3D SINGLE SOURCE LOCALIZATION BASED ON EUCLIDEAN DISTANCE MATRICES

Klaus Brümann, Simon Doclo

Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4all,
University of Oldenburg, Germany

2022 International Workshop on Acoustic Signal Enhancement (IWAENC) | 978-1-6654-6867-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/IWAENC53105.2022.9914726

ABSTRACT

A popular approach for 3D source localization using multiple microphones is the steered-response power method, where the source position is directly estimated by maximizing a function of three continuous position variables. Instead of directly estimating the source position, in this paper we propose an indirect, distance-based method for 3D source localization. Based on properties of Euclidean distance matrices (EDMs), we reformulate the 3D source localization problem as the minimization of a cost function of a single variable, namely the distance between the source and the reference microphone. Using the known microphone geometry and estimated time-differences of arrival (TDOAs) between the microphones, we show how the 3D source position can be computed based on this variable. In addition, instead of using a single TDOA estimate per microphone pair, we propose an extension that enables to select the most appropriate estimate from a set of candidate TDOA estimates, which is especially relevant in reverberant environments with strong early reflections. Experimental results for different source and microphone constellations show that the proposed EDM-based method consistently outperforms the steered-response power method, especially when the source is close to the microphones.

Index Terms— Source localization, Euclidean distance matrix, Gram matrix, rank, time-difference of arrival

1. INTRODUCTION

The location of a speech source, relative to some microphones (e.g., in mobile phones or hearing aids), is a widely used spatial feature for speech enhancement or speaker extraction. Often, the localization constitutes estimating the source direction of arrival using compact microphone arrays, where it can be assumed that the source is in the far field. In this paper, we focus on 3D localization, where the far field assumption is not made, i.e., using spatially distributed microphones of an acoustic sensor network.

Source localization methods [1–4] can be broadly categorized into direct (one-step) and indirect (two-step) approaches. The steered-response power with phase transform (SRP-PHAT) method [1] is a direct approach, which exploits the generalized cross-correlations [5] between all microphone pairs and can be interpreted as a delay-and-sum beamformer, steered towards all possible 3D source positions, and has gained much popularity due to its robustness against noise and reverberation. A drawback is that it requires the optimization of three continuous position variables, for which in practice a discrete 3D grid search is used. Various methods have been proposed to reduce the computational complexity while achieving comparable localization performance [6–10].

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2177/1 - Project ID 390895286 and Project ID 352015383 - SFB 1330 B2.

Instead of directly estimating the source position, in this paper we propose an indirect estimation method based on a Euclidean distance matrix (EDM) [11, 12], containing both the distances between the microphones (assumed to be known) and the (unknown) distances between the source and the microphones. We propose to decompose the unknown distances into the distance between the source and the reference microphone and a distance component which is proportional to the time-differences of arrival (TDOAs) between the reference microphone and the other microphones. Assuming estimates of the TDOAs to be available allows us to formulate the EDM and the related Gram matrix as a function of a single variable, representing the distance between the source and the reference microphone. Exploiting the rank property of the Gram matrix, we propose to minimize a cost function which depends on this variable. The estimated relative source position can be reconstructed from the Gram matrix, which minimizes the cost function, and can then be aligned to the estimated source position using orthogonal Procrustes analysis [12, 13]. Since in reverberant environments early reflections may result in large TDOA estimation errors, we propose a method to select the best TDOA estimate from a set of multiple candidate TDOA estimates, based on the same rank property of Gram matrices.

Experimental results for different source and microphone constellations in noisy and reverberant environments show that the proposed EDM-based 3D source localization method outperforms SRP-PHAT and results in significantly smaller estimation errors when the source is close to the microphones. Furthermore, we show that the proposed TDOA selection method leads to a reduction in the number of large localization errors.

