Sādhanā (2020)45:204
https://doi.org/10.1007/s12046-020-01453-8

A TDOA-based multiple source localization using delay density maps

RITU BOORA* and SANJEEV KUMAR DHULL

Guru Jambheshwar University of Science and Technology, Hisar 125001, India
e-mail: rituboora@gmail.com; sanjeevdhull2011@yahoo.com

MS received 8 April 2020; revised 21 June 2020; accepted 5 July 2020

Abstract. The higher computational efficiency of the time difference of arrival (TDOA) based sound source localization makes it a preferred choice over steered response power (SRP) methods in real-time applications. However, unlike SRP, its implementation for multiple source localization (MSL) is not straight forward. It includes challenges as accurate feature extraction in unfavourable acoustic conditions, association ambiguity involved in mapping the feature extractions to the corresponding sources and complexity involved in solving the hyperbolic delay equation to estimate the source coordinates. Moreover, the dominating source and early reverberation make the detection of delay associated with the submissive sources further perplexing. Hence, this paper proposes a proficient three-step method for localizing multiple sources from delay estimates. In step 1, the search space region is partitioned into cubic subvolumes, and the delay bound associated with each one is computed. Hereafter, these subvolumes are grouped differently, such that whose associated TDOA bounds are enclosed by a specific delay interval, are clustered together. In step 2, initially, the delay segments and later each subvolume contained by the corresponding delay segment are traced for passing through estimated delay hyperbola. These traced volumes are updated by the weight to measure the likelihood of a source in it. The resultant generates the delay density map in the search space. In the final step, localization enhancement is carried out in the selected volumes using conventional SRP (C-SRP). The validation of the proposed approach is done by carrying out the experiments under different acoustic conditions on the synthesized data and, recordings from SMARD & Audio Visual 16.3 Corpus.

Keywords. Sensor arrays; acoustic signal processing; time difference of arrival (TDOA); multiple source localization; steered response power-phase transform (SRP-PHAT); generalized cross-correlation (GCC).

1. Introduction

The localization of an acoustic event is the fundamental prerequisite for developing applications such as human-machine interaction, robotics, videoconferencing, hands-free communication, military surveillance, acoustic scene analysis, hearing aids, smart environments and many more. With the growing demand for the proficient methods of sound source localization (SSL), many efforts have been devoted by researchers to examine this field and develop several methodologies to solve this central problem of signal processing. Amongst the existing methods, those based on TDOA are fast, but their performance deteriorates in challenging conditions. Hence, there is continuous effort to make them robust and computationally simple [1–4].

TDOA based techniques involve a 2-step procedure wherein initially the received signal at the sensor array is processed to compute the TDOA between the microphone pairs. Thereafter the estimated delay for each sensor pair is mapped to the corresponding source positions by some mathematical manipulation. The key to the effectiveness of these localizers is an accurate and robust TDOA estimators and hence, to overcome this issue, several mathematical procedures like Adaptive Eigen Value Decomposition (AED) [5], GCC-PHAT [6], a geometric approach using non-coplanar arrays [7], a modified version of Maximum Likelihood [8], a combination of linear interpolation and Cross-Correlation (CC) [9], etc. have been adopted. Amongst these, GCC-PHAT is the most widely used method to estimate TDOA as it is conceptually simple [10–12]. As shown in figure 1, each TDOA measurement in the realizable delay interval $[-D/v, D/v]$ (where D is inter sensor distance and v is the velocity in medium) of a sensor pair corresponds to a distinct half hyperbolic branch in 2-D with sensor locations as its foci's. The estimated source position is the intersection point of the hyperbolic branches corresponding to estimated TDOA at different sensor pairs. However, solving the nonlinear hyperbolic complex and moreover, the solution becomes inconsistent with inaccuracies in TDOA measurements. Henceforth, researchers are continuously involved to suggest numerous approaches to solve these hyperbolic equations which include planar

*For correspondence
Published online: 13 August 2020

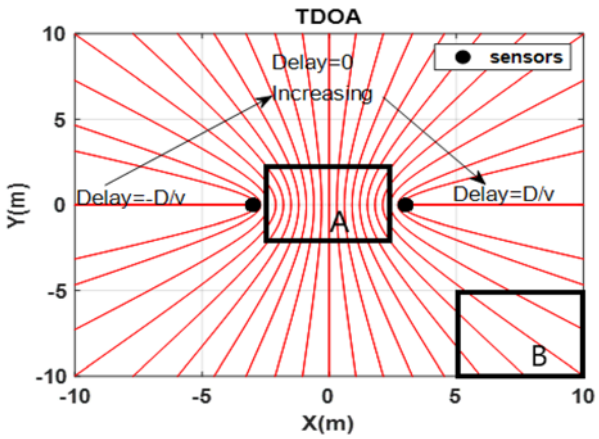

Figure 1. Confocal hyperbolas for a range of physically realizable TDOAs' of a sensor pair.

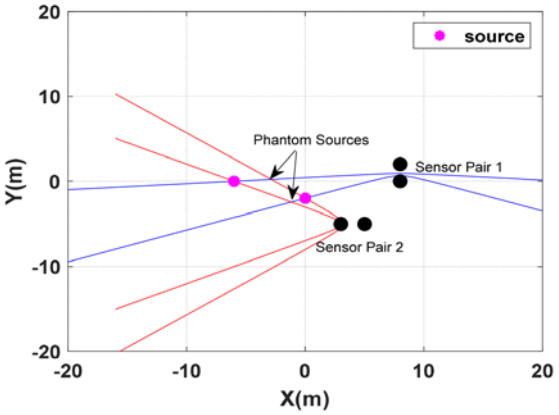

Figure 2. Two hyperbolas from each sensor pair (black dots) in respect of two sources (purple dots).

intersection (PI), spherical intersections (SX), spherical interpolation (SI) [13], least square and weighted least square (WLS), global branch and bound [14], space-range reference frames [15], WLS with cone tangent plain constrain [16], etc.

1.1 *Challenges in TDOA based multiple source localization (MSL)*

In a realistic scenario, there are often situations where multiple sources are simultaneously active, and there is a need to localize them simultaneously. Several authors have proposed different methods based on Steered Response Power (SRP) and Time Difference of Arrival (TDOA) to localize multiple sources; some of these are discussed in the literature section. However, unlike SRP, none of the existing TDOA based localization methods can be directly extended to multiple sources due to several challenges discussed below.

- *Multi-source ambiguity*: In an acoustic setup with K sources ($K > 1$), it is generally expected that each sensor pair GCC will contain at least K local extrema, corresponding to each source. Identifying the correct TDOA estimate and matching to the corresponding source is a big challenge called multi-source ambiguity. As displayed in figure 2, the hyperbolas corresponding to the estimated delays of both the sources at each sensor pair gives four intersections, where only two corresponds to the true source and remaining are the phantom sources. Subsequently, it gets complicated to differentiate the correct locations from the phantom sources. Furthermore, the TDOA mapping becomes more problematic with reverberation when the indirect path GCC peak of a source is higher than that of its direct path.
- *Dominant source*: In a multi-sourced environment, due to different frequency spectrum or propagation parameters, one source may supersede other sources. The work in [17] has ascertained the relationship between the correlation values and source characteristics. The submissive sources show less coherence at most of the microphone pairs, and thus, it is overtaken by early reverberations more often. As a result, the GCC values at these secondary peaks might not provide the correct likelihood of a source.
- *Spatial resolution*: The two sources may become indistinguishable when narrowly positioned in the lower sensitive area. It can be analysed from figure 1, where the sensor pair shows varying spatial sensitivity in the search space. The subregion B, when compared to subregion A, has lower spatial sensitivity as it has only a very few and widely spaced hyperbolas passing through it. Thus, two narrowly separated sources in the subregion B may show the same delay to the sensor pair and therefore become indistinguishable.

