, Guilin 541004, China
mwang@glut.edu.cn

[2] School of Information Science and Engineering, Guilin University of Technology, Guilin 541006, China

Abstract. The sound-source localization method derived by combining the steered-response power phase transform (SRP-PHAT) method with stochastic region contraction (SRC) is one of the most effective localization methods currently available. It can yield accurate target sound-source localization results in weak-noise and moderate-reverberation environments. However, owing to the unstructured room space (which entails many points to be searched), the SRP-PHAT-SRC localization method using grid searches is computationally heavy and exhibits poor real-time performance. Therefore, we propose an improved method using insights from the indoor acoustic environment; our model estimates the room geometry via acoustic scene reconstruction and triangulates this estimated geometry via (offline) Delaunay triangulation for structured room space volumes [the volumes are selectively assigned (online) to SRC initialization when positioning]; the method searches for the target sound source within a more trusted space in the room and thereby eliminates computationally wasteful searches in regions where the sound source does not appear. When the position estimates of the motion target are updated, an increased number of points (position estimates) can be used to update the room space-volume structure (these are local updates that do not increase online positioning complexity), making the selective volume more efficient for the next searching task. We have verified the feasibility of the proposed method in improving positioning time consumption and demonstrated its good practical application prospects.

Keyword: Sound-source localization; SRP-PHAT-SRC; acoustic scene reconstruction; Delaunay triangulation

1 Introduction

Indoor sound-source localization (ISSL) is a fundamental task in location-based services (LBS) [1–5]; there are several representative examples, including map services for LBS users, message push, voice enhancement in online conferencing systems, and targeted

J. Wang (Ed.): MobiMedia 2023, LNICST 576, pp. 137–161, 2025.
https://doi.org/10.1007/978-3-031-60347-1_13

listening. In all ISSL tasks, the continuous positioning of the target sound source represents a key problem, owing to the need to track targets to obtain location information for the application requirements.

Microphone-array-based sound-source localization methods can better perceive the surrounding environment (and thereby locate the sound source) than those with a single-microphone. For example, simultaneous localization and mapping (SLAM) robots are equipped with one or multiple arrays to reconstruct the trajectory of the moving sound source [6–12] (i.e., to locate the moving sound sources by interacting with a room map whilst also using the trajectories of the moving sound sources to complete that room map). However, this method is a type of Bayesian filtering and depends upon data association; it often contains significant uncertainty in data associations, owing to online data losses and mismatching; it is unsuitable for real-time ISSL. Another typical example is the research into deep-learning-based sound-source positioning and tracking, which is effective in product research and scene applications that do not require rapid response times; however, if the online speech transcription application or immersive auditory experience application requires rapid responses, the deep-learning-based method is not beneficial [13–15].

According to the relevant theories of architectural acoustics, if the room geometry [i.e., not the room shape in detail but the locations of the most important reflectors (e.g., ceiling, floor, and walls)] is known, then we can localize the sound source, predict where its echoes will originate from using simple geometric rules, determine the positions of the image sound sources, and obtain the real sound-source position. The process of determining the room geometry is called acoustic scene reconstruction in the SCENIC project proposal [16].

However, audio processing tasks are complicated when dealing with complex indoor environments. The main problem is the indefinite number of noise sources and the general ambient noise (which pollutes the sound signal originating from the source of interest) [17]. Traditional array-based ISSL solutions help to overcome many of these shortcomings [18–21]; they exploit spatial diversity using multiple microphones to simultaneously acquire different versions of the emitted source signals, which are then jointly processed. Generally, these solutions can be classified into three categories: time-difference-of-arrival [22], high-resolution spectral estimation [23, 24], and steered-response power (SRP) [25]. Compared with the first two, the SRP approach with a power-phase transform (SRP-PHAT) is very attractive in acoustic applications because of its robustness under noise and reverberation conditions [26]. Despite its robustness, real-time processing remains challenging because the SRP search space contains many local extrema, and the method employs an exhaustive grid search scheme to examine numerous candidates.

Many improved methods have been developed for real-time SRP-PHAT [27–38]. One type of method is to use the idea of modifying the SRP function [27, 28] to reduce the search space of the algorithm. Although the computational cost is reduced, the final accuracy is affected by the smaller spatial resolution [29].

Another type of method is based upon an iterative strategy, such as stochastic region contraction (SRC) [30], coarse-to-fine region contraction [31], hierarchical search [32],

and vectorization [33]. Although these methods have been studied to accelerate SRP scanning, they usually limit the search to 2D surfaces [19].

In this study, we propose an iterative strategy for tracking a moving sound source; this is of interest for practical indoor robot-based human–computer interaction LBS applications [34, 35]. The proposed iterative strategy uses insights into the acoustic environment to eliminate the information gap between the building room and indoor sound source. Therefore, acoustic room geometry can be integrated into each step of the moving sound-source positioning.

The proposed improved iterative strategy for the SRP-PHAT-SRC-based method consists of first reconstructing the room geometry using the acoustic scene reconstruction method (to ensure the feasibility of the SRC) and then optimizing the initial searching region for target positioning, using the proposed Delaunay triangulation-based search volume reconstruction (DTSVR) to obtain the chosen spatial resolution. Finally, the method involves integrating the SRC with the optimized initial search volume to accelerate the SRP-PHAT.

The contributions of our proposed SRP-PHAT-SRC are summarized as follows:

First, it analyzes the ISSL problem from a geometric perspective [39].

The SRC starts from a known initial search volume (always initialized as room geometry, which we assume to be unknown). Such methods fail in the first step. Moreover, as an important part of the acoustic environment, room geometry can be used to reverse the trend of decreasing positioning accuracy as reverberation increases [40]. This demonstrates that the offline acoustic scene reconstruction does not burden the online target positioning but helps to improve its positioning performance by interacting with the room acoustics [41].

Second, it finds a compromise between the reducing complexity and ensuring accurate room geometries, using Delaunay triangulation [42].

Because the first step of SRC is to generate a set of random points in a room [30], we used the proposed DTSVR to fully exploit these random points and adaptively optimize the initial search volume of the sound-source position. Such an operation can constrain undesired interference to within a controllable range and reduce the uncertainty in the position estimation process, resulting in improved positioning accuracies.

The remainder of this paper is organized as follows. Section 2 presents two challenges in the proposed improved iterative strategy for an SRP-PHAT-SRC combined with acoustic scene reconstruction and DTSVR. Section 3 provides an overview of the proposed improved iterative strategy for the SRP-PHAT-SRC-based method. The process of applied acoustic scene reconstruction is presented in Sect. 4. Details of the DTSVR are explained in Sect. 5. Experiments and conclusions are presented in Sects. 6 and 7, respectively.

2 Challenges

In this section, the signal model of mixed speech is introduced. Following that, the acoustic features used in the experiment are introduced, and finally the dilated convolution process of the TCN is introduced.

The ISSL accuracy depends upon the choice of localization algorithm and microphone array topology, which directly affect the resolution and directivity of the array [43]. Therefore, a 3D star format array similar to that described in [44] was chosen for this research. We assume a setup in which a sound source and a 3D star format microphone array are placed in a room (the room coordinate origin is O_{room}), as shown in Fig. 1.

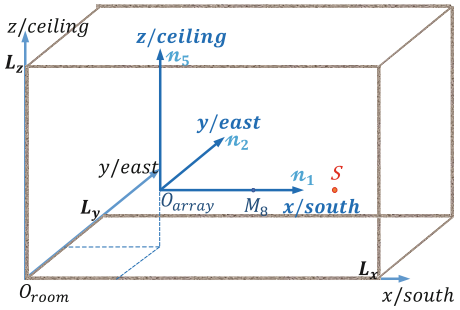

Fig. 1. Illustration of the room coordinates and room wall directions. We denote the room size as $[L_x, L_y, L_z]$, which is the unknown value to be solved.

The used array contains m microphones in a 3D grid and m microphones for each axis; the k-th microphone is denoted as M_k, $k = 1, ..., m$; O_{array} is the array origin; and Δd is the distance between adjacent microphones on each axis.

To reduce the influence of reverberation upon acoustic scene reconstruction (ASR), some operations can be easily implemented in actual experiments; for example, we assume that the array coordinate system is consistent with the room coordinate system, and that the initial position of the real sound source S lies along the x-axis.

Using the acoustic image model (AIM) [45], we define the first-order image sound source of S corresponding to the l-th wall as $\tilde{S}_l$, where $l = 1, ..., 6$ is the room wall index; the corresponding outward-pointing unit normal vectors of the wall are n_l, $l = 1, ..., 6$. The array coordinate system is consistent with the room coordinate system; hence, n_l lies along the same direction as the geophysical direction, as shown in Fig. 1; expressed otherwise, the positive x-axis is directed toward the south ($l = 1$), the positive y-axis is directed to the east ($l = 2$), the negative x-axis is directed to the north ($l = 3$), the negative y-axis is directed to the west ($l = 4$), the positive z-axis is directed to the ceiling ($l = 5$), and the negative z-axis is directed to the floor ($l = 6$).

