

# Incremental Averaging Method to Improve Graph-Based Time-Difference-of-Arrival Estimation

Klaus Brümmer<sup>1</sup>, Kouei Yamaoka<sup>2</sup>, Nobutaka Ono<sup>3</sup>, Simon Doclo<sup>1</sup>

<sup>1</sup>Carl von Ossietzky Universität, Oldenburg, Germany   <sup>2</sup>University of Tokyo, Tokyo, Japan

<sup>3</sup>Tokyo Metropolitan University, Tokyo, Japan

**Abstract**—Estimating the position of a speech source based on time-differences-of-arrival (TDOAs) is often adversely affected by background noise and reverberation. A popular method to estimate the TDOA between a microphone pair involves maximizing a generalized cross-correlation with phase transform (GCC-PHAT) function. Since the TDOAs across different microphone pairs satisfy consistency relations, generally only a small subset of microphone pairs are used for source position estimation. Although the set of microphone pairs is often determined based on a reference microphone, recently a more robust method has been proposed to determine the set of microphone pairs by computing the minimum spanning tree (MST) of a signal graph of GCC-PHAT function reliabilities. To reduce the influence of noise and reverberation on the TDOA estimation accuracy, in this paper we propose to compute the GCC-PHAT functions of the MST based on an average of multiple cross-power spectral densities (CPSDs) using an incremental method. In each step of the method, we increase the number of CPSDs over which we average by considering CPSDs computed indirectly via other microphones from previous steps. Using signals recorded in a noisy and reverberant laboratory with an array of spatially distributed microphones, the performance of the proposed method is evaluated in terms of TDOA estimation error and 2D source position estimation error. Experimental results for different source and microphone configurations and three reverberation conditions show that the proposed method considering multiple CPSDs improves the TDOA estimation and source position estimation accuracy compared to the reference microphone- and MST-based methods that rely on a single CPSD as well as steered-response power-based source position estimation.

## 1. INTRODUCTION

In various speech communication applications such as videoconferencing and smart speakers, microphone arrays are used to localize speech sources in noisy and reverberant environments, more in particular estimate their direction-of-arrival or position [1]–[5]. Several model-based and learning-based methods have been proposed, e.g., methods using time-differences-of-arrival (TDOAs) [1], [6]–[11], the steered response power with phase transform (SRP-PHAT) method [12]–[15], subspace-based methods [16], [17], and deep neural network-based methods [18], [19]. In this paper, we focus on TDOA-based methods, where the source localization relies on prior estimation of the TDOAs. A few methods exist to estimate the TDOA between a given microphone pair [20], [21]. Here we use the widely adopted method [22] based on maximizing the generalized cross-correlation with phase transforms (GCC-PHAT) function. However, noise and reverberation may introduce additional peaks in the GCC-PHAT function, affecting the accuracy of the estimated TDOAs and therefore also the source localization [23].

Since consistency relations exist for TDOAs between multiple microphone pairs, generally only a small subset of microphone pairs, i.e., the minimal set [24], [25], are used for source localization, to reduce the likelihood of including TDOA outliers. A simple, commonly used method to determine the minimal set of TDOAs is to relate them to a common reference microphone [26], [27], where the choice of the

reference microphone may have a large impact on the reliability of the GCC-PHAT functions [7]. Using graph theory, a more robust method has recently been proposed in [28] to determine the optimal minimal set in terms of GCC-PHAT reliability. This minimal set is determined as the minimum spanning tree (MST) of the signal graph and does not require choosing a common reference microphone. Nevertheless, even though the MST method has been shown to be more robust in terms of avoiding TDOA outliers, strong noise or reverberation in one or more microphones may mean that it is not possible to avoid TDOA outliers in the minimal set of microphone pairs.

Instead of computing each GCC-PHAT function for each microphone pair in the MST based on a single cross-power spectral density (CPSD), in this paper, we propose an incremental method to average over multiple CPSDs from multiple microphone pairs. Addressing a different microphone pair of the MST-based minimal set in each step, we order the microphones based on their GCC-PHAT reliabilities and incrementally re-estimate the TDOAs using the averaged CPSDs. In each step, we propose to incorporate an additional CPSD, estimated indirectly via another microphone which was used in previous steps, which generally has a similar, or more reliable GCC-PHAT function. To include these additional CPSDs into the average requires phase alignment, which is accomplished using a phase shift based on TDOAs which were re-estimated in previous steps.

The performance of the proposed TDOA estimation method is compared to baseline reference microphone- and MST-based methods, as well as the steered-response power-based position estimation, in terms of TDOA estimation error and source 2D position estimation error. Based on recorded noisy and reverberant speech signals, experimental results for three reverberation conditions demonstrate that the proposed TDOA estimation method outperforms all of the considered baseline methods in a range of reverberant environments, demonstrating that it is beneficial to base TDOA estimation on an average of multiple CPSDs instead of just a single CPSD, which is possible through our incremental method.