A proficient TDOA based method to overcome the above-said ambiguities is proposed in the paper. However, before proceeding with the proposed method, the related literature and their significant contributions are discussed below. The remaining part of the paper is delivered as follows. The literature on multiple source localization is elaborated in section 2. Section 3 introduces the proposed method and, the results and discussion are presented in section 4. Conclusions are provided in section 5.

2. Literature on multiple source localization

2.1 *TDOA based*

TDOA based clustering method [18] suggests a three stage strategy which includes pre-whitening the signal, computing the TDOA of the direct path and early reflections by the time delay estimation method and, clustering the hyperbolic intersection while rejecting the outliers. The intersection

points calculated from all the permutation of the estimated TDOAs are clustered based on some criteria. The centre of each cluster is then considered as the estimated source location. Later, a few variants of clustering algorithm as short clustering for moving speakers [19], K-means++ [20], competitive K-means [21], multi-path matching pursuit [10] were also explored. However, these algorithm results in voluminous permutations of clusters in a reverberant environment which subsequently tempts to phantom source locations. Authors in work [22–24] suggested solution based on consistent graph along with filters like raster condition, Zero-sum condition, upper bound on TDOA etc. to eliminate the undesirable TDOA. However, the cumbersome process involved in implementing the graphs limits their applicability. Work in [25] uses multiple hypothesis frameworks for TDOA disambiguation, but its performance degrades with reverberations.

A hybrid method based on TDOA and SRP is introduced in work [12] where a spatial enclosure is partitioned into several elemental regions such that each partition contains at most one source. The characteristics delay interval associated with each elemental region maps to an area circumscribed by the pair of hyperboloid branches called Inverse Delay Interval Region (IDIR) or characteristic IDIR. However, the foremost challenge here is to decide the shape of the elemental region such that it contains at most one source and approximates the intersection of IDIR of all the sensor pairs. The factors like unequal characteristics intervals associated with each partition, reverberations, additive noise, etc., adversely affect in estimating the accurate active region. It is illustrated in figure 3, where a source placed at (1.5, 3.3) is erratically estimated in another elemental region at (3.1, 7.3). Further, for TDOA disambiguation, the delay information is discarded from the sensors pairs whose characteristics IDIR contain multiple active regions. It consequently restricts the usage of competent methods to solve the hyperbolic equations and hence affects the accuracy within the active sector. It is also demonstrated in figure 3, where a source placed at coordinates (1, 5) is localized at (1.7, 4.8). Moreover, this detection of IDIR with the single active region is computationally expensive.

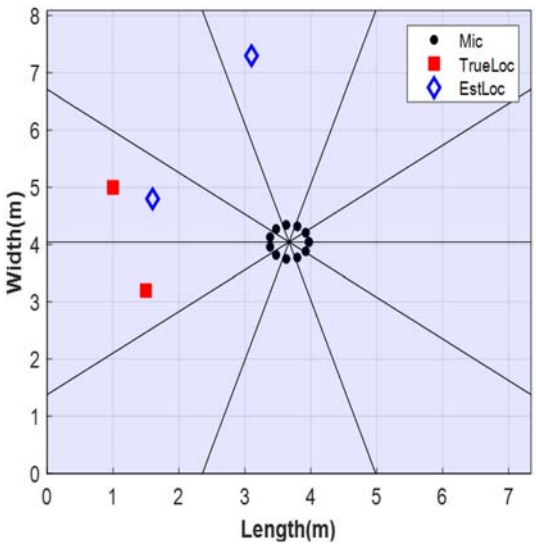

Figure 3. Active sector and fine localization error using IDIR.

2.2 *SRP based*

The beam steering methods are very robust but, evaluating the objective functions at all the candidate locations is quite cumbersome and subsequently limits their practical deployment. To accelerate SRP based localization, the work in [26–28] offers hierarchical search procedure for coarse to fine identification of potential source locations. However, they tend to miss the source locations before the final stage as after each iteration they prune the candidate source locations with lesser probability.

Work presented in [29] manipulates the GCC-PHAT measurements to extract the information of submissive sources. In this method, the dominating speaker location is deduced straightforwardly by maximizing the Global Coherence Field (GCF) acoustic map. Further, the dominant source is de-emphasized by reducing the magnitude of the GCC at delays associated with its locations. The GCF acoustic map is re-computed from de-emphasized GCC to extract the information of the submissive source. Hence, this method is referred by Global Coherence Field De-emphasized (GCF-D) for convenience in further text. However, the prime challenge here is to design the optimum de-emphasizing function, which accentuates the GCC at desired delays without affecting the remaining. As illustrated, the inappropriate de-emphasization of the dominating source shifts the second peak of the GCF acoustic map from the location (4.9, 3.2) in figure 4a to (4.2, 3.5) in figure 4b. Furthermore, a narrow de-emphasis filter may also not remove the relevant GCC effectively and adversely affect the performance of the method. Besides this, the acoustic map is to be computed times' the number of sources present and hence, demands high computations and necessitates source prior information. However, none of the current methods has achieved the desired characteristics to meet the growing demand in numerous applications.

3. Proposed method

In the proposed method, initially, the search region is divided into cubic subvolumes, and the associated delay bound with each one is computed. Hereafter, the physically realizable delay interval of each sensor pair is segmented into small intervals and, the subvolumes are assembled in a way that whose associated delay bound lies in a segmented

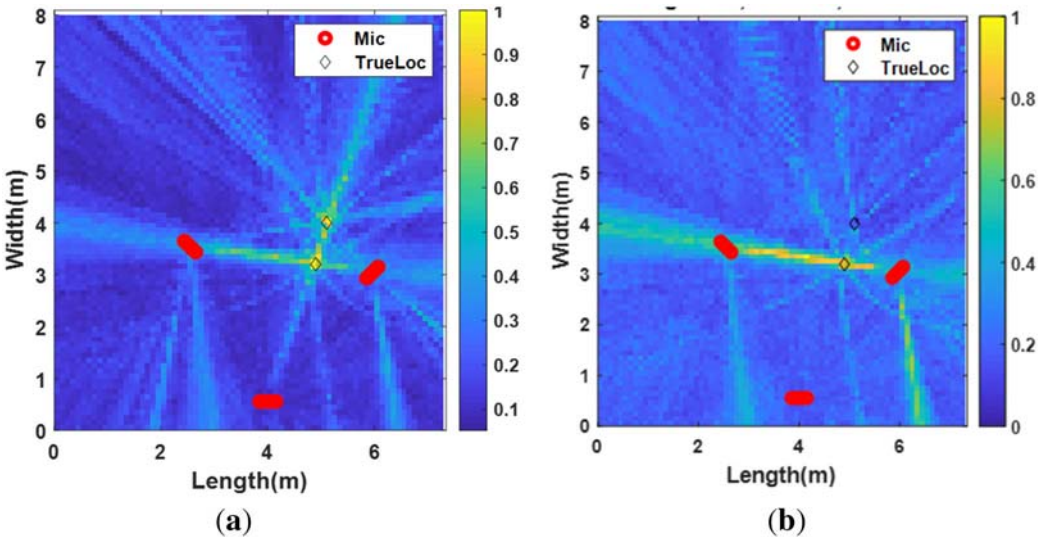

Figure 4. GCF map (**a**) Prior to De-emphasis. (**b**) After GCC De-emphasis.

interval are clustered together. Following this, the delay segments through which estimated delay hyperbola of the corresponding sensor pair passes are identified. The volumes contained by these delay segments are scanned and updated by weight only if the estimated delay hyperbola passes through them. When executed for all the sensor pairs, it provides the likelihood of a source in any volume. Towards the end, the localization refinement is performed using C-SRP in the subvolumes that are more likely to contain the source. The proposed method is detailed in the following subsections.