2.1 Challenge 1: Distinguish Real First-Order Echoes from Spurious Peaks

The first challenge is to estimate the room geometry, to satisfy the feasibility requirements of the improved iterative strategy for the SRP-PHAT-SRC method. Finding the image sound sources implies confirming the surfaces in the room [46]; hence, the ASR problem can be transformed into a problem of localizing the first-order image sound sources.

Several AIM-based ASR studies have been proposed to determine the first-order echo paths [50–55]. The Euclidean distance matrix (EDM)-based first-order echo-sorting

method proposed in [53] is one of the most effective AIM-based ASR methods. This indicates that estimating the image sources is equivalent to knowing the walls; thus, we can search for image points instead of room walls.

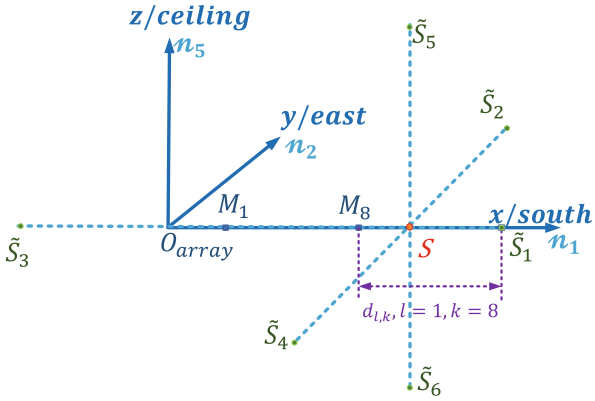

Fig. 2. Illustration of the microphone array elements on the positive x-axis and the real sound source S and its first-order image sound source $\tilde{S}_l$, where $l = 1, ..., 6$. The symbol $d_{l,k}$ denotes the distance between the l-th first-order image sound source $\tilde{S}_l$ and the k-th microphone M_k.

Figure 2 is a position diagram containing the real sound source S, the first-order image sound sources $\tilde{S}_l$ of S, and the array shown in Fig. 1; if we know the positions of the six first-order image sound sources, we can obtain the room geometry, as represented by the room size $[L_x, L_y, L_z]$, where

$$\begin{cases} L_x = \left(\left\|\tilde{S}_3 - \tilde{S}_1\right\|\right)/2 \\ L_y = \left(\left\|\tilde{S}_4 - \tilde{S}_2\right\|\right)/2 \\ L_z = \left(\left\|\tilde{S}_6 - \tilde{S}_5\right\|\right)/2 \end{cases}. \tag{1}$$

$\|\cdot\|$ denotes the Euclidean distance. According to AIM, $\tilde{S}_l$ for the l-th wall is computed as

$$\tilde{S}_l = S + 2 < S - \mathcal{p}_l, \mathcal{n}_l > \mathcal{n}_l, l = 1,2, ..., 6\ , \tag{2}$$

where $\mathcal{n}_l$is the unit normal vector, and $\mathcal{p}_l$is any point on the l-th wall.

The first-order image sound sources correspond to the first-order echoes; in practice, finding a good combination of first-order echoes is far more important than sorting correctly selected echoes [53]; therefore, the EDM-based first-order echo sorting method uses EDM as a mold, to ensure the required echo combination: if you can tightly fit a tuple of echoes in it, then the echoes must be correct.

However, room impulse responses (RIRs) contain peaks introduced by various sources of noise, and they do not correspond to any wall. Furthermore, some second-order echoes may arrive before first-order ones; the image sound sources corresponding to second- or higher-order echoes will be mis-estimated as first-order image sound

sources [56]. To avoid such mis-estimations, the EDM-based first-order echo sorting method is applied to determine the correct first-order echo combination at the cost of intensive searches.

It may be challenging to quickly and effectively extract the correct first-order echo combination from spurious peaks.

2.2 Challenge 2: Adaptive Small Search Volume for SRP-PHAT-SRC Initialization

In solving the challenge, one can only ensure the feasibility of SRP-PHAT-SRC: the efficiency problem of SRP-PHAT-SRC still remains to be solved.

The SRP-based SSL method is inherently robust when it is combined with the phase transform (PHAT) method, which is not sensitive to the surrounding environment; thus, the SRP-PHAT-based SSL method shows good robustness when combating room reverberations and noise. Because the noise term $v_k(n)$ is unrelated to $s(n)$, $\mathrm{P}_n(\overrightarrow{s})$ in SRP-PHAT can be expressed as

$$\mathrm{P}_n(\overrightarrow{s}) = \sum_{p=1}^{M} \sum_{q=p+1}^{M} \int_{-\infty}^{\infty} \psi_{pq}(\omega) X_p(\omega) X_q{}^*(\omega) e^{j\omega(\tau(\overrightarrow{s},p)-\tau(\overrightarrow{s},q))} d\omega, \tag{3}$$

where $X_p(\omega)$ and $X_q(\omega)$ are the discrete Fourier transforms of $x_p(n)$ and $x_q(n)$, respectively; superscript $*$ is a complex conjugate operator; and $\psi_{pq}(\omega) \equiv \frac{1}{|X_p(\omega)X_q{}^*(\omega)|}$ is the PHAT weighting. τ is the steered time-delay estimation of the array to the sound source, which can compensate for each delay from candidate sound-source position $\overrightarrow{s}$ directly to M_k. It can be computed using the popular generalized cross-correlation (GCC)-PHAT [57].

Therefore, we obtain the required sound-source position as follows:

$$S = arg\max_{\overrightarrow{s}} \mathrm{P}_n(\overrightarrow{s}). \tag{4}$$

SRC is a coarse-grain parallel-processing method used to compute the optimal global solution of the target function; hence, we adopted it to speed up the processing of SRP-PHAT.

However, one prerequisite for applying SRC is there must be an initial rectangular sound-source search volume posited a priori [58]; this is denoted as V_0, where

$$V_0 \leftarrow [L_x, L_y, L_z]. \tag{5}$$

The $\leftarrow$ is an assignment symbol.

In the SRC procedure, the point set used to generate the volume for the next iteration is defined as

$$N_0 = \left\{ \overrightarrow{s} \,|P_n(\overrightarrow{s}) \geq \epsilon_0, \overrightarrow{s} \in V_0 \right\}, \epsilon_0 = E(J_0), \tag{6}$$

where $E(J_0)$ is the mean power of the random points in point set J_0 within V_0. More specifically, $N_0 < J_0$. In each iteration i, $N_i < J_i$ indicates contraction. Regardless of the location of the source, the iterative process of SRP-PHAT-SRC starts from V_0.

Consider a volume (denoted as $\widehat{V}$) that adaptively changes following the movement of a sound source (i.e., a walking speaker) (thus, $\widehat{V}$ belongs to the room volume and contains the requested global optimum, i.e., $\widehat{V} \in V_0$); the judgement and decision steps in SRC will save significant time because the searching scope of the source will be reduced. Saving time helps improve the SSL efficiency. It may be challenging to determine a $\widehat{V}$ that follows an appropriate room-space triangulation.

3 System Overview

We designed an indoor sound-source continuous-positioning system under the SRP-PHAT-SRC framework. Figure 3 shows a block diagram of this system.

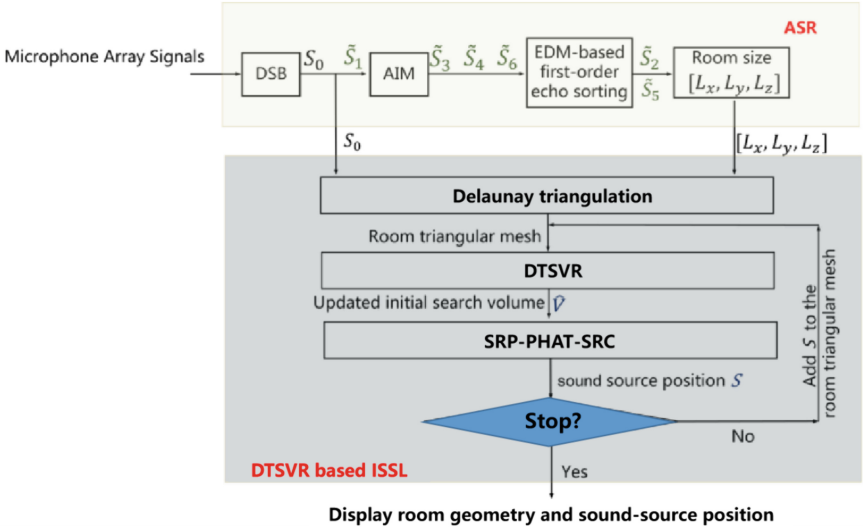

Fig. 3. Block diagram of the proposed scheme for real-time continuous positioning of an indoor sound source. ASR: Acoustic scene reconstruction. ISSL: Indoor sound-source localization. DSB: Delay and sum beamforming. AIM: Acoustic image model. EDM: Euclidean distance matrices. DTSVR: Delaunay triangulation based search volume reconstruction. SRP-PHAT-SRC: Steered response power-phase transform-stochastic region contraction.