## 2. STATE-OF-THE-ART TDOA ESTIMATION

We consider a noisy and reverberant acoustic environment with a single speech source at position  $\mathbf{p}$  and a microphone array with  $M$  microphones at positions  $[\mathbf{m}_1, \dots, \mathbf{m}_M] \in \mathbb{R}^{P \times M}$ , where  $P$  is the dimensionality of the acoustic scenario. Assuming synchronized microphones and free-field transmission, i.e., no objects between the source and the microphones, the TDOA between microphones  $i$  and  $j$  is given by  $\tau_{i,j}(\mathbf{p}) = (\|\mathbf{p} - \mathbf{m}_i\|_2 - \|\mathbf{p} - \mathbf{m}_j\|_2) / \nu$ , where  $\nu$  denotes the speed of sound. The  $M \times M$ -dimensional TDOA matrix, containing the TDOAs between all microphone pairs, is defined as  $\mathbf{T}(\mathbf{p}) = [\tau_{i,j}(\mathbf{p})]$ , and is an anti-symmetric matrix with rank 2 [25]. Since the TDOAs across different microphone pairs satisfy consistency relations (i.e.,  $\tau_{i,j} = \tau_{i,m} - \tau_{j,m}, \forall m$ ), it has been shown in [25], [29] that TDOA matrices can be written as

$$\mathbf{T}(\mathbf{p}) = \boldsymbol{\tau}_m(\mathbf{p}) \mathbf{1}_M^T - \mathbf{1}_M \boldsymbol{\tau}_m^T(\mathbf{p}) \quad \forall m, \quad (1)$$

This work was funded by the German Academic Exchange Service (DAAD), the Federal Ministry of Education and Research (BMBF) - Project ID 57711070, and the German Research Foundation (DFG) under Germany's Excellence Strategy - EXC 2177/1 - Project ID 390895286.

with  $\tau_m(\mathbf{p}) = [\tau_{1,m}(\mathbf{p}), \dots, \tau_{M,m}(\mathbf{p})]^T$  the  $M$ -dimensional vector of TDOAs relative to the  $m$ -th reference microphone (with  $\tau_{m,m}(\mathbf{p}) = 0$ ) and  $\mathbf{1}_M$  an  $M$ -dimensional vector of ones.

A commonly used method to estimate the TDOA  $\tau_{i,j}(\mathbf{p})$  is based on the GCC-PHAT function [22] between microphones  $i$  and  $j$ , defined as

$$\xi_{i,j}(\tau) = \int_{-\infty}^{\infty} \phi_{i,j}(\omega) \exp(j\omega\tau) d\omega, \quad (2)$$

with imaginary number  $j = \sqrt{-1}$ , radial frequency  $\omega$  and time lag  $\tau$ . The phase transform (PHAT)-weighted CPSD  $\phi_{i,j}(\omega)$  is given by

$$\phi_{i,j}(\omega) = \frac{\psi_{i,j}(\omega)}{|\psi_{i,j}(\omega)|}, \quad (3)$$

where  $\psi_{i,j}(\omega) = \mathbb{E}\{Y_i(\omega)Y_j^*(\omega)\}$  denotes the CPSD between the microphones  $i$  and  $j$ ,  $Y_m(\omega)$  denotes the  $m$ -th microphone signal in the continuous-time Fourier transform domain, and  $\mathbb{E}\{\cdot\}$  denotes the expectation operator. The estimated TDOA  $\tilde{\tau}_{i,j}$  is computed as the time lag that maximizes  $\xi_{i,j}(\tau)$ , i.e.,

$$\tilde{\tau}_{i,j} = \underset{\tau}{\operatorname{argmax}} \xi_{i,j}(\tau). \quad (4)$$

Due to noise and reverberation, the GCC-PHAT functions may exhibit additional peaks, which result in TDOA estimation errors if they are higher than the peak corresponding to the direct path signal [23]. Due to estimation errors, the resulting estimated TDOA matrix is not guaranteed to be consistent due to estimation errors. To avoid discrepancies between inconsistent estimated TDOAs, various methods have proposed to use only the estimated TDOAs between a minimal set of  $M-1$  microphone pairs, from which a consistent estimated TDOA matrix can be constructed. In the following subsections, we will discuss two classes of methods to determine the minimal set of microphone pairs, either based on graph theory in Section 2.1 or using a common reference microphone in Section 2.2.