3.1 *TDOA estimation*

The received signal, $s_i(t)$ at ith sensor and at instant t in a multi-path and multi-source environment of K sources can be modelled using Eq. (1) where $x_k(t)$ is the kth source signal transmitted, $h_{k,i}(t)$ is impulse response from kth source to ith sensor, and $\mathcal{N}_i(t)$ is Additive White Gaussian Noise (AWGN).

$$s_i(t) = \sum_{k=1}^{K} h_{k,i}(t) * x_k(t) + \mathcal{N}_i(t) \tag{1}$$

From a set of M sensors, the P different sensors pair are constituted as $M(M-1)/2$. The TDOA (in samples) associated with a source placed at $X(x, y, z)$ and the two sensors of pth pair positioned at X_{p1}, X_{p2} respectively, is evaluated from Eq. (2) where f_s is the sampling frequency and v is the propagation speed of the signal.

$$\tau_p(X) = round\left(\left|\left|X_{p1} - X\right|\right| - \left|\left|X_{p2} - X\right|\right|/v\right) * f_s \tag{2}$$

The set of physically attainable delay for any sensor pair is bounded by the distance, D between the sensor pair and is characterized as $\{-\tau_{p,max}, -\tau_{p,max}+1, \ldots, -1, 0, 1, \ldots, \tau_{p,max}-1, \tau_{p,max}\}$ with $\tau_{p,max} = round(f_s * D/v)$. The negative delay means that the source is closer to the sensor $p1$ than $p2$ of pth pair while a positive delay means a vice-versa. Consequently, the delay measured between any two microphones signals using GCC-PHAT is restricted by the delay interval $[-\tau_{p,max}, \tau_{p,max}]$ to eliminate the undesired computations [11].

The coherence, $\psi_p[\tau]$ between the two received signals, $s_{p1}(t)$ and $s_{p2}(t)$ at pth sensor pair is computed using Eq. (3) with S_{p1}, S_{p2} as their respective N point Discrete Fourier Transform (DFT) where * represents a complex conjugate. The estimated relative delay $\widehat{\tau_p}$, evaluated by Eq. (4) for pth sensor pair is the value of τ that maximizes the normalized coherence function $\psi_p[\tau]$ [6].

$$\psi_p[\tau] = \sum_{m=0}^{N_{DFT}-1} \frac{S_{p1}[m]S_{p2}^*[m]}{|S_{p1}[m]S_{p2}^*[m]|} e^{j\frac{2\pi\tau m}{N_{DFT}}} \tag{3}$$

$$\widehat{\tau_p} = \underbrace{\arg\max}_{\tau\in[-\tau_{max},\tau_{max}]} \psi_p[\tau] \tag{4}$$

3.2 *Constant and distinct delay hyperbolas in the search space*

The relative delay is given in Eq. (2) and maps any point X from the spatial search space to 1-D TDOA vector. Presence of several locations with similar TDOA on a half hyperboloid branch with foci at two microphones, it is inferred that this mapping is not unique. This geometrical interpretation of a constant TDOA to a hyperboloid provides the locus of the possible source location. The relative delay is continuous in X, and henceforth, the source locations with

different TDOA will lead to distinct and confocal half hyperboloid branches, as shown in figure 1, for 2-D.

3.3 *Delay bound associated with a cubic volume*

The TDOA associated with a volume can take values only in a range defined by its boundary surface. Therefore, the delay bounds related to a volume can be calculated from the maximum and minimum TDOA values on its boundary surface. However, for the accurate associated TDOA range measurement, it requires computing the TDOA on a dense grid on all the bounding surfaces of the volume, making it computationally expensive [30]. Therefore, in this paper, a gradient-based approach [31] has been followed that is equally efficient and involves a much lesser number of computations.

Initially, a given search volume V that contains the source(s) is segmented into a set of $\sum$ uniform cubic subvolumes, V_{sub}. These subvolumes are created such that each grid point of a uniform spatial grid in the search space defines the center of a subvolume. The associated boundary to each subvolume is the symmetrical region of half the grid resolution around each point.

Next, the direction of the maximum increase of delay at each grid point is calculated by the gradient function, given in Eq. (5). The gradient in the individual direction (x, y and z) is computed by Eq. (6) where $\gamma \in (x, y, or\, z)$ [31].

$$\nabla\tau_p(X) = \left[\nabla_x\tau_p(X), \nabla_y\tau_p(X), \nabla_z\tau_p(X)\right] \tag{5}$$

$$\nabla_\gamma\tau_p(X) = \frac{\partial\tau_p(X)}{\partial\gamma} = \frac{1}{c}\left(\frac{\gamma-\gamma_{p1}}{||X-X_{p1}||} - \frac{\gamma-\gamma_{p2}}{||X-X_{p2}||}\right) \tag{6}$$

Thereafter, the TDOA bounds for a symmetrical subvolume V_{sub} surrounding a grid location are calculated from Eq. (7) and Eq. (8) where d, defined in Eq. (9) is the distance from the center to the boundary of V_{sub} and r is the grid size (i.e., is the distance between the centers of two adjacent V_{sub}).

$$\tau_{p,min}^{V_{sub}} = \tau_p(X) - ||\nabla\tau_p(X)|| \cdot d \tag{7}$$

$$\tau_{p,max}^{V_{sub}} = \tau_p(X) + ||\nabla\tau_p(X)|| \cdot d \tag{8}$$

$$d = \frac{r}{2}\min\left(\frac{1}{|\sin\theta cos\phi|}, \frac{1}{|\sin\theta sin\phi|}, \frac{1}{|cos\theta|}\right) \tag{9}$$

Where $\theta = cos^{-1}\left(\frac{\nabla_z\tau_p(X)}{||\nabla\tau_p(X)||}\right)$ and $\phi = atan_2(\nabla_y\tau_p(X), \nabla_x\tau_p(X))$.

3.4 *Delay density maps*

By definition of associated TDOA to a volume, for any source located in volume V in the search space, the TDOA at a sensor pair for this source also lies in a range, tightly bounded by the two hyperboloid branches. With ∂V as the boundary surface of this given search volume [31], the minimum and maximum delay comes out to be $\tau_{min}^V = \min_{X\in V}\tau_p(X \in \partial V)$ and $\tau_{max}^V = \max_{X\in V}\tau_p(X \in \partial V)$ respectively. Using this, we compute the Delay Density Map of the entire search region.

Therefore, an estimated kth delay, τ_p^k from the GCC of the pth sensor pair is mapped to a subvolume V_{sub} only if it is contained by its associated TDOA bound, $[\tau_{p,min}^{V_{sub}}, \tau_{p,max}^{V_{sub}}]$. Furthermore, only the subvolumes which are mapped by the estimated delay in the search space are updated by a unity weight. It is mathematically represented by Eq. (10).

Henceforth, the objective function for each volume, $W(V_{sub})$ is evaluated from Eq. (11), which involves summing the projections of Eq. (10) for the K delays over the comprehensive set of sensor pairs. When computed across all the $V_{sub} \in \sum$, it represents the density of passing through estimated delay hyperbolas through each and hence named delay density map (DDM). Thus, the DDM displays the likelihood of a source in each subvolume.

$$\chi_p\left(\tau^k, V_{sub}\right) = \begin{cases} 1, for\, \tau_{p,min}^{V_{sub}} \le \tau_p^k \le \tau_{p,max}^{V_{sub}} \end{cases} \tag{10}$$

$$W(V_{sub}) = \sum_{p=1}^{P}\sum_{k=1}^{K}\tau_p\left(\tau^k, V_{sub}\right) \tag{11}$$

Assigning equal weights to the mapped subvolumes for each K number of GCC peaks benefits to recover the sources with small coherence values. Moreover, it helps to suppress the early reverberation with large GCC peak values. It is demonstrated with the help of figure 5, which displays the localization of three sources in a frame for various techniques when RT60 is 0.55 sec. In figure 5a and b, the acoustic map of the search region is developed using C-SRP and Mean Modified SRP (MMSRP) [32] respectively while figure 5c displays the delay density map. From figure 5a and b, it is observed that each acoustic map has many high SRP points at locations other than the source and hence couldn't localize all the three sources. In contrast, figure 5c shows the high likelihood only in the regions that contain the source.