2. SOURCE LOCALIZATION USING SRP-PHAT

We consider a reverberant and noisy acoustic environment with a single static speech source and a spatially distributed microphone array with $M > 3$ microphones, where $\mathbf{m}_m \in \mathbb{R}^3$ denotes the position of the m-th microphone. The aim is to estimate the source position $\mathbf{s} \in \mathbb{R}^3$ relative to the microphone positions $\mathbf{M} = [\mathbf{m}_1, \dots, \mathbf{m}_M]$, which are assumed to be known. Assuming synchronized microphones and free field transmission, i.e., no object or head between the source and the microphones, the TDOA of the direct speech component between the i-th and j-th microphone is equal to $\tau_{i,j}(\mathbf{s}) = (||\mathbf{s}-\mathbf{m}_i|| - ||\mathbf{s}-\mathbf{m}_j||)/\nu$ with ν the speed of sound.

A common approach to estimate the TDOAs between the microphone pairs is based on the time-domain generalized cross correlation with phase transform (GCC-PHAT) function [5, 14, 15], defined between microphone i and j as

$$\xi_{i,j}(\tau) = \int_{-\omega_0}^{\omega_0} \psi_{i,j}(\omega) e^{\jmath\omega\tau} d\omega \,, \tag{1}$$

with radial frequency $-\omega_0 \leq \omega \leq \omega_0$ and time lag τ. The frequency-

domain GCC-PHAT function $\psi_{i,j}(\omega)$ in (2) is given by

$$\psi_{i,j}(\omega) = \frac{\mathbb{E}\{Y_i(\omega)Y_j^*(\omega)\}}{|\mathbb{E}\{Y_i(\omega)Y_j^*(\omega)\}|}, \tag{2}$$

where $Y_m(\omega)$ denotes the m-th microphone signal in the frequency-domain and $\mathbb{E}\{\cdot\}$ the expectation operator. The PHAT weighting in (2) has been shown to improve robustness against reverberation and noise [14–16]. The TDOA $\hat{\tau}_{i,j}$ between the i-th and j-th microphone is estimated by maximizing $\xi_{i,j}(\tau)$, i.e.,

$$\hat{\tau}_{i,j} = \underset{\tau}{\operatorname{argmax}}\, \xi_{i,j}(\tau). \tag{3}$$

Building upon GCC-PHAT, the SRP-PHAT method [1] is a popular method for 3D source localization. The SRP-PHAT functional for the 3D position $\mathbf{p} = [p_x, p_y, p_z]^{\mathrm{T}}$ is defined as

$$\Psi(\mathbf{p}) = \sum_{i,j:i>j} \int_{-\omega_0}^{\omega_0} \psi_{i,j}(\omega) e^{\jmath\omega\tau_{i,j}(\mathbf{p})}\, d\omega, \tag{4}$$

where $\tau_{i,j}(\mathbf{p})$ denotes the TDOA corresponding to a source at position $\mathbf{p}$ and the TDOAs between all microphone pairs are considered. The source position is estimated as

$$\hat{\mathbf{s}}_{\text{SRP-PHAT}} = \underset{\mathbf{p}}{\operatorname{argmax}}\, \Psi(\mathbf{p}), \tag{5}$$

which requires the optimization of three continuous variables, i.e., $0 \le p_x \le P_x$, $0 \le p_y \le P_y$ and $0 \le p_z \le P_z$, with P_x, P_y and P_z the room dimensions.

3. EDM-BASED SOURCE LOCALIZATION

In this section, we show how to determine the source position, by constructing an EDM of the distances between the microphones and between the source and the microphones (Section 3.1) in a way, which, together with the TDOAs, allows us to build a cost function to determine the (unknown) distance between the source and the reference microphone (Section 3.2). Furthermore, we propose a method to select the best TDOA estimate out of multiple candidate estimates (Section 3.3).