### 2.1. Minimal Set Based on MST

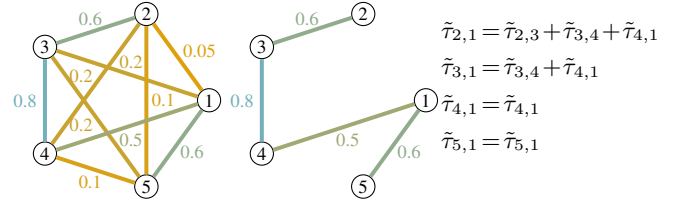
Using graph theory, a method was recently proposed in [28] to determine the minimal set of microphone pairs based on GCC-PHAT reliability. The undirected signal graph  $G = (\mathcal{M}, \Theta)$  is characterized by the set of vertices  $\mathcal{M} = \{1, \dots, M\}$ , representing the microphones, and the set of edges  $\Theta = \{i, j | i, j \in \mathcal{M} \cap i \neq j\}$  (corresponding to the cost, defined as the negative reliability). Similarly as in [28], reliability is defined in this paper as the maximum value of the GCC-PHAT function (2), i.e.,

$$R_{i,j} = \xi_{i,j}(\tilde{\tau}_{i,j}). \quad (5)$$

In the MST method, the minimal set of microphone pairs is determined by the edges  $\Theta^{\text{MST}} \subseteq \Theta$ , which are chosen by the Prim method [30] such that  $G^{\text{MST}} = (\mathcal{M}, \Theta^{\text{MST}})$  forms a spanning tree and the total negative reliability, i.e.,  $\sum_{i,j \in \Theta^{\text{MST}}} -R_{i,j}$  is minimized. As described in [28], the set of TDOAs corresponding to these microphone pairs is then rewritten relative to an arbitrarily chosen reference microphone in a TDOA vector  $\tilde{\tau}_m$ , which is generally how they are used for TDOA-based position estimation. An exemplary MST and the corresponding set of TDOAs are shown in Fig. 1.

### 2.2. Minimal Set Based on Common Reference Microphone

A commonly used minimal set of microphone pairs for TDOA estimation is the set of microphone pairs relative to a selected reference microphone. In this paper, we consider three baseline methods for choosing the reference microphone index  $m$ . In the arbitrary method (Ref-A), the reference microphone is chosen randomly, i.e.,  $\hat{m}_A \sim \mathcal{U}(\{1, 2, \dots, M\})$ . In the centroid method (Ref-C),



**Fig. 1:** Left: Exemplary graph of GCC-PHAT reliability values between  $M=5$  microphones. Middle: Corresponding minimum spanning tree. Right: Corresponding TDOAs, rewritten relative to the first reference microphone.

the reference microphone is chosen as the microphone that is closest to the centroid of the microphone array as in [1], i.e.,  $\hat{m}_C = \operatorname{argmin}_m \|\mathbf{m}_m - \frac{1}{M} \sum_{i=1}^M \mathbf{m}_i\|_2$ . In the reliability-based method (Ref-R), the reference microphone is chosen based on the GCC-PHAT reliability. Rather than selecting the reference microphone based on the average GCC-PHAT reliability between this microphone and all other microphones, we propose to select the reference microphone for which the minimum GCC-PHAT reliability between this microphone and all other microphones is highest, i.e.,

$$\hat{m}_R = \underset{m}{\operatorname{argmax}} \{ \underset{i \neq m}{\operatorname{argmin}} R_{i,m} \}. \quad (6)$$

This reduces the risk that a highly erroneous estimated TDOA is included in the minimal set.

## 3. TDOA ESTIMATION BY INCREMENTALLY AVERAGING MULTIPLE CPSDS

Although the MST-based minimal set of microphone pairs reduces the chances of including a TDOA outlier compared to other minimal sets, it should be realized that each TDOA is computed from a single PHAT-weighted CPSD. Therefore, including microphone pairs with TDOA outliers may be unavoidable when one or more microphones are subject to high levels of noise and reverberation. To further improve the accuracy of the estimated TDOAs, in this section we propose an incremental method to re-estimate the TDOAs corresponding to the MST based on an average of multiple CPSDs from multiple microphone pairs. This averaging of CPSDs is based on a directly computed CPSD and indirectly estimated CPSDs which are computed via other microphones from previous steps. After introducing indirect CPSD estimation in Section 3.1, in Section 3.2 we describe the proposed incremental method for averaging CPSDs.

### 3.1. Indirect CPSD Computation

Considering only the direct source component (i.e., in noise-free and anechoic conditions) the CPSD between microphones  $i$  and  $j$  can be written as  $\psi_{i,j}(\omega) = \exp(-j\omega\tau_{i,j}(\mathbf{p})) / (16\pi^2 d_i d_j)$ , with  $d_i = \|\mathbf{p} - \mathbf{m}_i\|_2$  the distance between the source and the  $i$ -th microphone [2], [3], [5]. Since for any microphone  $k$  it can be easily shown that  $\exp(-j\omega\tau_{i,j}(\mathbf{p})) = \exp(-j\omega\tau_{i,k}(\mathbf{p})) \exp(j\omega\tau_{j,k}(\mathbf{p}))$ , the CPSD between microphones  $i$  and  $j$  can be indirectly computed by applying a phase shift to the CPSD between microphones  $i$  and  $k$  and compensating for the distance-related attenuation, similarly to [31], i.e.,

$$\psi_{i,j}(\omega) = \frac{d_k}{d_j} \psi_{i,k}(\omega) \exp(j\omega\tau_{j,k}(\mathbf{p})). \quad (7)$$