3.5 *Computing DDM using subvolume clustering*

Up to now, each $V_{sub} \in \sum$ in the enclosure needs to be scanned for mapping an estimated TDOA to a respective V_{sub} and therefore, it is a computationally expensive process. Hence, this section introduces a delay segmentation and subvolume clustering approach to minimize the DDM computational expense.

3.5a *Delay segmentation:* As discussed, the distinct TDOA to spatial mapping appears as continuous confocal half hyperboloid branches when the delay is traversed from $-\tau_{p,max}$ to $\tau_{p,max}$ for a sensor pair p. Consequently, a

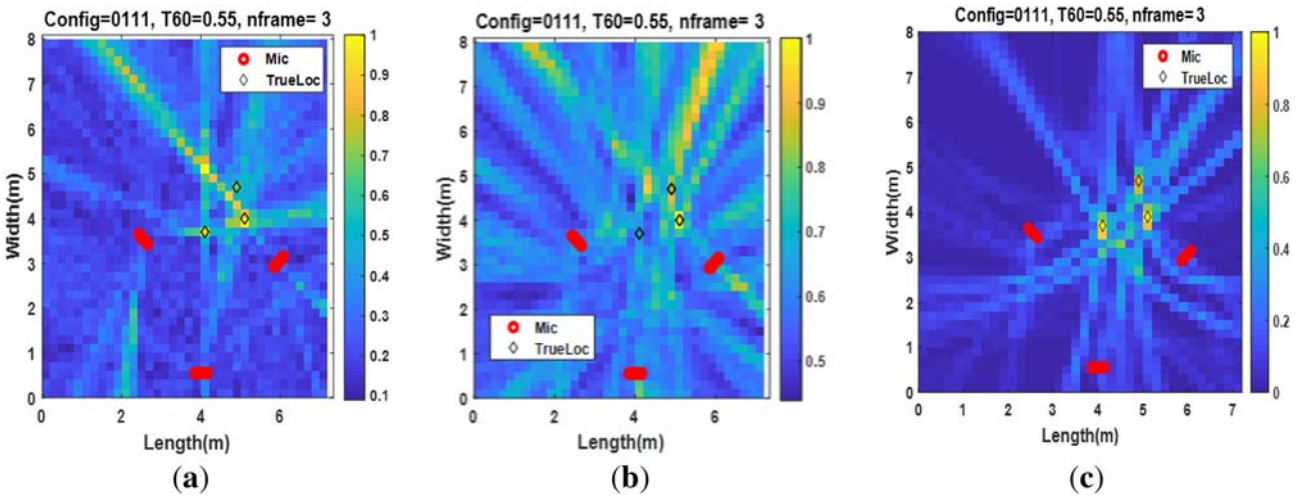

Figure 5. (**a**) Acoustic map using C-SRP. (**b**) Acoustic map using MMSRP. (**c**) DDM using the proposed method.

segmented delay interval $[I_{p,min}: I_{p,max}] \subseteq [-\tau_{p,max}: \tau_{p,max}]$ is mapped to a region bounded by two half hyperboloids $h_p(I_{p,min})$ and $h_p(I_{p,max})$ corresponding to $I_{p,min}$ and $I_{p,max}$ respectively. Henceforth, the entire range of realizable delay for a sensor pair is segmented into T intervals, as given in Eq. (12). As illustrated in figure 6, each region corresponds to a segmented delay interval and therefore has hyperbolic spatial boundaries.

$$[-\tau_{p,max} : \tau_{p,max}] \leftrightarrow [I^1_{p,min} : I^1_{p,max} \cup I^2_{p,min} : I^2_{p,max} \cup \cdots I^T_{p,min} : I^T_{p,max}] \tag{12}$$

3.5b *Subvolume clustering as per delay segmentation:* All the search space subvolumes are grouped such that their respective upper TDOA bound, $\tau^{V_{sub}}_{p,max}$ lying in one delay interval $[I^t_{p,min}: I^t_{p,max}]$ are clustered together. It is mathematically represented by Eq. (13) where $t \in [1, 2, \ldots, T]$, i.e. the tth delay interval. In this equation, $\sum \leftarrow \sum - V_{sub}$ removes a subvolume from the complete set of subvolumes $\sum$ once it becomes a part of any cluster. This step ensures that each group G^t_p has unique elements where G^t_p is the tth group for the pth sensor pair.

$$\forall V_{sub} \in \sum, G^t_p \leftarrow G^t_p \cup V_{sub} \; and \; \sum \leftarrow \sum - V_{sub} \; if \left(I^t_{p,min} < \tau^{V_{sub}}_{p,max} \leq I^t_{p,max}\right) \tag{13}$$

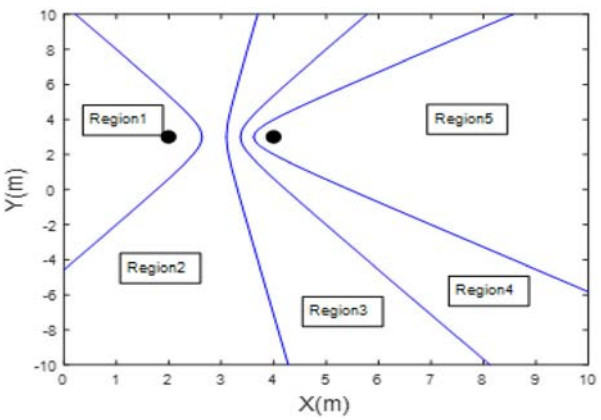

Figure 6. Delay segmentation and corresponding hyperbolas of a sensor pair.

3.5c *Redefining the lower bound for each delay segment:* In the previous subsection, the elements of a group G^t_p have been clustered considering only their upper TDOA bound $\tau^{V_{sub}}_{p,max}$. However, since each V_{sub} has its own lower delay bound, $\tau^{V_{sub}}_{p,min}$ and it may overlap with the delay interval of the previous segment $(t-1)$. It necessitates redefining the lower TDOA bound of each segment exclusively such that lower bound of the clustered volumes is also considered, as given in Eq. (14). The Eq. (13) and Eq. (14) collectively guarantee that there remains no V_{sub} which is not part of any cluster and no subvolume is mapped more than once for an estimated TDOA.

$$I^t_{p,min} = min\left(min\left(\tau^{V_{sub}}_{p,min}\right), I^t_{p,min}\right) \forall V_{sub} \in G^t_p \tag{14}$$

3.5d *Delay density maps:* Hereafter, for each sensor pair, initially, the delay segments and later the volume in this delay segments are traced for passing through corresponding estimated delay hyperbola. These traced V_{sub} are updated by a weight using Eq. (11) to measure the likelihood of a source in it. This whole process minimises the required TDOA mapping from the set $\sum$ to a subset of subvolumes contained by the traced delay segments. It thus makes the DDM computationally inexpensive.

3.6 *Refined position estimation*

In the end, the refined position is evaluated in the K higher likelihood selected subvolumes using C-SRP [11]. The source location, X_k in kth sub-volume V_{sub}^k is estimated using Eq. (15).

$$X_k = \underbrace{argmax}_{X \in V_{sub}^k} \sum_{p=1}^{P} \psi\left(\tau_p(X)\right) \tag{15}$$

4. Results and discussions

In this section, the proposed method is evaluated on various simulated acoustic settings, SMARD database and AV16.3 corpus.