The first block is ASR. It consists of three steps: delay and sum beamforming (DSB), acoustic image modeling (AIM), and Euclidean distance matrix (EDM)-based first-order echo sorting. First, the DSB takes the microphone array signals as inputs and outputs the sound-source initial position S_0 and its first-order image sound source $\tilde{S}_1$ pointing toward the south wall. Next, the first-order image sound sources $\tilde{S}_3$, $\tilde{S}_4$, and $\tilde{S}_6$, pointing to the north, west, and floor, respectively, are resolved via AIM analysis. Then, with the solved S_0 and the $\tilde{S}_1$, $\tilde{S}_3$, $\tilde{S}_4$, and $\tilde{S}_6$ (four of six first-order image sources of S), the EDM-based first-order echo sorting is used to search the remaining two first-order image sound sources, $\tilde{S}_2$ and $\tilde{S}_5$, which point to the east and ceiling, respectively. Finally,

taking all the positions of the first-order image sound sources, the ASR is solved using Eq. (1).

In this phase, Challenge 1 is solved and the complexity of first-order echo matching is reduced from matching six first-order image sources to matching only two first-order image sound sources ($\tilde{S}_2$ and $\tilde{S}_5$) using the EDM-based method, because the other four first-order image sound sources ($\tilde{S}_1$, $\tilde{S}_3$, $\tilde{S}_4$, and $\tilde{S}_6$) have been solved via the simple spatially symmetric structure analysis, with the help of DSB and AIM.

The second block is DTSVR-based ISSL. It accelerates the SRP-PHAT-SRC using the proposed DTSVR algorithm and addresses the real-time ISSL efficiency problem.

First, a room triangular mesh is created offline via Delaunay triangulation; this takes a set of random points within an initial volume V_0 and the sound-source initial position S_0 as inputs and outputs a room triangular mesh.

According to the actual procedure of the SRP-PHAT-SRC algorithm, during the layer-by-layer searching of the sound-source position, it is necessary to first generate a set of random points within the search volume (i.e., J_0 in V_0) during initialization. We do not change the initialization of SRP-PHAT-SRC but only use Delaunay triangulation to form these random points in set J_0 into a regularized triangular mesh; this is offline and its cost is $O(nlogn)$, where n is the point number in J_0.

Second, during online localization, provided the positioning request has not been stopped, each sound-source position estimate (taken as a point) is fed back to refine the room triangular mesh; this refining is low-cost, owing to the characteristics of the empty circumcircle.

By continuously supplementing the position points of the moving the sound source into J_0, J_0 will become increasingly large, and the volume of the regularized tetrahedrons will shrink; the smaller the regularized tetrahedron volume, the finer the room triangular mesh. With this continually optimized room mesh as the input, the proposed DTSVR algorithm (detailed in Sect. 5) outputs an adaptively smaller $\widehat{V}$ for the next moment sound source searching via SRC; therefore, the efficiency of the traditional SRP-PHAT-SRC is improved.

4 Acoustic Scene Reconstruction

In this section, we describe the application of the ASR method to room geometry. First, the array topology is used to obtain the initial position of the sound source S_0; then, the geometric symmetry of the room structure is analyzed to obtain the four first-order image sound-source positions corresponding to S_0; next, the EDM method is applied to obtain the position of the remaining two first-order image sound sources; finally, the geometric information (represented as room size) is obtained.

4.1 DSB for S_0, $\widetilde{S}_1$, $\widetilde{S}_3$, $\widetilde{S}_4$, and $\widetilde{S}_6$

In this study, we set $m = 24$; thus, the number of microphones along the x-axis is eight (i.e., from M_1 to M_8; we use these eight microphones as a linear array). Because the sound-source initial position S_0 lies along this linear array (as shown in Fig. 3, where the real sound source S is placed at S_0), we use the DSB method to enhance the

target signals (i.e., signals from S and $\tilde{S}_1$) and suppress interference and noise in the non-targeted directions.

We denote $s(n)$ as the source signal emitted by the real sound source S [i.e., $s(n)$ is the sound-source template signal]; $h_k(n)$ as the RIR, which models the channel between the real sound source S and k-th microphone; and $v_k(n)$ as a sum of acoustic multipath reflection interferences and ambient noise [unrelated to $s(n)$]. The signal received by the k-th microphone can be computed as follows:

$$x_k(n) = s(n) * h_k(n) + v_k(n),\ k = 1,2,...,p,q,...,m\ , \tag{7}$$

where $*$ denotes the linear convolution operator.

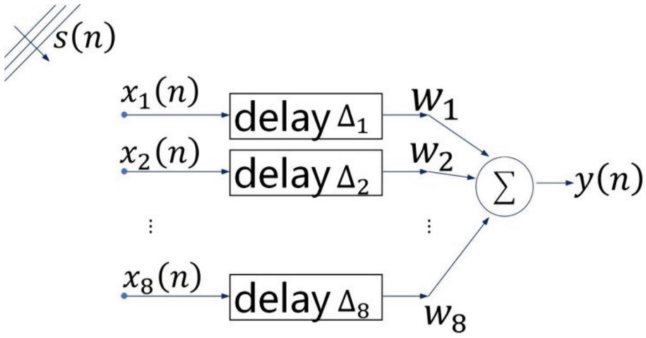

Fig. 4. Illustration of delay and sum beamforming (DSB).

As shown in Fig. 4, with the average weighting factor $w_k, k = 1, ..., 8$, the output $y(n)$ is specified as

$$y(n) = \sum_{k=1}^{8} w_k x_k(n - \Delta_k). \tag{8}$$

Δ_k is the time delay of the k-th microphone; that is, $\Delta_k = (8-k)\Delta d/c$, where c is the speed of sound propagation.

According to AIM, reflections from the walls can be replaced with direct signals from the corresponding image sound sources, the time of flight (TOF) from S_0 to M_k can be expressed as $t_{0,k}$, and the TOF from $\tilde{S}_l$ to M_k can be expressed as $t_{l,k}$; then, $t_{0,8}$is the TOF from S_0 to M_8, $t_{1,8}$is the TOF from $\tilde{S}_1$ to M_8, and the distances $d_{0,8}$ and $d_{1,8}$ are therefore

$$d_{0,8} = \|S_0 - M_8\| = ct_{0,8}/f_s\ , \tag{9}$$

$$d_{1,8} = \left\|\tilde{S}_1 - M_8\right\| = ct_{1,8}/f_s\ , \tag{10}$$

where f_s is the sample frequency, and $t_{0,8}$ and $t_{1,8}$are

$$t_{0,8} = \underset{\tau}{argmax}\left(R(\tau)\right)\ , \tag{11}$$

$$t_{1,8} = \underset{\tau}{argmax}\left(R(\tau) - max\big(R(\tau)\big)\right). \tag{12}$$

$R(\tau)$ is the GCC-PHAT between the sound-source signal $s(n)$ and output $y(n)$, which is expressed as follows:

$$R(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \psi(\omega)G(\omega)e^{j\omega\tau}d\omega, \tag{13}$$

where $\psi(\omega) \equiv \frac{1}{|S(\omega)Y^*(\omega)|}$, $G(\omega) = S(\omega)Y^*(\omega)$, and $S(\omega)$ and $Y(\omega)$ are the discrete Fourier transforms of $s(n)$ and $y(n)$, respectively. Thus, the positions of S_0 [denoted by $S_0(S_x, S_y, S_z)$], $\tilde{S}_1$ [denoted by $\tilde{S}_1(\tilde{S}_{1x}, \tilde{S}_{1y}, \tilde{S}_{1z})$], $\tilde{S}_3$ [denoted by $\tilde{S}_3(\tilde{S}_{3x}, \tilde{S}_{3y}, \tilde{S}_{3z})$], $\tilde{S}_4$ [denoted by $\tilde{S}_4(\tilde{S}_{4x}, \tilde{S}_{4y}, \tilde{S}_{4z})$], and $\tilde{S}_6$ [denoted by $\tilde{S}_6(\tilde{S}_{6x}, \tilde{S}_{6y}, \tilde{S}_{6z})$] were computed as follows:

$$\begin{cases} S_x = M_{8x} + d_{0,8} \\ S_y = M_{8y} \\ S_z = M_{8z} \end{cases}, \tag{14}$$

$$\begin{cases} \tilde{S}_{1x} = S_x + \big(d_{1,8} - \|S_0 - M_8\|\big) \\ \tilde{S}_{1y} = S_y \\ \tilde{S}_{1z} = S_z \end{cases}, \tag{15}$$

$$\begin{cases} \tilde{S}_{3x} = -S_x \\ \tilde{S}_{3y} = S_y \\ \tilde{S}_{3z} = S_z \end{cases}, \tag{16}$$

$$\begin{cases} \tilde{S}_{4x} = S_x \\ \tilde{S}_{4y} = -S_y \\ \tilde{S}_{4z} = S_z \end{cases}, \tag{17}$$

$$\begin{cases} \tilde{S}_{6x} = S_x \\ \tilde{S}_{6y} = S_y \\ \tilde{S}_{6z} = -S_z \end{cases}. \tag{18}$$

4.2 EDM-Based First-Order Echo Sorting for $\tilde{S}_2$ and $\tilde{S}_5$

We consider the microphones and sound-source setup illustrated in Fig. 3 as an example to demonstrate the application of the EDM-based first-order echo sorting method proposed in [53]. For ease of explanation and understanding, we used symbols that were consistent with the original text wherever possible.