3.1. Properties of EDM Matrices

We define the $3 \times (M+1)$-dimensional positions matrix as $\mathbf{P} = [\mathbf{M}|\mathbf{s}]$. The corresponding $(M+1) \times (M+1)$-dimensional EDM $\overline{\mathbf{D}}$ is defined as

$$\overline{\mathbf{D}} = \left[\begin{array}{c|c} \mathbf{D} & \mathbf{d} \\ \hline \mathbf{d}^{\mathrm{T}} & 0 \end{array}\right]. \tag{6}$$

This matrix contains the inter-microphone EDM $\mathbf{D} = \left[D_{i,j}^2\right]$, with $D_{i,j} = ||\mathbf{m}_i - \mathbf{m}_j||$ the distances between the i-th and j-th microphones, and the vector of squared distances $\mathbf{d} = \left[d_1^2, ..., d_M^2\right]^{\mathrm{T}}$, with $d_m = ||\mathbf{m}_m - \mathbf{s}||$ the distance between the source and the m-th microphone.

In [12, 17], it was shown that an EDM corresponding to a 3D geometry can be transformed to a Gram matrix, whose rank is at most 3, as

$$\mathbf{G} = -\frac{1}{2}(\mathbf{I} - \mathbf{1}\mathbf{e}^{\mathrm{T}})\overline{\mathbf{D}}(\mathbf{I} - \mathbf{e}\mathbf{1}^{\mathrm{T}}), \tag{7}$$

where $\mathbf{I}$ denotes the identity matrix, $\mathbf{1}$ denotes a vector with ones, and $\mathbf{e}$ denotes a vector with zeros except for the element corresponding to the reference microphone (chosen as the first microphone without loss of generality), equal to one. The Gram matrix can be written using the relative microphone and source positions $\mathbf{P}_{\text{rel}}$ as $\mathbf{G} = \mathbf{P}_{\text{rel}}^{\mathrm{T}}\mathbf{P}_{\text{rel}}$, where the absolute positions $\mathbf{P}$ are related to the relative positions $\mathbf{P}_{\text{rel}}$ via a translation which places the reference microphone at the origin, and the remaining array is arbitrarily rotated and/or reflected (preserving the inter-microphone distances). Realizing that the positive semi-definite Gram matrix has at most 3 positive eigenvalues which are not equal to zero, i.e., $\lambda_1 \ge ... \ge \lambda_3 \ge 0$ and $\lambda_4 = \cdots = \lambda_{M+1} = 0$, the relative positions $\mathbf{P}_{\text{rel}}$ can be written using the eigenvalue decomposition of $\mathbf{G}$ as

$$\mathbf{P}_{\text{rel}} = \left[\operatorname{diag}\left(\sqrt{\lambda_1}, ..., \sqrt{\lambda_3}\right) | \, \mathbf{0}_{3\times((M+1)-3)}\right]\mathbf{U}^{\mathrm{T}}, \tag{8}$$

where $\mathbf{0}_{3\times((M+1)-3)}$ is a $3 \times ((M+1)-3)$ dimensional matrix of zeros and $\mathbf{U}$ denotes the matrix containing the eigenvectors of $\mathbf{G}$. The relative positions $\mathbf{P}_{\text{rel}}$ can be aligned with the absolute positions $\mathbf{P}$ using orthogonal Procrustes analysis [12, 13] by aligning the relative microphone positions $\mathbf{M}_{\text{rel}}$ with the known microphone positions $\mathbf{M}$. This simultaneously aligns the relative source position $\mathbf{s}_{\text{rel}}$ with the absolute source position $\mathbf{s}$.

3.2. EDM-Based Cost Function

Defining α_s as the distance between the source and the reference microphone (i.e., $\alpha_s = d_1$), the distance between the source and the m-th microphone can be written as

$$d_m = \alpha_s + \nu\tau_{m,1}(\mathbf{s}), \quad m = 1, ..., M, \tag{9}$$

where $\tau_{m,1}(\mathbf{s})$ denotes the TDOA between the m-th microphone and the reference microphone. Assuming for now that the TDOAs are known and considering the distance variable α, we can define $d_m(\alpha)$ similarly to (9), i.e.,