4.1 *Performance evaluation on simulated data*

The simulated dataset is generated using Image method [33, 34] in a room of dimensions $[7.34\,\text{m} \times 8.01\,\text{m} \times 2.87\,\text{m}]$ for various speech and music signals from TSP speech database [35] and MIS database respectively [36].

Before considering an array configuration, its spatial sensitivity was examined using Geometrically Sampled Grid procedure [37], as shown in figure 7. Each array configuration has nine sensors which are equally spaced for small UCA and Large-UCA (L-UCA) of figures 7a and b, respectively. Whereas, the Distributed Linear Array (DLA) in figure 7c has three linear arrays with equally distributed sensors. As the physically realizable delay interval expands with an increase in the inter-sensor distance; the number of hyperbolas passing through the region intensifies and improves the spatial resolution [38]. It explains the enhanced sensitivity of Large-UCA over small UCA and DLA (cf. figure 7).

Furthermore, the delay density maps of the search region are developed using Eq. (11) for small UCA, L-UCA and DLA, respectively in figure 8a, b and c. As observed, the three sources are accurately localized with DLA and L-UCA while small UCA shows comparatively higher delay density only in the source direction. Henceforth, the performance of the proposed method for spatial localization is evaluated using DLA and L-UCA array configuration.

Moreover, as exhibited in figure 9, the spatial sensitivity of a sensor array configuration also impacts the minimum resolvable distance between the two sources. The two sources show the relative spread of the peak in the delay density map (DDM) when placed in the lower sensitive area as in figure 9b in comparison to figure 9a. Hence, considering the spread of the peaks to the neighbouring subvolumes, the two sources should be kept set apart by at least twice the grid resolution for efficient localization. While evaluating the DDM on a coarse grid of 0.2 m, this minimum distance is only 0.4 m and hence can come across the practical requirements. Thus, the spatial resolution of the proposed method is much higher than that of IDIR, where it is limited by the size of its elemental region. Moreover, this resolution can be improved further by evaluating the DDM on a denser grid with a marginal compromise in the computational cost.

The performance of the proposed method is evaluated and compared with the 2-D state-of-art methods, IDIR [12] and GCF-D [39] under various simulated acoustic settings. The proposed method (Prop) and GCF-D have been simulated on DLA and large UCA. At the same time, IDIR is limited to DLA configuration for competitive comparison as it yields poor results with L-UCA. It is mainly due to inappropriate intersections of characteristics IDIR with sectored elemental regions for large UCA. The methods are simulated for localization with a sampling frequency of 44.1 KHz, frame length of 4096 samples with 50% overlap and tested for 200 trials over random source locations. The proposed method is implemented with the initial grid resolution of 0.2 m to compute DDM and a final grid resolution of 0.01 m in the selected K regions. The GCF-D is computed with a grid resolution of 0.01 m whereas IDIR is implemented with 20 elemental regions in the form of sectors.

Table 1 presents the Root Mean Square Error (RMSE) in the localization of the two sources separately under different RT60 (sec). As observed from these results, the dominating source is localized with lower error with all the methods. However, the proposed method shows a considerable improvement in the localization of the submissive source and a slower degradation to reverberation as compared to other methods. The array geometry is another critical factor which impacts the performance of these methods. The source localization performance of the proposed method and GCF-D are superior with large UCA when compared to that of DLA, being enhanced spatial resolution provided by the earlier. Table 2 presents the averaged RMSE over the number of sources present for different reverberant settings. As clearly shown by these results, the proposed method shows superior localization results for the multiple sources, whereas GCF-D is highly influenced by it.

The robustness to uncorrelated noise is evaluated by adding White Gaussian Noise (WGN) to each received microphone signal to achieve different SNR. Performance comparison of the proposed method with existing methods is presented in figure 10 for a range of SNR.

Noticeably, the proposed method shows the enhanced results over the state-of-art methods. Tables 1, 2 and figure 10 together validate the competency of the proposed method in adverse acoustic conditions over the existing methods.

4.1a *Computational cost:* Table 3 displays the comparison of the execution time (in sec) to localize the two

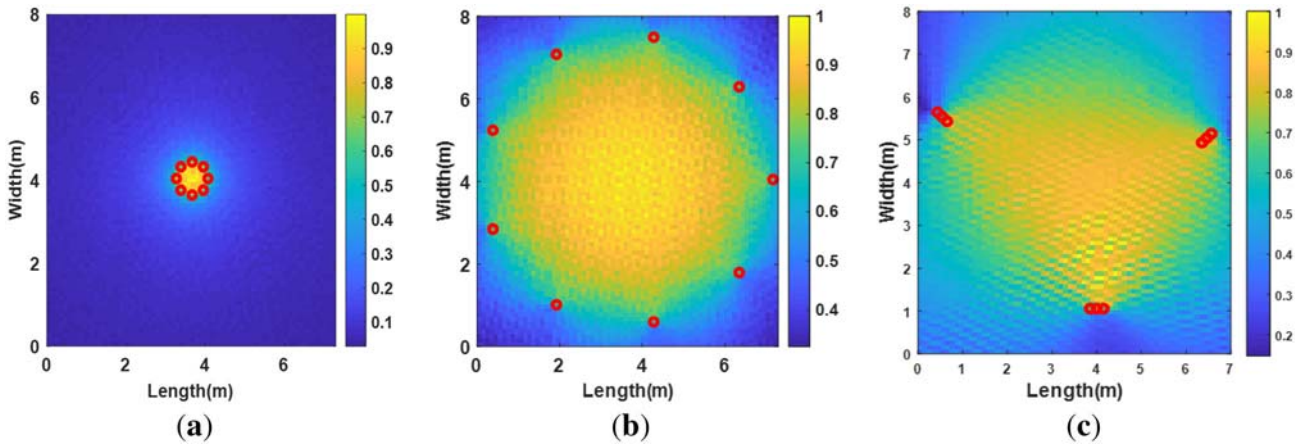

Figure 7. Analysing the spatial sensitivity of different sensor array geometries. (**a**) Small UCA, (**b**) Large UCA and (**c**) Distributed Linear Array (DLA).

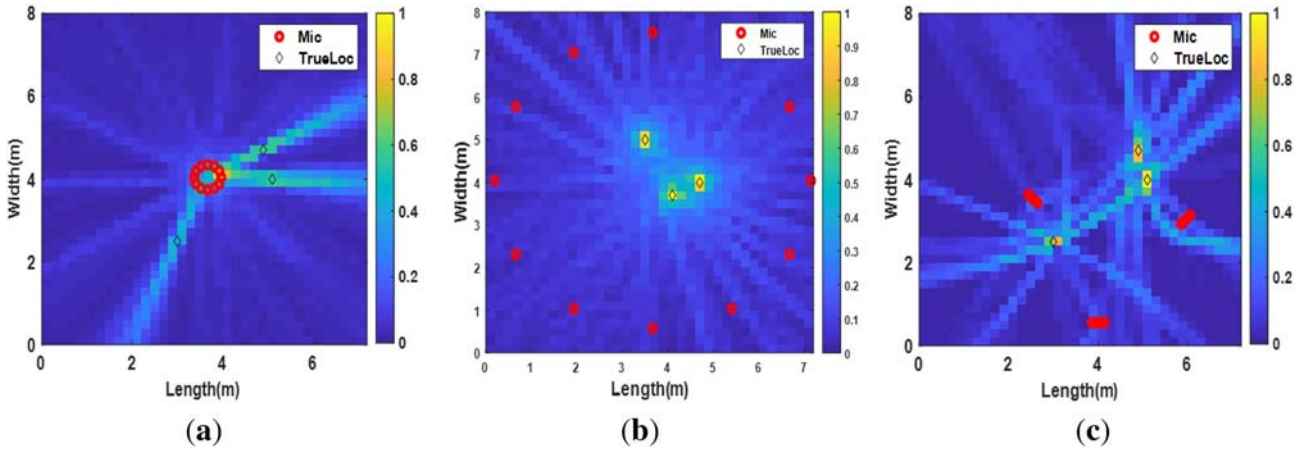

Figure 8. Delay density map with different geometries (M = 12) for RT60 of 0.5 sec. (**a**) Small UCA, (**b**) Large-UCA and (**c**) Distributed Linear Array.