$D \in \mathbb{R}^{8\times 8}$ is a matrix whose entries are the squared distances between microphones, $D[i, j] = \|M_i - M_j\|^2$, $1 \le i, j \le 8$. D is the EDM corresponding to the illustrated setup; it is symmetric and has zero diagonal and positive off-diagonal entries. We augment matrix D with a column vector $\mathbb{d}_l, l = 1, \ldots,$ 6to obtain matrix D_{aug}:

$$D_{aug} = \begin{bmatrix} D & \mathbb{d}_l \\ \mathbb{d}_l{}^T & 0 \end{bmatrix}. \tag{19}$$

The column vector $\mathbb{d}_l$is constructed as a combination of eight unlabeled squared distances as

$$\mathbb{d}_l = \mathbb{d}\left[d_{l,k}\right] = \begin{bmatrix} d_{l,1}^2 \\ d_{l,2}^2 \\ d_{l,3}^2 \\ d_{l,4}^2 \\ d_{l,5}^2 \\ d_{l,6}^2 \\ d_{l,7}^2 \\ d_{l,8}^2 \end{bmatrix}, l = 1, ..., 6, k.... . , 8 \ , \tag{20}$$

where $\mathbb{d}_{l,k}$ can be obtained using the Pythagorean theorem, as shown in

$$d_{l,k} = \tilde{S}_l - M_k = \begin{cases} d_{1,8} + (8-k)\Delta d, k = 1, ..., 7, l = 1 \\ \sqrt{S_0 - M_k^2 + S_0 - \tilde{S}_2^2}, k = 1, ..., 8, l = 2 \\ \left|\tilde{S}_{3x}\right| + M_{kx}, k = 1, ..., 8, l = 3 \\ \sqrt{S_0 - M_k^2 + S_0 - \tilde{S}_4^2}, k = 1, ..., 8, l = 4 \\ \sqrt{S_0 - M_k^2 + S_0 - \tilde{S}_5^2}, k = 1, ..., 8, l = 5 \\ \sqrt{S_0 - M_k^2 + S_0 - \tilde{S}_6^2}, k = 1, ..., 8, l = 6 \end{cases} . \tag{21}$$

Once the $\mathbb{d}_2$and $\mathbb{d}_5$are confirmed, the position of $\tilde{S}_2$ [denoted as $\tilde{S}_2(\tilde{S}_{2x}, \tilde{S}_{2y}, \tilde{S}_{2z})$], and $\tilde{S}_5$ [denoted as $\tilde{S}_5(\tilde{S}_{5x}, \tilde{S}_{5y}, \tilde{S}_{5z})$] can be solved by

$$\begin{cases} \tilde{S}_{2x} = S_x \\ \tilde{S}_{2y} = S_y + \sqrt{d_{2,k}^2 - \|S - M_k\|^2} \\ \tilde{S}_{2z} = S_z \end{cases}, \tag{22}$$

$$\begin{cases} \tilde{S}_{5x} = S_x \\ \tilde{S}_{5y} = S_y \\ \tilde{S}_{5z} = S_z + \sqrt{d_{5,k}^2 - \|S - M_k\|^2} \end{cases} . \tag{23}$$

Thus, it can be seen that D_{aug} is interpreted as an object encoding a particular selection of echoes $\mathbb{d}$. The selected echo combination only corresponds to the same image source when $rank(D_{aug}) < 6$ holds or, more specifically, when D_{aug} is an EDM. Finally, the room geometry is solved using

$$\begin{cases} L_x = S_x + (\|S_0 - \tilde{S}_1\|)/2 \\ L_y = S_y + (\|S_0 - \tilde{S}_2\|)/2 \\ L_z = S_z + (\|S_0 - \tilde{S}_5\|)/2 \end{cases}, \tag{24}$$

where $\|\cdot\|$ denotes the Euclidean distance operator. It should be noted that the meaning expressed by Eq. (24) is consistent with that expressed by Eq. (1), albeit in a different form.

The advantages of our ASR method are as follows:

1) Owing to the advantages of the star array topology, it dose not need to test numerous echoes via the EDM mold for the image sound source $\tilde{S}_1$; instead, we use simple equations [Eq. (15)].
2) In the room geometric symmetry structure analysis, we do not need to test numerous echoes for the image sound sources $\tilde{S}_3$, $\tilde{S}_4$, and $\tilde{S}_6$, because they can be calculated using the symmetrical structure [via Eqs. (16–18)].
3) We test the EDM mold upon the image sound sources $\tilde{S}_2$ and $\tilde{S}_5$. However, the aforementioned advantages reduce the testing number from six unknown first-order image sound sources to two. For echo matching, this markedly reduces the computation time for ASR.

The simulation verification of this acoustic reconstruction scheme has been published in our open paper [54]; it fully proves the theoretical feasibility of the model. However, the simulation and verification of the complexity of the sound field environment can only be accomplished by adjusting the reverberation time (RT_{60}) and signal-to-noise ratio (SNR). In fact, RT_{60} and SNR differ across real sound fields, and unpredictable mutual signal interferences arise. Therefore, this study fully verifies the validity and reliability of ASR scheme in a real room.

5 Delaunay Triangulation-Based Search Volume Reconstruction Algorithm

In this section, we detail the proposed DTSVR, to show how knowledge of the 3D room geometry information can improve the efficiency of SRP-PHAT-SRC.

5.1 Delaunay Triangulation

Delaunay triangulation produces a Euclidean graph that contains only a linear number of edges [45]. The graph"s convex polygonal shell is the basic structure of computational geometry and represents a useful tool for constructing other geometric shapes; hence, it is widely applied in practical problems such as 3D surface reconstruction.

Triangulation involves creating a set of non-overlapping, triangularly bounded facets from sample points; the vertices of the triangles are the input sample points.

For example, in our task, we first create a set of sample points within the search volume (i.e., J_0 in V_0, where $V_0 \leftarrow [L_x, L_y, L_z]$) and then apply the Delaunay triangulation method to generate the room triangular mesh, as shown in Fig. 5(a).

To add a new point [e.g., the green point in Fig. 5(b)] to the Delaunay triangulated mesh, we need only to delete all triangles containing this point and then connect it with all other existing points such that it does not intersect the other edges after connection. As a result, the Delaunay triangulation algorithm is simple and has a low time complexity; new points can be easily added to the pre-existing room triangular mesh at any time, without the need to re-triangulate all sample points. This is very useful for the automatic refinement of the room mesh because we can import each historical location of the moving sound source into the room mesh; from this, the room mesh can produce a more accurate estimate of the subsequent time position of the sound source.

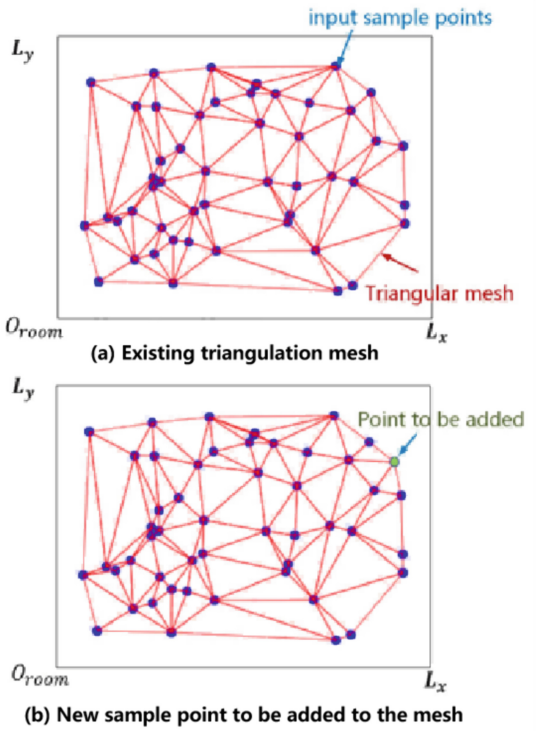

Fig. 5. Top view of the Delaunay triangulation on the *xoy* plane.

5.2 Proposed DTSVR Algorithm

Initially, S is placed at S_0, and the next moment position of S is very likely to be within a circle with the initial position S_0 as its center and step length r_0 as its radius, as shown by the green circles in Fig. 6. Naturally, the volume of the three-dimensional sphere of this circle can be regarded as the desired small search volume of the sound-source position at the next moment (i.e., $\widehat{V} \leftarrow 4\pi {r_0}^3/3$).

However, this assumption does not always hold in complex indoor acoustic environments, and its influence upon the cumulative error of the ISSL cannot be neglected. When the position of the sound source changes with time, the error transmission causes the center of the sphere to deviate from the current correct position of the sound source; hence, the sphere volume does not contain the subsequent moment position of the sound source.