$$d_m(\alpha) = \alpha + \nu\tau_{m,1}(\mathbf{s}). \tag{10}$$

Using $d_m(\alpha)$, we can construct the vector of squared distances $\mathbf{d}(\alpha) = \left[d_1^2(\alpha), ..., d_M^2(\alpha)\right]^{\mathrm{T}}$, the EDM $\overline{\mathbf{D}}(\alpha)$, and its Gram matrix $\mathbf{G}(\alpha)$. As mentioned in Section 3.1, the rank of $\mathbf{G}(\alpha)$ is equal to 3 if $\alpha = \alpha_s$. Motivated by the idea of minimizing the rank of a matrix as in [18], we now define the cost function

$$\boxed{J(\alpha) = \sum_{i=3+1}^{M+1} |\lambda_i(\alpha)|} \tag{11}$$

which considers all but the three largest eigenvalues $\lambda_i(\alpha)$ of $\mathbf{G}(\alpha)$. The absolute values of the eigenvalues are used, since it can not be guaranteed that the eigenvalues of $\mathbf{G}(\alpha)$ are positive for all values of α (e.g., in case of a mismatch with the TDOAs). If $\alpha = \alpha_s$, then all but the three largest eigenvalues of $\mathbf{G}(\alpha_s)$ are equal to zero, such that $J(\alpha_s) = 0$. The optimal value α_s can hence be found as

$$\alpha_s = \underset{\alpha}{\operatorname{argmin}}\, J(\alpha). \tag{12}$$

3.3. TDOA Selection

In practice, the TDOAs, of course, are not available, so we now rewrite (10) to take into account the estimated TDOAs. If the source or a microphone is close to a wall or the corner of a room, super-positions of acoustic reflections may lead to peaks in the time-domain GCC-PHAT function (1), which are higher than the peak corresponding to the direct path. Basing the TDOA estimate on these erroneous peaks can result in large errors in the source localization. We propose to consider C candidate TDOA estimates per microphone pair, corresponding to the C highest local peaks in the time-domain GCC-PHAT function. The index $c_m \in \{1, ..., C\}$ denotes the candidate TDOA estimate $\hat{\tau}_{m,1}^{c_m}$ between the m-th microphone and the reference microphone. This means that the distance variable $d_m(\alpha, \hat{\tau}_{m,1}^{c_m})$ now becomes a function of the distance variable α as well as the estimated candidate TDOA estimates, i.e.,

$$d_m(\alpha, \hat{\tau}_{m,1}^{c_m}) = \alpha + \nu\hat{\tau}_{m,1}^{c_m}. \tag{13}$$

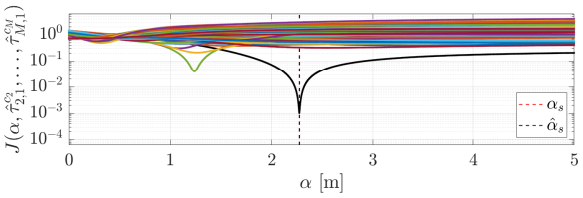

Fig. 1. Example of the cost function (14) using estimated TDOAs, considering $C=3$ candidate TDOA estimates, with $M=6$ microphones (i.e., $C^{M-1}=3^5=243$ total combinations), for a distance $\alpha_s=2.28$ m between the source and the reference microphone.

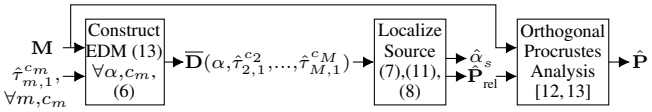

Fig. 2. Overview of EDM-Based Source Localization

Similarly to Section 3.2, we now construct the vector of squared distances $\mathbf{d}(\alpha,\hat{\tau}_{2,1}^{c_2},\ldots,\hat{\tau}_{M,1}^{c_M})$, the EDM $\overline{\mathbf{D}}(\alpha,\hat{\tau}_{2,1}^{c_2},\ldots,\hat{\tau}_{M,1}^{c_M})$ and its Gram matrix $\mathbf{G}(\alpha,\hat{\tau}_{2,1}^{c_2},\ldots,\hat{\tau}_{M,1}^{c_M})$, and determine the optimal distance variable $\hat{\alpha}_s$ with (11) for all C^{M-1} possible combinations of candidate TDOA estimates, taking the value with the minimal cost, i.e.,

$$\hat{\alpha}_s \;=\; \underset{\alpha,c_2,\ldots,c_M}{\operatorname{argmin}} \; J(\alpha,\hat{\tau}_{2,1}^{c_2},\ldots,\hat{\tau}_{M,1}^{c_M}) \tag{14}$$