In practice, the distances  $d_j$  and  $d_k$  are of course unavailable and are difficult to estimate. Since we are mainly interested in the phase, we assume  $d_j = d_k$ , and approximate the indirectly computed CPSD as

$$\tilde{\psi}_{i,j}^{[k]}(\omega) = \psi_{i,k}(\omega) \exp(j\omega\tau_{j,k}(\mathbf{p})) \quad (8)$$

### 3.2. Incremental Method for Averaging CPSDs

Based on the vertices and edges of the MST, the proposed method builds up a consistent TDOA matrix with re-estimated TDOAs from scratch in  $M-1$  steps, where  $H$  denotes the step index. To improve the estimation of CPSDs corresponding to less reliable GCC-PHAT functions in later steps, we begin with CPSDs corresponding to the highest GCC-PHAT reliabilities. Therefore, before introducing the method, the edges are ordered based on reliability for which we define an  $M-1$ -dimensional vector  $\mathbf{r}$  of reliabilities and a corresponding  $(M-1) \times 2$ -dimensional matrix  $\mathbf{V}$  containing  $M-1$  rows of MST edges  $[i, j]$ ,  $i, j \in \Theta^{\text{MST}}$ . The first entry of  $\mathbf{r}$  contains the highest reliability value of the MST and the first row of  $\mathbf{V}$  contains the pair of vertices connected by the associated edge. Subsequent entries of  $\mathbf{r}$  can only correspond to an edge connected to one of the vertices from previous rows of  $\mathbf{V}$ , in order of descending reliability. For the example in Fig. 1, the reliability vector would be  $\mathbf{r}_{\text{MST}} = [0.8, 0.6, 0.5, 0.6]^T$  and the rows of  $\mathbf{V}$  would be  $[4, 3]$ ,  $[3, 2]$ ,  $[4, 1]$ , and  $[5, 1]$ . The indices  $i[H]$ ,  $j[H]$  refer to the microphones in the  $H$ -th row of  $\mathbf{V}$ . In each step  $H$ , the proposed method (referred to as MST+) re-estimates the TDOA between the microphones  $i[H]$  and  $j[H]$ .

In the first step ( $H = 1$ ), the TDOA between  $i(1)$  and  $j(1)$  is estimated based on the same GCC-PHAT function (2) as used for computing the reliability in (5). In other words, no TDOA re-estimation is performed. In subsequent steps ( $H > 1$ ), we propose to re-estimate the TDOAs by incorporating knowledge from previous steps. Considering noisy and reverberant speech signals, we assume that the noise and reverberation destructively interfere across multiple microphone pairs, in contrast to the direct path component. Therefore, we propose to average out the spurious GCC-PHAT function peaks, corresponding to TDOA outliers, by averaging over the directly computed CPSD  $\psi_{i[H],j[H]}(\omega)$  and  $H-1$  indirectly computed CPSDs via all microphones used in previous steps. As such, the PHAT-weighted CPSD between microphones  $i[H]$  and  $j[H]$  is computed as

$$\tilde{\phi}'_{i[H],j[H]}(\omega) = \frac{\psi_{i[H],j[H]}(\omega) + \sum_{h=1}^{H-1} \tilde{\psi}_{i[H],j[H]}^{[j[h]]}(\omega)}{|\psi_{i[H],j[H]}(\omega) + \sum_{h=1}^{H-1} \tilde{\psi}_{i[H],j[H]}^{[j[h]]}(\omega)|} \quad (9)$$

where, based on (8), the CPSDs in (9) are computed using TDOAs  $\tilde{\tau}'_{j[h],j[h]}$  which were re-estimated in previous steps  $j[h]$  as

$$\tilde{\psi}_{i[H],j[H]}^{[j[h]]}(\omega) = \psi_{i[H],j[h]}(\omega) \exp(j\omega \tilde{\tau}'_{j[h],j[h]}), \quad 1 \leq h < H. \quad (10)$$

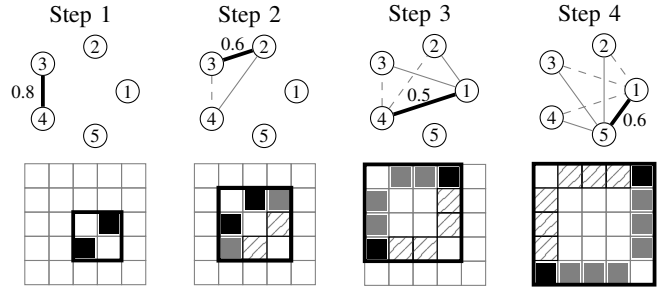
By incorporating CPSDs indirectly estimated via microphones corresponding to edges with generally higher GCC-PHAT reliabilities, it is expected that, in addition to mitigating spurious peaks corresponding to TDOA outliers, also the TDOA estimation accuracy can be improved. The TDOA  $\tilde{\tau}'_{i[H],j[H]}$  is then re-estimated as the time lag maximizing the GCC-PHAT function in (2) using the PHAT-weighted CPSD in (9). To use the re-estimated TDOAs for the phase alignment of the indirectly estimated CPSDs in (10) in subsequent steps, we re-estimate additional TDOAs for each of the microphones used in previous steps  $j[h]$  as

$$\tilde{\tau}'_{i[h],j[h]} = \tilde{\tau}'_{i[h],j[h]} - \tilde{\tau}'_{j[h],j[h]}, \quad 1 \leq h < H. \quad (11)$$