To avoid ISSL failures attributable to error transmission, we adopt Delaunay triangulation to fully utilize the initial position and room geometry of the sound source when generating a room triangular mesh; then, the proposed DTSVR uses this mesh to generate an optimized and adaptive volume $\widehat{V}$.

The proposed DTSVR for identifying the desired $\widehat{V}$ is as follows:

1) Initialize pedestrian"s step index: $i = 0$.

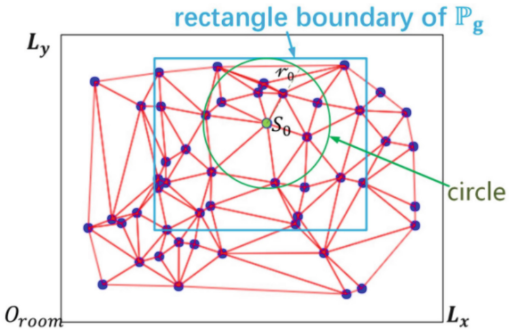

Fig. 6. Top view of the room mesh. The green circle is that mentioned in Step 3 of the proposed DTSVR for $i = 0$; the blue rectangle is that mentioned in Step 6 of the DTSVR.

2) Set the initial parameters RM, S_i, and r_i. RM is the room triangular mesh from the Delaunay triangulation module, S_i is the current position of S [S_0 can be computed from Eq. (14)], $r_i$is the step length of S and is set as a constant r(i.e., $r_i = r$, meaning that the pedestrian walks with a fixed step length).
3) Generate a circle with S_i as its center and $r_i$as its radius.
4) If $i \neq 0, RM \leftarrow RM + S_i$; otherwise, $RM \leftarrow RM$. Using this RM, identify all sample points within the three-dimensional sphere that correspond to the circle formed in Step 3; denote this point set as $\mathbb{P}_s$.
5) Identify all the edge points of $\mathbb{P}_s$ related to the mesh; denote the edge point set as $\mathbb{P}_g$.
6) Assign the volume of the three-dimensional cuboid (comprising the rectangle boundary of $\mathbb{P}_g$) to $\widehat{V}$ [i.e., $\widehat{V}_i \leftarrow (x_{max}(\mathbb{P}_g) - x_{min}(\mathbb{P}_g)) \times (y_{max}(\mathbb{P}_g) - y_{min}(\mathbb{P}_g)) \times (z_{max}(\mathbb{P}_g) - z_{min}(\mathbb{P}_g))$].
7) $i = i + 1$, go to Step 3 until the request stops.

It should be emphasized that when $i = 0$, $S_i = S_0$; when $i \neq 0$, the S_i is assigned according to the value outputted by the SRP-PHAT-SRC module, which takes the $\widehat{V}_i$ derived in Step 6 as its input.

Taking the speaker"s current position estimate into account and expanding $\widehat{V}$ from a three-dimensional sphere to a room mesh-based search space can make $\widehat{V}$ more robust against error transmissions. It is understood from Delaunay triangulation that when the speaker walks more, more position point information can be fed back to the room mesh; this refines the room mesh and improves the reliability of the generated $\widehat{V}$. Moreover, because $\widehat{V}$ is always smaller than V_0, $\widehat{V}$ can improve the SRP-PHAT-SRC localization efficiency by alleviating the SRC search task.

To calculate $\widehat{V}$ whilst ensuring the accuracy of sound-source location estimation, we developed the DTSVR algorithm to accelerate sound-source localization (we applied for Chinese patent protection for this scheme in 2018). The scheme was made public in 2018 and officially authorized in April 2022 (ZL 201810560142.1). Since publication, the scheme has been recognized by other teams. For example, [55] proposed a spherical structure to reduce the spatial search range of the sound source and accelerate the positioning process following a similar research scheme.

6 Experiments

We conducted three experiments to evaluate the performance of the proposed system and methods: A) an experiment conducted in a real room, to examine the efficacy of the array topology analysis-based ASR; B) an experiment using a reconstructed room, to evaluate the feasibility of the Delaunay triangulation; and C) an robustness evaluation of the proposed DTSVR-based SRP-PHAT-SRC.

To evaluate the performance of our approach in the presence of actual obstacles, we collected data across different days and times of day over a period of three months. During collection, the students and staff walked around normally. We collected data from rooms in the No.7 Teaching Building at the Jin Ji Campus of Guilin University of Electronic Technology (GUET), Guilin, Guangxi Zhuang Autonomous Region, China. Here, we consider Room 7317 (shown in Fig. 7) as an example, to demonstrate the performance of our proposed scheme.

Fig. 7. Illustration of a small office (Room 7317) in GUET. The office room contains six sets of office desks and chairs, a sofa, a conference table, and multiple lockers. The east side of Room 7317 is a wall with glass windows, the south and north sides are plain walls, the door is on the west side of the room, the ceiling is covered with plaster, and the floor is covered with ordinary tiling. The sound source is an indoor pedestrian (talker) and the sound receiver is in a star format array (on the conference table).

To simplify measurement of the sound-source template signal $s(n)$ and eliminate the ISSL errors introduced via the $s(n)$ measurement, we used a Samsung Note 5 smartphone (carried by a student) as a real sound source; it repeatedly played a 3 s piece of female speech audio (recorded in an anechoic chamber) as $s(n)$. The purpose of this setting relates to actual LBS applications [59]. The $s(n)$ signal can be obtained through head-mounted devices; however, this fall under the field of speech recognition and enhancement, which lies outside the scope of the current study.

The sound receiver used in our experiment was a star-format microphone array, as shown on the right-hand side of Fig. 7. The received signal $x_k(n)$ was obtained through an NI PXIe-4497 data acquisition module and the Single Pad audio data processing software available from KeyGoTech Inc., Shanghai, China.

The true values of the relevant parameters (for error analysis only) were set as follows: initial position (S_0): [5,3.5,1.6](m); true room size: [6.9740, 7.3520, 3.2950](m); position of the O_{array}: [3,3.5,1.6](m); and Δd : 0.1 m.

6.1 Acoustic Scene Reconstruction Results

1) Results of S_0 and its first-order image sound sources

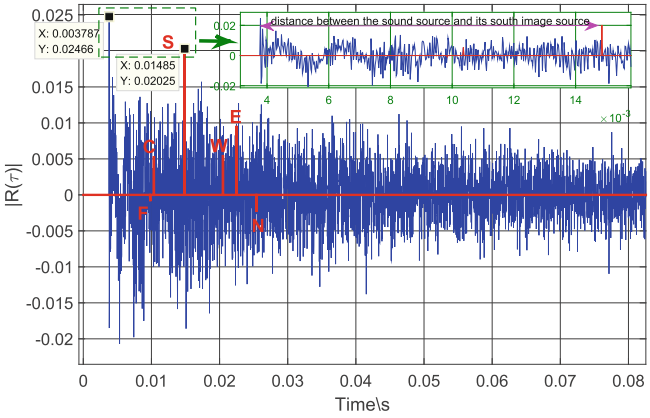

Fig. 8. Illustration of the GCC-PHAT between $s(n)$ and $y(n)$ as well as the first-order echo sorting results. F: Floor. C: Ceiling. S: South. W: West. E: East. N: North. The upper right corner in the figure shows the enlarged waveform between the first two highest peaks in the GCC-PHAT.

First, through the DSB analysis [Eqs. (8–13)], we get the first- and second-highest peaks of GCC-PHAT correspond to the direct paths from S_0 to M_8 and from $\widetilde{S}_1$ to M_8, respectively. These two peaks are indicated by the dotted green rectangle in Fig. 8.

Thus, the TOFs corresponding to the two peaks can be obtained from the figure as $t_{0,8} = 0.003787(\text{s})$ and $t_{1,8} = 0.01485(\text{s})$. Setting $c = 340m/s$, we obtained the position estimates of S_0 and $\widetilde{S}_1$ as $S_0 = [5.0875,3.5,1.6](\text{m})$ and $\widetilde{S}_1 = [8.849,3.5,1.6](\text{m})$.

Second, from Eqs. (16)–(18), we obtained position estimates for the first-order image sound sources corresponding to the north and west walls and the floor, respectively, as follows: $\tilde{S}_3 = [-5.0875,3.5,1.6](\text{m})$, $\tilde{S}_4 = [5.0875, -3.5,1.6](\text{m})$, and $\tilde{S}_6 = [5.0875,3.5, -1.6](\text{m})$.

Third, from the above data and Eqs. (19)–(21), we obtained $d_{2,8} = 7.8748(\text{m})$ and $d_{5,8} = 3.6823(\text{m})$; thus, the corresponding positions of the first-order image sound sources were as follows: $\tilde{S}_2 = [5.0875,11.2689,1.6](\text{m})$ and $\tilde{S}_5 = [5.0875,3.5,5.0498](\text{m})$.

2) Room geometry information estimation

The ASR effect is shown in Fig. 9, where the reconstructed room is seen to be consistent with the real one, the estimated one is [6.9683,7.3844,3.3249](m), the corresponding error is [0.0057,0.0324,0.0299](m).