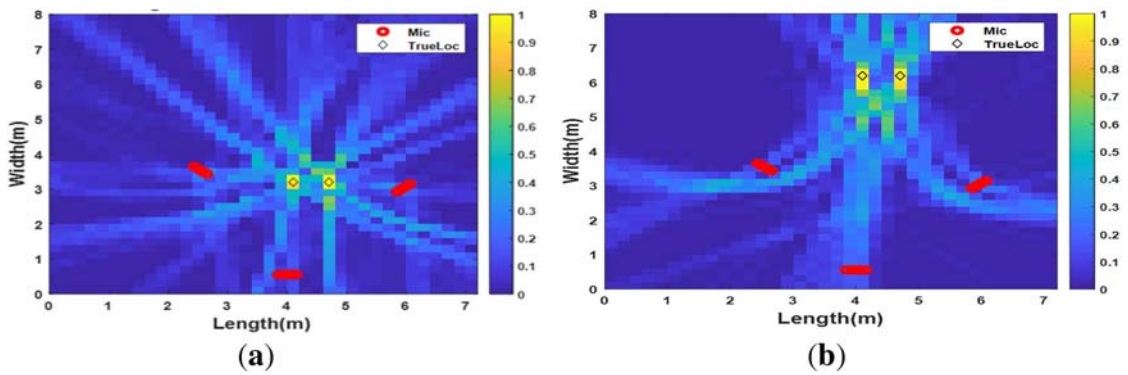

Figure 9. DDM in the different sensitive regions. (**a**) Higher sensitive region. (**b**) A comparatively lower sensitive region.

sources using a single frame of length 1 sec for the discussed methods. The methods are simulated in MATLAB and run on a system with Intel Core i5 processor (1.99 GHz) and 8 GB RAM.

As shown, the GCF-D technique has the highest run-time as it involves computing the SRP twice (as there are two sources) on a fine grid and removal of the dominating peak from GCC of each sensor pair before re-computing the

Table 1. RMSE (m) in localization shown distinctly for the two sources (Src1, Src2) at different RT60 and SNR = 25 dB.

	RT60 = 0.11 (sec)		RT60 = 0.33 (sec)		RT60 = 0.55 (sec)	
Technique	Src1	Src2	Src1	Src2	Src1	Src2
IDIR(DLA)	0.16	0.38	0.20	0.51	0.31	0.82
GCF-D(DLA)	0.21	0.61	0.32	0.82	0.55	1.01
Prop (DLA)	0.12	0.17	0.15	0.21	0.20	0.29
GCF-D(UCA-D)	0.15	0.44	0.26	0.76	0.41	0.94
Prop(UCA-D)	0.09	0.12	0.12	0.18	0.17	0.24

Table 2. RMSE in localization averaged over the number of active sources at various RT60 settings and SNR = 20 dB.

	RT60 = 0.33 sec			RT60 = 0.55 sec		
	Number of active sources			Number of active sources		
Technique	1	2	3	1	2	3
IDIR(DLA)	0.11	0.38	0.61	0.21	0.56	1.12
GCF-D(DLA)	0.13	0.62	1.21	0.28	0.82	1.42
Prop (DLA)	0.08	0.21	0.38	0.16	0.27	0.55
GCF-D(UCA-D)	0.11	0.57	1.01	0.24	0.72	1.22
Prop(UCA-D)	0.06	0.19	0.29	0.12	0.24	0.42

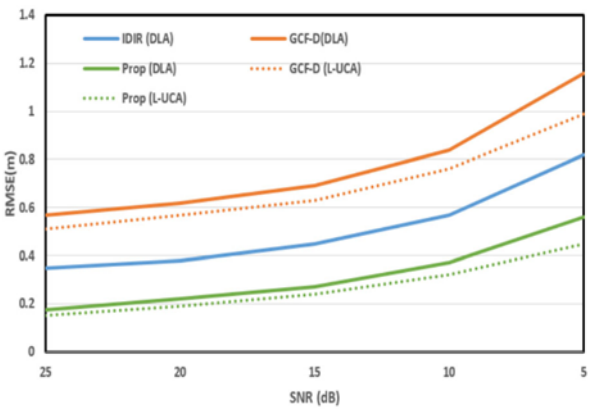

Figure 10. RMSE averaged over two active sources for a range of SNR.

acoustic map. The IDIR technique requires comparatively lower run-time, but its initialization process is quite burdensome which includes dividing the search space into elemental regions and finding the TDOA interval associated to each sector using a fine grid inside them. Moreover, to localize closely placed sources, the number of segments has to be increased to ensure that each segment has a maximum one source. It ultimately makes it computationally more expensive. The proposed method minimizes its initialization time by following a volumetric grid and gradient-based approach to estimate the TDOA interval associated with each volume. It further reduces the run time by following a

Table 3. Execution time (sec) to localize the two sources (Src1, Src2) in a frame of 1 sec using DLA.

		Techniques		
Aspects		GCF	IDIR	Prop
Initialize(s)		2	24.60	8.56
Evaluate GCC(s)		10.61	10.61	10.61
Location estimation from TDOAs	*Src1*	38.5	6.16	4.21
	Src2	46.6	6.24	4.21
Total Time (s)		97.71	47.61	27.59

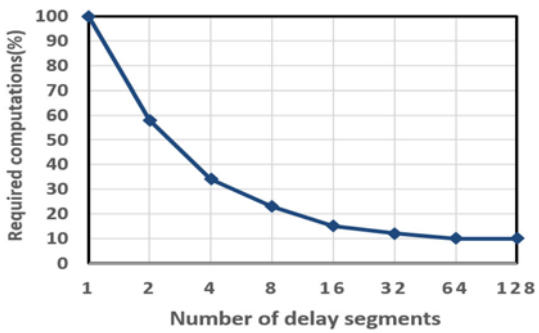

Figure 11. Reduction in computational cost with delay segmentation.

delay segmentation and sectoring approach for TDOA to spatial mapping. As displayed by the results, the proposed method takes 27.59 sec to localize the two sources using a

Table 4. RMSE (m) in localization shown distinctly for two sources (Src1, Src2) on SMARD database.

	Anechoic signals		Reverberant signals	
Technique	Src 1	Src 2	Src 1	Src 2
IDIR	0.18	0.41	0.34	0.86
GCF-D	0.23	0.63	0.59	1.08
Prop	0.11	0.13	0.19	0.25

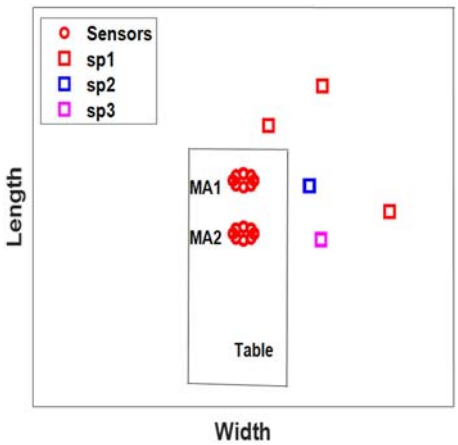

Figure 12. Physical set-up for AV16.3 and speakers locations for seq37-3p-0001.

ten delay segmentation for each pair, which is immensely lesser than the other two.

A further study is carried out to observe the impact of delay segmentation on the computation cost in the spatial mapping of the estimated TDOAs. Figure 11 presents the percentage of computation required with an increase in the number of delay segments. As shown, the percentage of necessary calculations initially reduces exponentially and afterwards follows the constant trajectory.