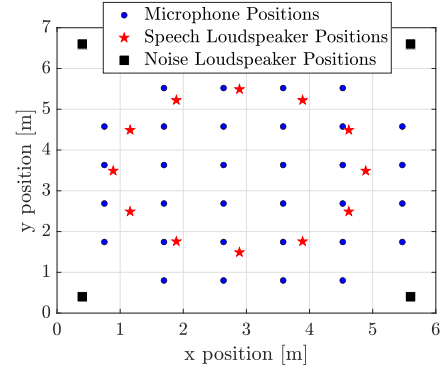
These steps are repeated until all TDOAs corresponding to the MST have been re-estimated. For the example in Fig. 1, Fig. 2 shows graphs and the relevant entries of the estimated TDOA matrix.

### 4. EXPERIMENTAL EVALUATION

Using real-world noisy and reverberant speech signals recorded using an array of spatially distributed microphones, in this section we



**Fig. 2:** Graphs and TDOA matrices for all steps of the proposed incremental TDOA re-estimation method, corresponding to the example in Fig. 1. The black lines and boxes correspond to the microphone pair  $i[H]$ ,  $j[H]$  used in step  $H$ , the solid gray lines and boxes correspond to the microphone pairs used for indirect CPSD estimation and the dashed gray lines and boxes correspond to microphone pairs used for phase alignment using TDOAs re-estimated in previous steps (i.e., entries within the black border from a previous step).



**Fig. 3:** Layout of the distributed microphones (●) and speech loudspeakers (★) in the BRUDEX laboratory

compare the performance of the proposed MST+ method with the baseline MST and reference microphone-based TDOA estimation methods and the steered-response power method, both in terms of TDOA estimation error as well as source position estimation error. Section 4.1 outlines the acoustic scenario, in Section 4.2, we discuss practical considerations for the implementation of the considered methods, and in Section 4.3 we discuss the experimental results.

#### 4.1. Acoustic Scenario

In our experiments, we used noisy and reverberant signals from the BRUDEX database [32], where the speech source and the noise were recorded separately by  $M=6$  distributed microphones in a laboratory with dimensions of about  $6 \text{ m} \times 7 \text{ m} \times 2.7 \text{ m}$ , whose layout can be seen in Fig. 3. A 10 s recorded speech signal randomly chosen from one of the three available recorded signals was considered for 120 different static source-microphone configurations (comprising 10 random microphone array configurations for each of the 12 available source positions) at a height of approximately 1.5 m. We considered three reverberation levels, corresponding to low ( $T_{60} \approx 310 \text{ ms}$ ), medium ( $T_{60} \approx 510 \text{ ms}$ ), and high ( $T_{60} \approx 1300 \text{ ms}$ ) reverberation. Diffuse-like babble noise produced by loudspeakers in the corners of the laboratory was added to the recorded speech signals at a signal-to-noise ratio of 5 dB averaged across the microphones.

#### 4.2. Practical Implementation

The algorithms were implemented with a sampling frequency  $f_s = 16 \text{ kHz}$  and a short-time Fourier transform (STFT) framework was used with a frame length of 1024 samples (corresponding to 64 ms), 50% overlap between frames, and a square-root-Hann analysis window.

The CPSDs between the microphone signals were calculated in the STFT domain using recursive smoothing as

$$\psi_{i,j}[k,l] = \lambda\psi_{i,j}[k,l-1] + (1-\lambda)Y_i[k,l]Y_j^*[k,l], \quad (12)$$

where  $k$  denotes the frequency bin index, and  $l$  denotes the time frame index. We used a recursive smoothing factor  $\lambda=0.98$ , corresponding to 1.58 s. The PHAT-weighted CPSD was computed as  $\phi_{i,j}[k,l] = \psi_{i,j}[k,l]/|\psi_{i,j}[k,l]|$ . The GCC-PHAT functions (based on an average of multiple CPSDs) were then computed using the inverse discrete Fourier transform for discrete time lags  $n=\tau f_s$ , i.e.,

$$\xi_{i,j}[n,l] = \frac{1}{(K-1)} \sum_{k=1}^{K-1} \phi_{i,j}[k,l] \exp\left(\frac{j2\pi nk}{K}\right), \quad (13)$$

where  $K$  denotes the DFT length. To achieve a more precise TDOA estimate, the GCC-PHAT functions were interpolated with an upsampling factor  $R=10$ . We only considered physically realistic time-lags, i.e.,  $n_{i,j}^{\min} \leq n_{i,j} \leq n_{i,j}^{\max}$ , with  $n_{i,j}^{\max} = -n_{i,j}^{\min} = Rf_s D_{i,j}/\nu$ , with  $D_{i,j} = \|\mathbf{m}_i - \mathbf{m}_j\|_2$  denoting the distance between the microphones  $i$  and  $j$ . The TDOA estimate for the  $l$ -th time frame was then obtained as  $\hat{\tau}_{i,j}[l] = \hat{n}_{i,j}[l]/(Rf_s)$ .