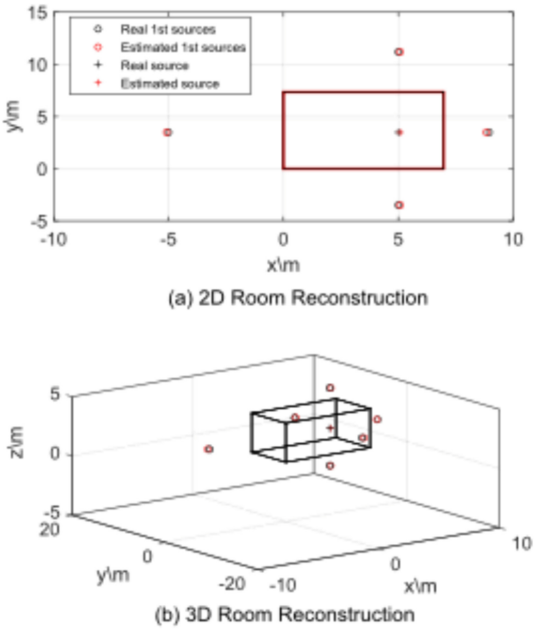

Fig. 9. ASR effects: (a) 2D form of the *xoy* plane for the reconstructed room, showing the differences between the real room geometry (marked in black) and the estimated one (marked in red). (b) Similar differences in 3D form.

Table 1. Average errors of room size

Error (m)	EDM-based ASR method proposed in [53]	Proposed array topology analysis-based ASR method
Horiz	0.05	0.0057
Vert	0.05	0.0324
Height	0.02	0.0299

Table 1 compares the reconstruction errors between the proposed method and that reported in [53]. Our horizontal reconstruction takes advantage of the array topology to produce a horizontal reconstruction result significantly better than that derived from EDM analysis. Because our vertical reconstruction method essentially comes from the EDM-based ASR method, its reconstruction effect is not significantly different from the EDM one. The difference is in the reconstruction time complexity, where we reduce the number of first-order image sound sources to be matched in the EDM mold and thereby simplify the EDM-based ASR method.

6.2 Results of Delaunay Triangulation

Figure 10 illustrates Delaunay triangulation. In our experiment, the input parameters for the Delaunay triangulation algorithm are $J_0 = 1000$ and $V_0 \leftarrow$ [6.9683,7.3844,3.3249](m). Figure 10(a) shows a set of random points J_0 (blue dots inside the cuboid) inside the reconstructed room, eight vertices (marked in blue) of the reconstructed room, and S_0 (marked in red). Figure 10(b) shows the room triangular mesh (i.e., the graphical representation of parameter RM) generated via Delaunay triangulation of the results shown in Fig. 10(a). RM is a $4 \times$ N matrix, where N denotes the number of regularized tetrahedrons.

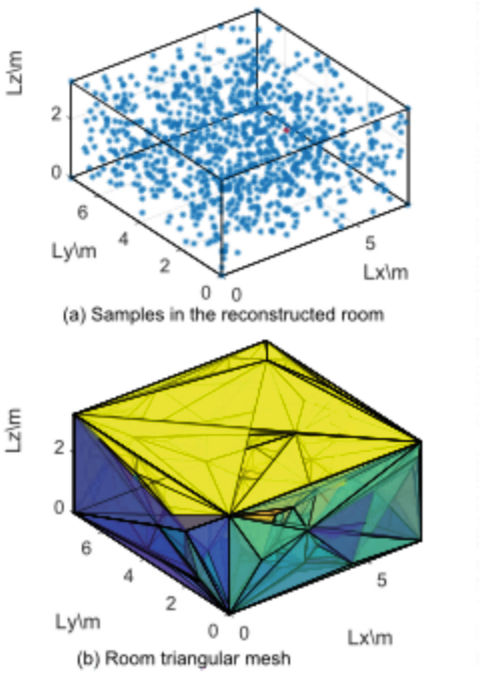

Fig. 10. Delaunay triangulation. Random point set J_0 is generated using the MATLAB random function, and the room triangular mesh can be obtained using the MATLAB Delaunay function. Both functions are built-in and have a lightweight arithmetic level.

6.3 Results and Analysis of the Proposed DTSVR

As discussed earlier, the SSL error leads to errors of the sound-source position; thus, the sphere volume centered upon the estimated value does not cover the next moment"s sound-source position, as shown in Fig. 11.

When time passes from time t to time $t+1$, the real sound source moves from the red plus sign in Fig. 11(a) to that in Fig. 11(b). Using the estimated value of the sound-source

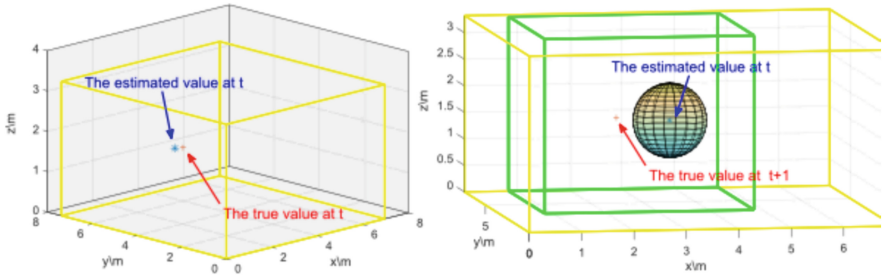

Fig. 11. Illustration of the proposed DTSVR: (a) The estimated sound-source position (blue star) and true sound-source position value (red plus) in the reconstructed room (yellow cuboid) at a previous time (e.g., time t). (b) The $\widehat{V}$ (green cuboid) generated by the proposed DTSVR, the circle-based three-dimensional volume (black sphere,$r = 1.0m$), the estimated sound-source position (blue star) at time t, and the true sound-source position (red plus) at time $t + 1$.

position at time t as the center of the sphere, and taking r as the radius, we can see from Fig. 11(b) that the sphere does not cover the true sound-source position at time $t + 1$.

Therefore, it is inappropriate to use the sphere volume as the search volume for the next moments sound-source position.

One could increase the step length to avoid such phenomena; however, the step length used in the result shown in Fig. 11 is $r = 1.0\ m$. Generally speaking, $r = 1.0\ m$is the largest step size for humans walking normally, according to experimental results regarding human preferred walking speeds [60]; thus, increasing the step size (to expand the sphere volume for the sound-source searching) does not fundamentally improve the efficiency of the SRP-PHAT-SRC-based SSL algorithm.

The phenomenon can also be avoided by using the adaptive search volume generated by the proposed DTSVR algorithm. More details are shown in the 24 sub-graphs in Fig. 12. These subgraphs show the real-time continuous positioning results of a moving sound source from Position 2 to Position 25 (Position $1 = S_0$) in Room 7317, according to the proposed scheme. The generated search volume $\widehat{V}$ can adaptively and reliably follow a moving sound source.

We denote the reduction in search volume in Step i as ΔV_i; thus, it can be computed as

$$\Delta V_i = V_0 - \widehat{V}_i. \tag{25}$$

Hence, the search volume reduction at Step i is $\Delta V_i / V_0$. Figure 13 shows that the proposed DTSVR achieves a mean search volume reduction of 69.81%, which confirms that the operation of the improved SRP-PHAT-SRC algorithm implementing our proposed scheme is greatly reduced, thereby ensuring a superior real-time performance. The operation used here is equivalent to the functional evaluation (FE) determined in [30], which reflects the calculation of any particular point of $\mathrm{P}(\overrightarrow{s})$.

Figure 13(a) shows this reduction at every step (i.e., Positions 2–25). To understand the distribution of these reductions statistically, we show all $\Delta V_i / V_0$ values in a histogram, as shown in Fig. 13(b).

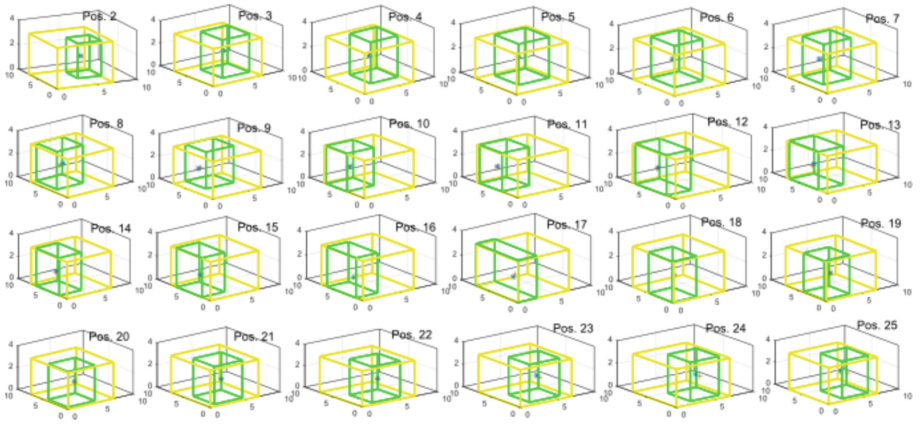

Fig. 12. Indoor continuous pedestrian positioning effect based on the proposed scheme. The reconstructed room is marked by a yellow cuboid, the estimated sound-source position value at the current position is marked with a blue star, the true sound-source position at the current position is marked with a red plus, and the green cuboid is the desired $\widehat{V}$ generated by the DTSVR using the sound-source position estimated from the previous moment. Pos.: Position.