4.2 *Performance evaluation with SMARD database*

SMARD database [40] has multichannel recordings of anechoic and reverberant signals with different sensor arrays configuration in a room with dimension 7.34 m × 8.09 m × 2.87 m at Aalborg University. The proposed method is tested on the recording of anechoic and reverberant signals on configuration 0011 and 0111 having sensors array placement identical to that of figure 8c. These configurations have the sources placed at (2.00, 6.50, 1.25) and at (3.50, 4.50, 1.50) with angle − 90° and − 45° respectively in XY-plane. The received signals at the respective sensors in these configurations are added to get the resultant of the two sources (cf. Eq. 1). The performance evaluation of the proposed and existing method for source localization is tabulated in Table 4 for anechoic and reverberant signals differently. From these results, it is observed that the proposed method outdoes the state-of-art methods.

4.3 *Evaluation using audio visual 16.3 (AV16.3) corpus*

In this section, experimental evaluation of the proposed method is done using sequences seq01-1p-0000 and seq37-3p-0001 for one and three speakers, respectively, selected from Audio Visual 16.3 corpus (AV 16.3) [41]. The corpus

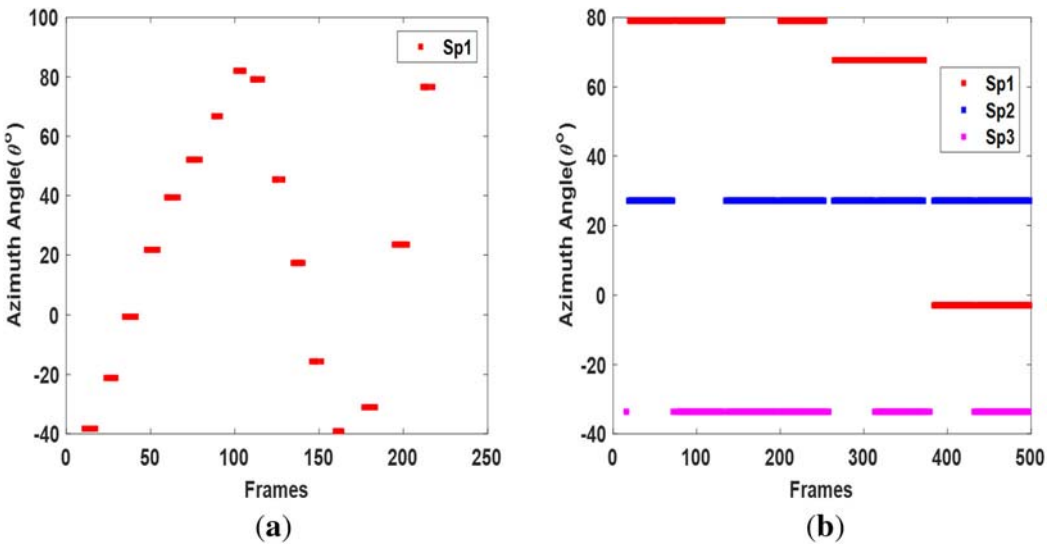

Figure 13. Ground truth of DOAs'of the sequences. (**a**) seq01-1p-0000 (1 speaker), (b) seq37-3p-0001 (3 speakers) from AV16.3 corpus.

Table 5. The RMSE (m) in localization of the speaker(s) in the sequences *seq01-1p-0000 and seq37-3p-0001* of AV 16.3 database.

Technique	*seq01-1p-0000*	*seq37-3p-0001*		
	Sp1	*Sp1*	*Sp2*	*Sp3*
IDIR	0.58	0.81	0.24	0.41
GCF-D	0.67	1.01	0.27	0.51
Prop	0.21	0.31	0.12	0.15

was recorded in IDIAP meeting room of size 8.2 m × 3.6 m × 2.4 m using the two circular microphone arrays (MA1 & MA2) where the possible speaker location is limited to an L shape area of 2 m × 3 m, as shown in figure 12. The ground truth of direction of arrival (DOA) of a single active speaker in the sequences *seq01-1p-0000* (217 sec) and three speakers in the *seq37-3p-0001* (511 sec) in different frames (frame length = 1 sec, sampling frequency = 16 KHz) are shown in figure 13a and b respectively. These DOAs were converted to spatial coordinates to find the localization error of the corresponding location estimates. For enhanced spatial sensitivity, the proposed method and GCF-D are implemented, selecting four alternating microphones from each array, i.e., MA1 and MA2, whereas IDIR is implemented with MA1 for competitive comparison. The spatial localization error (RMSE) of each source is tabulated in separate columns in Table 5 for both the sequences. As indicated by results, the localization error of the speakers is in the order $Sp2 < Sp3 < Sp1$ for the sequence *seq37-3p-0001*. Further, the proposed method has performed superior to the other two existing methods for both the sequences.

5. Conclusions

This paper intends to provide a proficient TDOA based method for localizing multiple sources. The method resolves the TDOA association ambiguities of the multiple sources using volumetric mapping and eliminates the need to solve the complex hyperbolic equations for position estimation. The calculation of TDOA bounds of each subvolume, delay segmentation and subvolume clustering is done initially only and hence cut-off the required run-time computations. Moreover, the latter two steps together diminish the necessary subvolume scanning for delay mapping significantly and make the method computationally more efficient. The localization of the multiple sources is dramatically improved in adverse acoustic conditions by assigning equal weights to the mapped subvolumes for all the estimated delays. Moreover, the method shows enhanced localization results over the state-of-the-art techniques in adverse acoustic conditions. Finally, any desired resolution can be achieved by implementing C-SRP in the selected subvolume.

References

[1] Cobos M, Antonacci F, Alexandridis A, Mouchtaris A and Lee B 2017 A survey of sound source localization methods in wireless acoustic sensor networks. *Wirel. Commun. Mob. Comput.* 2017: 1–24. https://doi.org/10.1155/2017/3956282

[2] Li P and Ma X 2009 Robust acoustic source localization with TDOA based RANSAC algorithm. In: *Proceedings of Emergency Intelligent Computing Technology and Applications ICIC 2009. Lecture Notes Computer Science*, vol. 5754. Springer, Berlin, Heidelberg, pp. 222–227

[3] Argentieri S, Danès P and Souères P 2014 A survey on sound source localization in robotics: from binaural to array processing methods. < *hal-01058575* > 1–32. https://doi.org/10.1016/j.csl.2015.03.003

[4] Shen H, Ding Z, Dasgupta S and Zhao C 2014 Multiple source localization in wireless sensor networks based on time of arrival measurement. *IEEE Trans. Signal Process.* 62: 1938–1949. https://doi.org/10.1109/TSP.2014.2304433

[5] Benesty J 2000 Adaptive eigenvalue decomposition algorithm for passive acoustic source localization. *J. Acoust. Soc. Am.* 107: 384–391. https://doi.org/10.1121/1.428310

[6] Knapp C H and Carter G C 1976 The generalized correlation method for estimation of time delay. *IEEE Trans. Acoust.* 24: 320–327. https://doi.org/10.1109/tassp.1976.1162830

[7] Alameda-Pineda X and Horaud R 2012 Geometrically-constrained robust time delay estimation using non-coplanar microphone arrays. In: *Proceedings of 20th EUSIPCO*, Bucharest, Romania, pp. 1309–1313

[8] Hosseini M S, Rezaie A H and Zanjireh Y 2017 Time difference of arrival estimation of sound source using cross correlation and modified maximum likelihood weighting function. *Sci. Iran* 24: 3268–3279. https://doi.org/10.24200/sci.2017.4355

[9] Benesty J, Chen J and Huang Y 2004 Time-delay estimation via linear interpolation and cross correlation. *IEEE Trans. Speech Audio Process.* 12: 509–519. https://doi.org/10.1109/TSA.2004.833008