To evaluate the usability of the estimated TDOAs from different methods for source position estimation, we used the spherical (SI) interpolation-based source position estimation method [1]. In each time-frame, the SI cost function (constrained to a distance of 5 m from the centre of the room) was iteratively solved with gradient descent. For comparison, we also considered the SRP-PHAT method used in [11], i.e., the SRP-PHAT functional was evaluated on a 2D grid, first at a resolution of 10 cm, then at a resolution of 1 cm in the vicinities of the three grid points with the highest SRP-PHAT functional values, to find the global functional maximum.

### 4.3. Comparison of TDOA and Position Estimation Performance

To evaluate the TDOA estimation performance of the considered methods, we used the mean TDOA estimation error

$$\bar{\sigma} = \frac{1}{S(M-1)} \sum_{s=1}^S \sum_{m=2}^M |\hat{\tau}_{m,1}[s] - \tau_{m,1}(\mathbf{p}[s])|, \quad (14)$$

over all snapshots (time frames with active speech)  $s=1, \dots, S$ , for each 10 s signal and all 120 source-microphone configurations (noting that the source position  $\mathbf{p}[s]$  was static for each configuration). For the SRP-PHAT method, the TDOA estimation error was computed based on the TDOAs corresponding to the estimated source position. In addition, to evaluate the usability of the estimated TDOAs for SI-based source position estimation, we considered measures based on the position estimation error  $\varepsilon[s] = \|\hat{\mathbf{p}}[s] - \mathbf{p}[s]\|_2$ , i.e., the mean position estimation error over snapshots  $\bar{\varepsilon} = \frac{1}{S} \sum_{s=1}^S \varepsilon[s]$  and the accuracy  $\text{Acc} = \frac{1}{S} \sum_{s=1}^S \mathcal{X}(\varepsilon[s] \leq 10 \text{ cm})$ , using the indicator function  $\mathcal{X}(\cdot)$ , which is equal to 1 if  $\varepsilon[s] \leq 10 \text{ cm}$  and 0 otherwise.

For all considered TDOA estimation methods, Table 1 shows the TDOA and source position estimation errors for the three considered reverberation levels. As can be observed, the SRP-PHAT method results in a relatively poor source position estimation (and TDOA estimation) performance (e.g.,  $\bar{\sigma} = 0.28 \text{ ms}$ ,  $\bar{\varepsilon} = 29.3 \text{ cm}$ , and  $\text{Acc} = 63.0 \%$  in medium reverberation). This is likely because the SRP-PHAT method has not been tested for such random source and microphone configurations, where it may not be ideal to weight the CPSD from each microphone pair equally within the SRP-PHAT functional. Considering the TDOA-based methods, randomly choosing the reference microphone (Ref-A method) also results in a relatively poor TDOA and source position estimation performance (e.g.,  $\bar{\sigma} = 1.34$

**Table 1:** Mean TDOA estimation error  $\bar{\sigma}$ , mean position estimation error  $\bar{\varepsilon}$  and position estimation accuracy  $\text{Acc}$  for the SRP-PHAT method, three reference microphone-based methods (Ref-A, Ref-C, Ref-R), the baseline MST method, and the proposed MST+ method for three reverberation conditions.

Reverb.	Error Metric	TDOA Estimation Method					
		SRP-PHAT	Ref-A	Ref-C	Ref-R	MST	MST+
Low	$\bar{\sigma}$ [ms]	0.18	0.49	0.16	0.04	0.04	<b>0.01</b>
	$\bar{\varepsilon}$ [cm]	20.0	28.1	8.5	2.9	2.9	<b>0.7</b>
	Acc [%]	69.5	81.1	93.9	98.2	98.3	<b>99.8</b>
Med.	$\bar{\sigma}$ [ms]	0.28	1.34	0.54	0.14	0.12	<b>0.08</b>
	$\bar{\varepsilon}$ [cm]	29.3	70.9	28.6	9.0	8.9	<b>5.0</b>
	Acc [%]	63.0	54.3	77.1	92.9	92.8	<b>95.8</b>
High	$\bar{\sigma}$ [ms]	0.44	1.77	0.80	0.44	0.40	<b>0.28</b>
	$\bar{\varepsilon}$ [cm]	37.9	90.1	46.1	24.3	20.3	<b>10.5</b>
	Acc [%]	59.6	44.2	62.7	82.2	85.0	<b>92.4</b>