This corresponds to the distance between the source and the reference microphone α and the combination of candidate TDOA estimates which best match with each-other in terms of constructing a 3D geometry. It should be noted that the minimum of (14) is not guaranteed to be 0 like in (11), due to possible estimation errors in the TDOAs. For an exemplary 3D source and microphone constellation with $\alpha_s=d_1=2.28$ m, Fig. 1 depicts the dependence of the cost function $J(\alpha,\hat{\tau}_{2,1}^{c_2},\ldots,\hat{\tau}_{M,1}^{c_M})$ on the distance variable α.

To reconstruct the estimated relative positions $\hat{\mathbf{P}}_{\text{rel}}$, only the three largest positive eigenvalues (for which the cost function (14) is minimized) and the corresponding eigenvectors are used in (8). The same alignment procedure is applied as described in Section 3 to map the estimated relative microphone position $\hat{\mathbf{s}}_{\text{rel}}$ to the estimated microphone position $\hat{\mathbf{s}}$. An overview of the EDM-based source localization is depicted in Fig. 2.

4. PRACTICAL IMPLEMENTATION

In this section, we discuss practical considerations to implement the previously discussed localization algorithms from Sections 2 and 3 in the short-time Fourier transform (STFT) domain.

4.1. Implementation of SRP-PHAT

In practice, the maximization of the SRP functional $\Psi(\mathbf{p})$ in (5), which depends on 3 continuous variables, is approximated through an exhaustive search on a discrete grid $\mathbf{p}'$. First, the integral in (4) is approximated by a sum over STFT frequency bins, i.e.,

$$\Psi[l](\mathbf{p}') \;=\; \sum_{i,j:i>j}\sum_{k=0}^{K-1}\psi_{i,j}[k,l]e^{\jmath 2\pi f_s\tau_{i,j}(\mathbf{p}')k/K}\,, \tag{15}$$

with frequency bin index $k\in\{0,\ldots,K-1\}$, where K is the Fourier transform length, and frame index $l\in\{1,\ldots,L\}$. Assuming a static source, the source position is then estimated by maximizing the sum of the SRP-PHAT functionals over L frames, i.e.,

$$\hat{\mathbf{s}}'_{\text{SRP-PHAT}} \;=\; \underset{\mathbf{p}'}{\operatorname{argmax}} \sum_{l=1}^{L}\Psi[l](\mathbf{p}')\,. \tag{16}$$

Since we perform a summation over frames in (16), we use instantaneous estimates of $\psi_{i,j}[k,l]$ in (15) (i.e, the expectation operation in (2) constitutes an average over a single frame) in order to not perform two temporal averaging operations.

Exhaustively searching for the 3D source position estimate at a high grid resolution can be computationally demanding. Similarly to coarse-to-fine region contraction in [6], we first evaluate (16) on a coarse 3D grid and then in the vicinity of a few points where SRP-PHAT yields the highest values, we evaluate the SRP-PHAT functional on a fine grid in those regions.