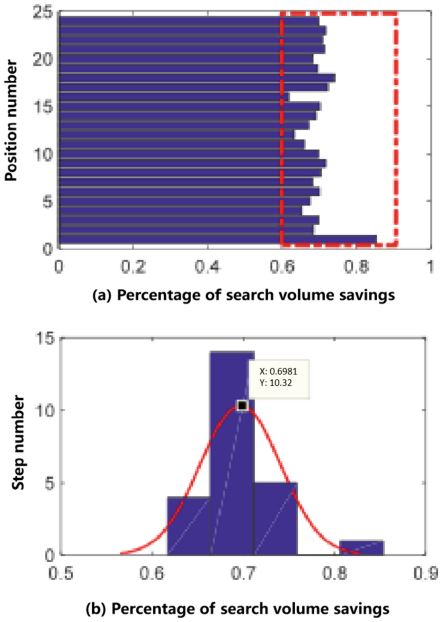

Fig. 13. Search volume saving: (a) The distribution of the reduction in search volume and (b) the reduction in search volume at every position.

We evaluated the effects of ASR and Delaunay triangulation upon SRP-PHAT-SRC, by comparing the localization error and operations of SRP-PHAT-SRC using the room space V_0 and the proposed adaptive $\widehat{V}$ as the initial search volume.

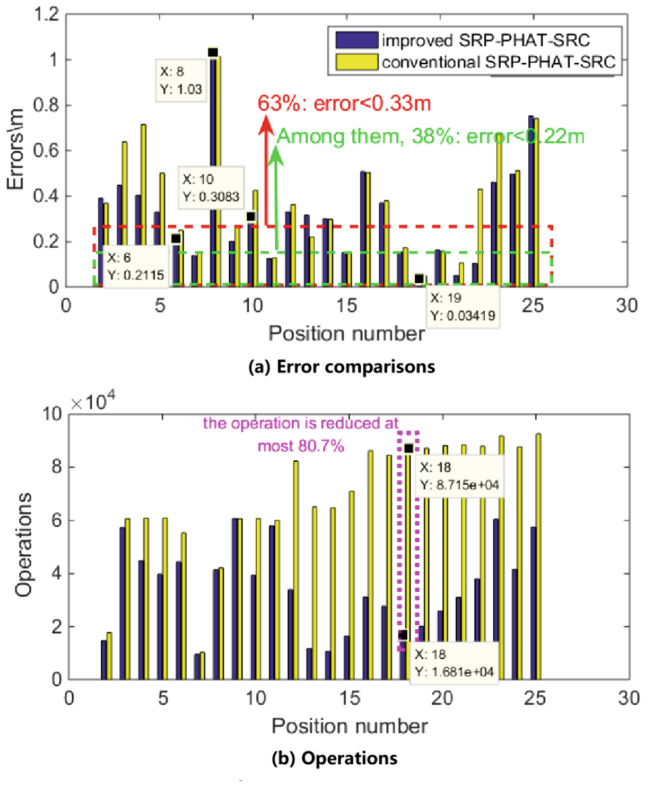

Fig. 14. Illustration of the positioning accuracy and efficiency: (a) The localization error comparison between the conventional SRP-PHAT-SRC (yellow block) and that (blue block) based on our proposed scheme. (b) Operation comparison between the two SSL algorithms.

As shown in Fig. 14(a), 17 of the 24 positions exhibited error reductions; thus, the accuracy of SSL is improved by 71% when using the adaptive $\widehat{V}$ as the initial search volume. Moreover, 63% of the position errors were 0.33 m or less; of these, 38% were 0.22 m or less. These results confirmed the accuracy of the improved SRP-PHAT-SRC. Meanwhile, the operation results shown in Fig. 13(b) indicate that using adaptive $\widehat{V}$ as the search volume can reduce the time consumption by as much as 80.7%, thereby ensuring the efficiency of the improved SRP-PHAT-SRC. These results demonstrate the efficacy of our proposed scheme for the real-time continuous positioning of an indoor pedestrian.

7 Conclusion

In this study, we used ASR and Delaunay triangulation to obtain indoor spatial knowledge and thereby generated the spatial correlations of the moving source position. Further, we used the spatial correlation to construct an adaptive search space for moving sound sources and to establish their motions and the temporal correlation of their locations. This transformed a large-scale search (i.e., for a large number of candidate sound sources) into an iterative search and correction procedure within an adaptive space.

In future, we will apply this method to robot-based sound signal analysis to realize smarter human–machine services. For example, we hope to enhance the clarity and volume of a specific speaker during voice calls for maintaining a good listening experience regardless of the motion of the speaker.

Acknowledgements. This work was supported by the Ministry of Education, Key Laboratory of Cognitive Radio, and Information Processing.

This work was funded by the National Natural Science Foundation of China:62071135, the Project (CRKL200111 and CRKL210110) from Key Laboratory of Cognitive Radio and Information Processing, Ministry of Education (Guilin University of Electronic Technology).

References

1. Anguera, X., Wooters, C., Hernando, J.: Acoustic beamforming for speaker diarization of meetings. IEEE Trans. Audio Speech Lang. Process. **15**(7), 2011–2022 (2007)
2. Ward, D.B., Lehmann, E.A., Williamson, R.C.: Particle filtering algorithms for tracking an acoustic source in a reverberant environment. IEEE Trans. Speech Audio Process. **11**(6), 826–836 (2003)
3. Wang, L., Roggen, D.: Sound-based transportation mode recognition with smartphones. In: *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, pp. 930–934. (2019)
4. Zhuang, Y., Cao, Y.: Guest Editorial: special issue on toward positioning, navigation, and location-based services (PNLBS) for internet of things. IEEE Int. Things J. **5**(6), 4613–4615 (2018)
5. Yu, Z., Tian, M., Wang, Z., Bin, G., Tao, M.: Shop-type recommendation leveraging the data from social media and location-based services. ACM Trans. Knowl. Discov. Data **11**(1), 1 (2016)
6. Luo, R.C., Lai, C.C.: Enriched indoor map construction based on multisensory fusion approach for intelligent service robot. IEEE Trans. Ind. Electron. **59**(8), 3135–3145 (2012)
7. Havangi, R., Taghirad, H.D., Nekoui, M.A., Mohammad, T.: A square root unscented fast slam with improved proposal distribution and resampling. IEEE Trans. Ind. Electron. **61**(5), 2334–2345 (2014)
8. Lee, S.M., Jung, J., Kim, S., Kim, I.J., Hyun, M.: DV-SLAM (dual-sensor-based vector-field slam) and observability analysis. IEEE Trans. Ind. Electron. **62**(2), 1101–1112 (2015)
9. Hwang, S.Y., Song, J.B.: Monocular vision-based slam in indoor environment using corner, lamp, and door features from upward-looking camera. IEEE Trans. Ind. Electron. **58**(10), 4804–4812 (2011)
10. Herranz, F., Llamazares, A., Molinos, E., Ocana, M., Sotelo, M.A.: Wi-Fi SLAM algorithms: an experimental comparison. Robotica **34**(04), 837–858 (2016)