ms,  $\bar{\varepsilon} = 70.9 \text{ cm}$ , and  $\text{Acc} = 54.3 \%$  in medium reverberation). The results of the Ref-C method show that the TDOA and source position estimation performance can be easily improved by simply choosing the microphone closest to the centroid of the microphone array as the reference microphone (e.g.,  $\bar{\sigma} = 0.54 \text{ ms}$ ,  $\bar{\varepsilon} = 28.6 \text{ cm}$ , and  $\text{Acc} = 77.1 \%$  in medium reverberation). The reliability-based Ref-R and MST methods outperform the SRP-PHAT, Ref-A and Ref-C methods, showing the importance of taking into account the reliability of the GCC-PHAT functions for determining the minimal set of microphone pairs. In low and medium reverberation conditions, the Ref-R and MST methods perform very similarly. However, in high reverberation the MST method performs slightly better than the Ref-R method ( $\bar{\sigma} = 0.40 \text{ ms}$ ,  $\bar{\varepsilon} = 20.3 \text{ cm}$ , and  $\text{Acc} = 85.0 \%$  compared to  $\bar{\sigma} = 0.44 \text{ ms}$ ,  $\bar{\varepsilon} = 24.3 \text{ cm}$ , and  $\text{Acc} = 82.2 \%$ ), most likely because in the MST method no reference microphone needs to be chosen. The results in Table 1 clearly show that the proposed MST+ method outperforms all other considered methods in every reverberation condition (e.g.,  $\bar{\sigma} = 0.08 \text{ ms}$ ,  $\bar{\varepsilon} = 5.0 \text{ cm}$ , and  $\text{Acc} = 95.8 \%$  in medium reverberation). This suggests that it is beneficial to average over multiple CPSDs from multiple microphone pairs to accurately estimate TDOAs and improve source position estimation performance, rather than estimating each TDOA based on a single CPSD.

## 5. CONCLUSIONS

In this paper, we have proposed an incremental method, improving the time-difference of arrival (TDOA) estimation based on an average over multiple cross-power spectral densities (CPSDs) from multiple microphone pairs, compared to reference microphone- and minimum spanning tree (MST)-based methods which rely on a single CPSD. This method re-estimates the TDOAs corresponding to the MST of generalized cross-correlation with phase transform (GCC-PHAT) function reliabilities, in multiple steps, beginning with the edges with highest GCC-PHAT reliabilities. In each step, we incorporate an additional CPSD, estimated indirectly via another microphone from a previous step. Including the indirectly estimated CPSD requires a phase-alignment, which we achieve using a phase shift based on re-estimated TDOAs from previous steps. Based on noisy and reverberant speech signals recorded in a laboratory with an array of spatially distributed microphones, we evaluated the performance of the different methods in terms of TDOA estimation error and source position estimation error, for three reverberation conditions. Experimental results for different source and microphone configurations demonstrate that the proposed method considerably improves the TDOA and source position estimation performance compared to existing reference microphone-, MST-based, and steered-response power-based methods.