4.2. Implementation of EDM-Based Localization

In the STFT-domain, GCC-PHAT $\psi_{i,j}[k,l]$ is estimated in each frequency bin k and time frame l, and the the continuous Fourier transform in time-domain GCC-PHAT is approximated with an inverse discrete Fourier transform for discrete time-lags n (with $\tau=n/f_s$), i.e.,

$$\xi_{m,1}[n,l] \;=\sum_{k=0}^{K-1}\psi_{m,1}[k,l]e^{\jmath 2\pi nk/K}\,. \tag{17}$$

with f_s the sampling frequency. To achieve a more precise TDOA estimate, the time-domain GCC-PHAT function $\xi_{m,1}[n,l]$ can be interpolated between the discrete time-lags n with a factor $R\geq 1$. The lower and upper limits of the possible time-lags are dependent on the distances between the microphone pair, i.e., $|n_m|<Rf_sD_{m,1}/\nu$. Using the interpolated time-domain GCC-PHAT, the sample-delay between the m-th and the reference microphone is estimated as

$$\hat{n}_m \;=\; \underset{n_m}{\operatorname{argmax}} \sum_{l=1}^{L}\xi_{m,1}[n_m,l]\,, \tag{18}$$

corresponding to the estimated TDOA $\hat{\tau}_{m,1}=\hat{n}_m/(Rf_s)$. Similarly to the previous section, since we perform a summation over frames in (18), we use instantaneous estimates of $\psi_{i,j}[k,l]$ in (17).

5. EXPERIMENTAL EVALUATION

In this section, we experimentally compare the source localization performance of the proposed EDM-based method (for up to three candidate TDOA estimates C per microphone pair) with the SRP-PHAT method for four different distances between the source position and the centroid of the microphone positions.

5.1. Scenario and Algorithm Parameters

For the simulations, we considered a rectangular room with dimensions $6\times6\times2.4$ m and simulated room impulse responses using the image method [19,20], assuming equal reflection coefficients for all walls. The $M=6$ spatially distributed microphones were randomly positioned within a cube with cube length 2 m (with a minimum distance of 2 cm between the microphones) and the source was located at one of four fixed distances $\alpha_c\in\{0.5,1,2,3\}$ m from the centroid of the microphone positions (in a random direction). For each source distance α_c, we considered 100 acoustic scenarios, using a 5 s speech signal randomly selected from [21] (with equal probability for a male or female speaker) as the source signal. The reflection coefficients were set for each scenario such that the room impulse responses had an average direct-to-reverberant ratio of approximately 0 dB over the microphones. This was achieved by setting $T_{60}=0.60\pm0.14$ s for $\alpha_c=0.5$ m, $T_{60}=0.47\pm0.10$ s for

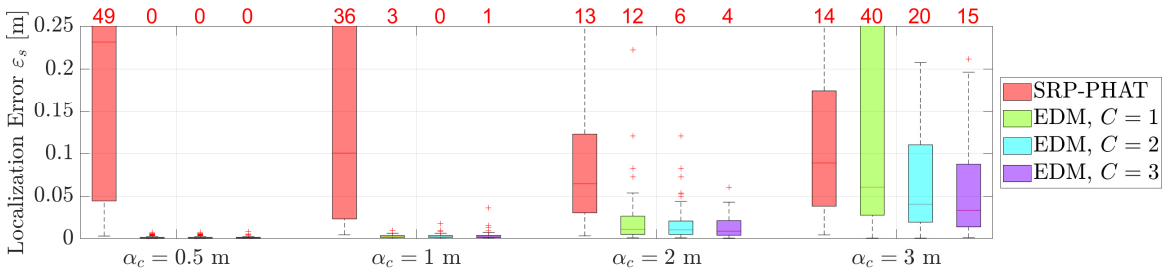

Fig. 3. Box plots of the localization errors ε_s (over 100 scenarios) for the SRP-PHAT method and the EDM-based method (with different numbers of candidate TDOA estimates C per microphone pair), for different distances α_c between the source and the centroid of the distributed microphones. The number of results outside of the plotted range are denoted by red numbers at the top.