11. Djugash, J., Singh, S., Kantor, G., Zhang, W.: Range-only SLAM for robots operating cooperatively with sensor networks. *IEEE Int. Conf. Robotics, Automation,* Orlando, Florida. p. 2078–2084 (2006)
12. Luo, R.C., Lai, C.C.: Multisensory fusion-based concurrent environment mapping and moving object detection for intelligent service robotics. IEEE Trans. Ind. Electron. **61**(8), 4043–4051 (2014)
13. Yuan, C., Yicheng, H, Mingsian, R.B.: Multi-channel end-to-end neural network for speech enhancement, source localization, and voice activity detection. ARXIV, (2022)
14. Yijun, G., Shupei, L., Xiao-Lei, Z.: End-to-end two-dimensional sound source localization with ad-hoc microphone arrays. ARXIV (2022)
15. Qing, W., et al.: Deep learning based audio-visual multi-speaker DOA estimation using permutation-free loss function. ARXIV (2022)
16. Annibale, P., Antonacci, F., Bestagini, P., et al.: The SCENIC project: environment-aware sound sensing and rendering. Proc. Comput. Sci. **7**(29), 150–152 (2011)
17. Astapov, S., Riid, A.: A hierarchical algorithm for moving vehicle identification based on acoustic noise analysis," *Proceedings of the 19th International Conference Mixed Design of Integrated Circuits and Systems - MIXDES*, Warsaw, pp. 467–472, May. 2012
18. Astapov,S., Preden, J., Berdnikova, J.: Simplified acoustic localization by linear arrays for wireless sensor networks. In: *2013 18th International Conference on Digital Signal Processing (DSP)*, Fira, Greece, pp. 1–6, Jul. 2013
19. Grondin, F., Glass, J.: SVD-PHAT: a fast sound source localization method. In: *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, 2019, pp. 4140–4144, https://doi.org/10.1109/ICASSP.2019.8683253
20. Salvati, D., Drioli, C., Foresti, G.L.: A low complexity robust beamforming using diagonal unloading for acoustic source localization. IEEE/ACM Trans. Audio Speech Lang. Proc. **1**(1), 99 (2018)
21. Padois, T.: Acoustic source localization based on the generalized cross-correlation and the generalized mean with few microphones. J. Acoust. Soc. Am. **143**(5), 393–398 (2018)
22. Laufer-Goldshtein, B., Talmon, R., Gannot, S.: A hybrid approach for speaker tracking based on TDOA and data-driven models. IEEE/ACM Trans. Audio Speech Lang. Process. **1**(1), 1 (2018)
23. Karabiyik, Y., Avdal, J., Ekroll, I.K., Fiorentini, S., Torp, H., Løvstakken, L.: Data-adaptive 2-d tracking doppler for high-resolution spectral estimation. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **67**(1), 3–12 (2020)
24. Wu, Y., Leshem, A., Jensen, J.R., Liao, G.S.: Joint pitch and DOA estimation using the ESPRIT method. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(1), 32–45 (2017)
25. Zotkin, D.N., Duraiswami, R.: Accelerated speech source localization via a hierarchical search of steered response power. IEEE Trans. Speech Audio Process. **12**(5), 499–508 (2004)
26. Dibiase, J.H., Silverman, H.F., Brandstein, M.S.: Robust localization in reverberant rooms. Microph. Arr. Signal Process. Techn. Appl. **2001**, 157–180 (2001)
27. Cobos, M., Marti, A., Lopez, J.J.: A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling. IEEE Signal Process. Lett. **18**(1), 71–74 (2011)
28. Lima, M.V.S., et al.: A volumetric SRP with refinement step for sound source localization. IEEE Signal Process. Lett. **22**(8), 1098–1102 (2015)
29. Marti, A., Cobos, M., Lopez, J.J., et al.: A steered response power iterative method for high-accuracy acoustic source localization. J. Acoust. Soc. Am. **134**(4), 2627 (2013)
30. Do, H., Silverman, H.F., Yu, Y.: A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array. *IEEE Int. Conf. Acoust., Speech Signal Process.*, Honolulu, HI, USA, Apr. 2007, pp. 121–124

31. H. Do, H. F. Silverman, "A Fast Microphone Array SRP-PHAT Source Location Implementation using Coarse-To-Fine Region Contraction (CFRC), In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, Oct. 2007
32. Grondin, F., Michaud, F.: Lightweight and optimized sound source localization and tracking methods for open and closed microphone array configurations. Robot. Auton. Syst. **113**, 63–80 (2019)
33. B. Lee and T. Kalker, A Vectorized Method for Computationally Efficient SRP-PHAT Sound Source Localization. In: *Int. Workshop on Acoustic Echo & Noise Control*, 2010
34. Salvati, D., Drioli, C., Foresti, G.L.: Exploiting a geometrically sampled grid in the steered response power algorithm for localization improvement. J. Acoust. Soc. Am. **141**(1), 586–601 (2017)
35. Salvati, D., Drioli, C., Foresti, G.L.: Sensitivity-based region selection in the steered response power algorithm. Signal Process. **153**, 1–10 (2018)
36. Dmochowski, J.P., Benesty, J., Affes, S.: A generalized steered response power method for computationally viable source localization. IEEE Trans. Speech Audio Proc. **15**(8), 2510–2526 (2007)
37. Do,H., Silverman, H.F.: Stochastic particle filtering: a fast SRP-PHAT single source localization algorithm. In: IEEE Workshop on Applications of Signal Processing to Audio & Acoustics. New Paltz, NY, USA, Oct. 2009
38. Traa, J., Wingate, D., Stein, N.D., Smaragdis, P.: Robust source localization and enhancement with a probabilistic steered response power model. IEEE/ACM Trans. Audio, Speech, Lang. Proc. **24**(3), 493–503 (2016)
39. Antonacci, F., Sarti, A., Tubaro, S.: Geometric reconstruction of the environment from its response to multiple acoustic emissions. *IEEE Int. Conf. Acoust., Speech Signal Process.*, Dallas, TX, USA, Mar. 2010, p. 2822 – 2825.
40. Cho, Y., Yook, D., Chang, S., Kim, H.: Sound source localization for robot auditory systems. IEEE Trans. Consumer Elect. **55**(3), 1663–1668 (2009)
41. Ribeiro, F., Zhang, C., Florêncio, D.A., Ba, D.E.: Using reverberation to improve range and elevation discrimination for small array sound source localization. IEEE Trans. Audio Speech Lang. Proc. **18**(7), 1781–1792 (2010)
42. Dokmanić, I., Scheibler, R., Vetterli, M.: Raking the cocktail party. IEEE J. Select. Top. Signal Proc. **9**(5), 825–836 (2015). https://doi.org/10.1109/JSTSP.2015.2415761
43. J. M. Keil, C. A. Gutwin, "The Delaunay Triangulation Closely Approximates the Complete Euclidean Graph," *Workshop on Algorithms & Data Structures*, Ottawa, Canada, Jan. 1989
44. D. V. Rabinkin, "Optimum sensor placement for microphone arrays," *Dissertations & Theses*, Jan. 1998
45. S. Tervo, T. Korhonen, Estimation of reflective surfaces from continuous signals. *IEEE Int. Conf. Acoust., Speech Signal Process*, Dallas, TX, USA, pp. 153–156 (2010)
46. Allen, J.B., Berkley, D.A.: Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. **65**(S1), 943–950 (1978)
47. Jager, I., Heusdens, R., Gaubitch, N.D.: Room geometry estimation from acoustic echoes using graph-based echo labeling. In: *IEEE Int. Conf. Acoust., Speech Signal Process.* Shanghai, China, pp. 1–5 (2016)
48. I. Dokmani, Y. M. Lu, M, Vetterli, Can one hear the shape of a room: The 2-D polygonal case. IEEE International Conference on Acoustics, Speech, and Signal Processing Prague, Czech. p. 321–324 (2011)
49. Antonacci, F., Sarti, A., Tubaro, S.: Geometric reconstruction of the environment from its response to multiple acoustic emissions. In: *IEEE Int. Conf. Acoust., Speech Signal Process.* Dallas, TX, USA, Mar. 2010, p. 2822–2825
50. Antonacci, F., et al.: Inference of room geometry from acoustic impulse responses. Audio, Speech, Language Process. **20**(10), 2683–2695 (2012)

51. Markovica, D., Antonacci, F., Sarti, A., Tubaro, S.: Estimation of room dimensions from a single impulse response. In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, Oct. 2013, pp. 1–4
52. Tervo, S., Tossavainen, T.: 3D room geometry estimation from measured impulse responses. *IEEE Int. Conf. Acoust., Speech Signal Process. Kyoto*, Japan p: 513–516 (2012)
53. Kelly, I.J., Boland, F.M.: Detecting arrivals in room impulse responses with dynamic time warping. Lang. Proc. **22**(7), 1139–1147 (2014)
54. Dokmanic, I., Parhizkar, R., Walther, A., Lu, Y.M. Vetterli, M.: Acoustic echoes reveal room shape. In: *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, np. 30, pp. 12186–12191, May. 2013
55. Song, X.Y., Wang, M., Qiu, H.: Room geometry reconstruction based on speech and acoustic image methodology. IEEE International Conference on Smart Internet of Things (SmartIoT), 2019, pp. 113–120
56. Liu, D., Cai, X., Yu, D., Qiao, Z., Dong, H., Wu, M.: Sound source localization methods based on lagrange-galerkin spherical grid. IEEE Int. Conf. Electr. Eng. Mechatron. Technol. (ICEEMT) **2021**, 665–670 (2021). https://doi.org/10.1109/ICEEMT52412.2021.9602760
57. Krekovic, M., Dokmanic, I., Vetterli, M.: EchoSLAM: simultaneous localization and mapping with acoustic echoes. *IEEE Int. Conf. Acoust., Speech Signal Process.* Shanghai, China, Mar. pp 11–15 (2016)
58. Knapp, C.H., Carter, G.C.: The generalized correlation method for estimation of time delay. Audio, Speech, Lang. Process. **24**(4), 320–327 (1976)
59. Berger, M.F., Silverman, H.F.: Microphone array optimization by stochastic region contraction. IEEE Trans. Signal Proc. **39**(11), 2377–2386 (1991). https://doi.org/10.1109/78.97993
60. Kumatani, K., McDonough, J., Raj, B.: Microphone array processing for distant speech recognition: from close-talking microphones to far-field sensors. IEEE Signal Proc. Magaz. **29**(6), 127–140 (2012)
61. Norris, J.A., Granata, K.P., Mitros, M.R., Byrne, E.M., Marsh, A.P.: Effect of augmented plantarflexion power on preferred walking speed and economy in young and older adults. Gait Posture **25**(4), 627 (2007)