[10] Liu H, Yang B and Pang C 2017 Multiple sound source localization based on TDOA clustering and multi-path matching pursuit. In: *Proceedings of IEEE ICASSP*, New Orleans, LA, pp. 3241–3245

[11] Dmochowski J P, Benesty J and Affes S 2007 A generalized steered response power method for computationally viable source localization. *IEEE Trans. Audio, Speech Lang. Process.* 15: 2510–2526. https://doi.org/10.1109/TASL.2007.906694

[12] Sundar H, Sreenivas T V and Seelamantula C S 2018 TDOA-based multiple acoustic source localization without association ambiguity. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 26: 1976–1990. https://doi.org/10.1109/TASLP.2018.2851147

[13] Smith J O and Abel J S 1987 Closed-form least-squares source location estimation from range-difference measurements. *IEEE Trans. Acoust.* 35: 1661–1669. https://doi.org/10.1109/TASSP.1987.1165089

[14] Alameda-Pineda X and Horaud R 2014 A geometric approach to sound source localization from time-delay estimates. *IEEE Trans. Audio, Speech Lang. Process.* 22: 1082–1095. https://doi.org/10.1109/TASLP.2014.2317989

[15] Bestagini P, Compagnoni M, Antonacci F, Sarti A and Tubaro S 2014 TDOA-based acoustic source localization in the space–range reference frame. *Multidim. Syst. Signal Process.* 25: 337–359. https://doi.org/10.1007/s11045-013-0233-8

[16] Jin B, Xu X and Zhang T 2018 Robust time-difference-of-arrival (TDOA) localization using weighted least squares with cone tangent plane constraint. *Sensors* 18: 1–16. https://doi.org/10.3390/s18030778

[17] Kwon B, Park Y and Park Y S 2010 Analysis of the GCC-PHAT technique for multiple sources. In: *Proceedings of International Conference on Control, Automation and Systems (ICCAS),* Gyeonggi-do, pp. 2070–2073

[18] Claudio E D Di, Parisi R and Orlandi G 2000 Multi-source localization in reverberant environments by root-music and clustering. In: *Proceedings of IEEE ICASSP,* Istanbul, Turkey, pp. 921–924

[19] Lathoud G and Odobez J 2007 Short-term spatio–temporal clustering applied to multiple moving speakers. *IEEE Trans. Audio Speech Lang. Process.* 15: 1696–1710

[20] Hu J S, Yang C H and Wang C K 2009 Estimation of sound source number and directions under a multi-source environment. In: *Proceedings of IEEE/RSJ International Conference on Intelligent Robot System* 1: 181–186. https://doi.org/10.1109/IROS.2009.5354706

[21] Lee B and Choi J S 2010 Multi-source sound localization using the competitive K-means clustering. In: *Proceedings of 15th IEEE Conference on Emerging Technologies and Factory Automation,* Bilbao, pp 1–7

[22] Scheuing J and Yang B 2007 Efficient synthesis of approximately consistent graphs for acoustic multi-source localization. In: *Proceedings of IEEE ICASSP*, Honolulu, HI, pp 501–504

[23] Zannini C M, Cirillo A, Parisi R and Uncini A 2010 Improved TDOA disambiguation techniques for sound source localization in reverberant environments. In: *Proceedings of IEEE International Symposium on Circuits and Systems: Nano-Bio Circuit Fabrics and Systems,* Paris, pp. 2666–2669

[24] Yang B and Kreißig M 2013 A graph-based approach to assist TDOA based localization. In: *Proceedings of 8th International Workshop on Multidimensional System (nDS'13)*, Erlangen, Germany, pp. 75–80

[25] Levy A, Gannot S and Habets E A P 2011 Multiple-hypothesis extended particle filter for acoustic source localization in reverberant environments. *IEEE Trans. Audio Speech Lang. Process.* 19: 1540–1555. https://doi.org/10.1109/TASL.2010.2093517

[26] Zotkin D N and Duraiswami R 2004 Accelerated speech source localization via a hierarchical search of steered response power. *IEEE Trans. Speech Audio Process.* 12: 499–508. https://doi.org/10.1109/TSA2004.832990

[27] Çöteli M B, Olgun O and Hacihabiboğlu H 2018 Multiple sound source localization with steered response power density and hierarchical grid refinement. *IEEE/ACM Trans. Audio Speech Lang. Process.* 26:2215–2229. https://doi.org/10.1109/TASLP.2018.2858932

[28] Hadad E and Gannot S 2018 Multi-Speaker Direction of Arrival Estimation using SRP-PHAT Algorithm with a Weighted Histogram. In: *Proceedings of International Conference on the Science of Electrical Engineering*, Israel, pp. 1–5

[29] Brutti A, Omologo M, Member E and Svaizer P 2010 Multiple source localization based on acoustic map de-emphasis. *EURASIP J. Audio, Speech, Music Process* 2010: 1–17. https://doi.org/10.1155/2010/147495

[30] Lima M V S, Martins W A, Nunes L O, Biscainho L W P, Ferreira T N, Costa M V M and Lee B 2015 A volumetric SRP with refinement step for sound source localization. *IEEE Signal Process. Lett.* 22: 1098–1102. https://doi.org/10.1109/LSP.2014.2385864

[31] Cobos M, Marti A and Lopez J J 2010 A Modified SRP-PHAT Functional for Robust Real-Time Sound Source Localization With Scalable Spatial Sampling. *IEEE Signal Process. Lett.* 18: 71–74. https://doi.org/10.1109/lsp.2010.2091502

[32] Cobos M 2014 A note on the modified and mean-based steered-response power functionals for source localization in noisy and reverberant environments. In: *Proceedings of IEEE 6th International Symposium on Communication, Control and Signal Process,* Athens, pp. 149–152

[33] Lehmann E A and Johansson A M 2010 Diffuse reverberation model for efficient image-source simulation of room impulse responses. *IEEE Trans. Audio, Speech Lang. Process.* 18: 1429–1439. https://doi.org/10.1109/TASL.2009.2035038

[34] Lehmann E A and Johansson A M 2015 Prediction of energy decay in room impulse responses simulated with an image-source model. *J. Acoust. Soc. Am.* 124: 269–277. https://doi.org/10.1121/1.2936367

[35] Kabal P 2002 *TSP speech database*. McGill Univ, Database Version, pp. 1–39

[36] Fritts L 1997 University of Iowa musical instrument samples. In: *Univ. Iowa*. http://theremin.music.uiowa.edu/MIS.html

[37] Salvati D, Drioli C and Foresti G L 2017 Exploiting a geometrically sampled grid in the steered response power algorithm for localization improvement. *J. Acoust. Soc. Am.* 141: 586–601. https://doi.org/10.1121/1.4974289

[38] Habets Emanuel and Sommen P C W 2002 Optimal microphone placement for source localization using time delay estimation. In: *Proceedings of Workshop Circuits Systems and Signal Process. ProRISC*, pp. 284–287

[39] Brutti A, Omologo M and Svaizer P 2008 Localization of multiple speakers based on a two step acoustic map analysis. In: *Proceedings of IEEE ICASSP*, Las Vegas, pp. 4349–4352

[40] Nielsen J K, Jensen J R, Jensen S H and Christensen MG 2014 The single-and multichannel audio recordings database (SMARD). In: *Proceedings of 14th International Workshop Acoustic Signal Enhancement (IWAENC)*, pp. 40–44. https://doi.org/10.1109/IWAENC.2014.6953334

[41] Lathoud G, Odobez J and Gatica-Perez D 2004 AV16.3: An audio- visual corpus for speaker localization and tracking. In: *Proceedings of MLMI. Lecture Notes Computer Science*, vol. 3361. Springer, Berlin, Germany, pp. 182–195