## REFERENCES

- [1] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone arrays: signal processing techniques and applications*, M. Brandstein and D. Ward, Eds. Springer, 2001, pp. 157–180.
- [2] Y. A. Huang, J. Benesty, and J. Chen, "Time delay estimation and source localization," in *Springer Handbook of Speech Processing*, J. Benesty, Y. A. Huang, and M. Christensen, Eds. Springer, 2008, pp. 1043–1063.
- [3] N. Madhu, R. Martin, U. Heute, and C. Antweiler, "Acoustic source localization with microphone arrays," in *Advances in Digital Speech Transmission*, R. Martin, U. Heute, and C. Antweiler, Eds. Chichester, UK: Wiley, 2008, pp. 135–170.
- [4] A. Brutti, M. Omologo, and P. Svaizer, "Multiple source localization based on acoustic map de-emphasis," *EURASIP J. Audio, Speech & Music Process.*, vol. 2010, pp. 1–17, 2010.
- [5] P. Pertilä, A. Brutti, P. Svaizer, and M. Omologo, "Multichannel source activity detection, localization, and tracking," in *Audio source separation and speech enhancement*, E. Vincent, T. Virtanen, and S. Gannot, Eds. Wiley, 2018, pp. 47–64.
- [6] Y. T. Chan and K. C. Ho, "A simple and efficient estimator for hyperbolic location," *IEEE Trans. Signal Process.*, vol. 42, no. 8, pp. 1905–1915, 1994.
- [7] T. Gustafsson, B. D. Rao, and M. Trivedi, "Source localization in reverberant environments: Modeling and statistical analysis," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 791–803, 2003.
- [8] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereau, "Real-time passive source localization: A practical linear-correction least-squares approach," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 9, no. 8, pp. 943–956, 2001.
- [9] J. Yang, X. Zhong, W. Chen, and W. Wang, "Multiple acoustic source localization in microphone array networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 334–347, 2020.
- [10] C. Evers, H. W. Löllmann, H. Mellmann, A. Schmidt, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA challenge: Acoustic source localization and tracking," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1620–1643, 2020.
- [11] K. Brümman and S. Doclo, "3D single source localization based on Euclidean distance matrices," in *Proc. IEEE International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Bamberg, Germany, 2022, pp. 1–5.
- [12] M. Cobos, A. Marti, and J. J. Lopez, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling," *IEEE Signal Process. Lett.*, vol. 18, no. 1, pp. 71–74, 2010.
- [13] L. O. Nunes, W. A. Martins, M. V. S. Lima, L. W. P. Biscainho, M. V. M. Costa, F. M. Gonçalves, A. Said, and B. Lee, "A steered-response power algorithm employing hierarchical search for acoustic source localization using microphone arrays," *IEEE Trans. Signal Process.*, vol. 62, no. 19, pp. 5171–5183, 2014.
- [14] T. Dietzen, E. De Sena, and T. van Waterschoot, "Low-complexity steered response power mapping based on Nyquist-Shannon sampling," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2021, pp. 206–210.
- [15] E. Grinstein, E. Tengan, B. Çakmak, T. Dietzen, L. Nunes, T. van Waterschoot, M. Brookes, and P. A. Naylor, "Steered response power for sound source localization: A tutorial review," *EURASIP J. Audio, Speech & Music Process.*, vol. 2024, no. 1, p. 59, 2024.
- [16] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.
- [17] J. P. Dmochowski, J. Benesty, and S. Affes, "Broadband MUSIC: Opportunities and challenges for multiple source localization," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY, USA: IEEE, 2007, pp. 18–21.
- [18] E. Grinstein, C. M. Hicks, T. van Waterschoot, M. Brookes, and P. A. Naylor, "The neural-SRP method for universal robust multi-source tracking," *IEEE Open J. Signal Process.*, vol. 5, pp. 19–28, 2023.
- [19] R. Varzandeh, S. Doclo, and V. Hohmann, "Improving multi-talker binaural DOA estimation by combining periodicity and spatial features in convolutional neural networks," *EURASIP J. Audio, Speech & Music Process.*, vol. 2025, no. 1, p. 5, 2025.
- [20] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Am.*, vol. 107, no. 1, pp. 384–391, 2000.
- [21] T. G. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Signal Process.*, vol. 85, no. 1, pp. 177–204, 2005.
- [22] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, 1976.
- [23] G. García-Barrios, E. L. Iglesias, J. M. Gutiérrez-Arriola, R. Fraile, N. Sáenz-Lechón, and V. J. Osma-Ruiz, "Exploiting spatial diversity for increasing the robustness of sound source localization systems against reverberation," *Appl. Acoust.*, vol. 202, p. 109138, 2023.
- [24] J. Scheuing and B. Yang, "Correlation-based TDOA-estimation for multiple sources in reverberant environments," in *Speech and Audio Processing in Adverse Environments*. Springer, 2008, pp. 381–416.
- [25] J. Velasco, D. Pizarro, J. Macias-Guarasa, and A. Asaei, "TDOA matrices: Algebraic properties and their application to robust denoising with missing data," *IEEE Trans. Signal Process.*, vol. 64, no. 20, pp. 5242–5254, 2016.
- [26] A. Canclini, P. Bestagini, F. Antonacci, M. Compagnoni, A. Sarti, and S. Tubaro, "A robust and low-complexity source localization algorithm for asynchronous distributed microphone networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 10, pp. 1563–1575, 2015.
- [27] T.-K. Le and T.-H. Le, "Rank properties for matrices constructed from time differences of arrival," *IEEE Trans. Signal Process.*, vol. 66, no. 13, pp. 3491–3503, 2018.
- [28] K. Yamaoka, T. Nakashima, Y. Wakabayashi, and N. Ono, "Minimum-spanning-tree-based time delay estimation robust to outliers," *IEEE Access*, vol. 11, pp. 121 284–121 294, 2023.
- [29] H. C. So, Y. T. Chan, and F. K. W. Chan, "Closed-form formulae for time-difference-of-arrival estimation," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2614–2620, 2008.
- [30] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, pp. 269–271, 1959.
- [31] K. Brümman and S. Doclo, "Steered response power-based direction-of-arrival estimation exploiting an auxiliary microphone," in *Proc. European Signal Processing Conference (EUSIPCO)*. Lyon, France: IEEE, 2024, pp. 917–921.
- [32] D. Fejgin, W. Middelberg, and S. Doclo, "BRUDEX database: Binaural room impulse responses with uniformly distributed external microphones," in *Proc. ITG Conference on Speech Communication*, Aachen, Germany, 2023, pp. 126–130.