Table 1. Median localization errors (over 100 scenarios) for the SRP-PHAT method and the EDM-based method, corresponding to the box plots in Fig. 3

	Median ε_s [m]			
$\alpha_c[m]$	SRP-PHAT	EDM, $C=1$	EDM, $C=2$	EDM, $C=3$
0.5	0.231	**0.001**	**0.001**	**0.001**
1	0.101	**0.002**	**0.002**	**0.002**
2	0.065	0.011	0.010	**0.009**
3	0.089	0.061	0.041	**0.033**

$\alpha_c = 1$ m, $T_{60} = 0.29 \pm 0.04$ s for $\alpha_c = 2$ m and $T_{60} = 0.25 \pm 0.03$ s for $\alpha_c = 3$ m. Spherically isotropic multi-talker babble noise was generated using [22] and added to the reverberant speech component in the microphones at 5 dB signal-to-noise ratio. The sampling frequency was equal to 16 kHz.

The algorithms were implemented using an STFT framework, with a frame length of 512 samples (corresponding to 32 ms), 50% overlap between frames, a discrete Fourier transform-length of 1024 samples and using a square-root-Hann analysis window.

For SRP-PHAT, the functional in (15) was evaluated first on a coarse grid with 10 cm resolution in x-, y-, and z-direction, and then reevaluated for the three grid points with the highest SRP-PHAT value on a fine grid with 1 cm resolution in each dimension. For the proposed EDM-based source localization method, the time-domain GCC-PHAT function in (17) was interpolated by a factor $R = 720$. To emphasize strong peaks, we weighted the time-domain GCC-PHAT function as $\widetilde{\xi}_{m,1}[n_m, l] = \exp(15\xi_{m,1}[n_m, l])$ prior to the TDOA estimation in (18). The candidate TDOAs were selected using a peak finding algorithm [23]. In (14), the exhaustive search for the optimal distance variable α_s was performed with a resolution of 1 mm, up to a maximal distance determined by the distance between opposite corners of the room (i.e., $\sqrt{6^2+6^2+2.4^2}$ m $\approx$ 8.82 m). For $C = 2$ candidate TDOA estimates per microphone pair, the number of combinations of TDOA estimates was $C^{M-1} = 2^5 = 32$, while for $C = 3$ the number of combinations was $C^{M-1} = 3^5 = 243$.

5.2. Performance Comparison

To analyze and compare the performance of the considered 3D source localization methods, we used the localization error

$$\varepsilon_s = ||\mathbf{s} - \hat{\mathbf{s}}||. \tag{19}$$

For different source distances α_c, Fig. 3 depicts the box plots of the localization error (over 100 scenarios) for the SRP-PHAT method and for the proposed EDM-based method, for different numbers of candidate TDOA estimates. Tab. 1 presents the corresponding median localization errors.

Considering source positions outside of the array of distributed microphones (i.e., $\alpha_c \geq 2$ m), it is clear from Fig. 3 that by increasing the number of candidate TDOA estimates, both the median localization errors as well as the number of errors larger than 25 cm are reduced. This suggests that the proposed procedure, considering multiple candidate TDOA estimates, is able to identify the TDOA corresponding to the direct path. Using $C = 2$ or $C = 3$ suffices, to halve the number of localization errors larger than 25 cm, compared to using $C = 1$, and the median localization error can be substantially reduced, especially for large source distances α_c. For source positions within the array of distributed microphones (i.e., $\alpha_c \leq 1$ m), considering more than $C = 1$ candidate TDOA estimates is not necessary, since the median localization error (in Tab. 1) is constantly at 1 mm or 2 mm, independently of C.

In Fig. 3 it can clearly be observed that when the distance between the source and the centroid of the microphones α_c is smaller than or equal to the cube length of the array of distributed microphones (i.e., $\alpha_c \leq 2$ m), the proposed EDM-based source localization method results in significantly lower localization errors than the SRP-PHAT method, regardless of the number of candidate TDOA estimates. For example, for $\alpha_c = 2$ m, the median localization error for the EDM-based method is 1 cm ± 1 mm (depending on C), whereas for the SRP-PHAT method the median localization error is 6.5 cm.

For $\alpha_c = 3$ m, the EDM-based method with $C = 3$ candidate TDOA estimates and the SRP-PHAT method have overlapping distributions of localization errors and a comparable number of errors larger than 25 cm, but the EDM-based method has a lower median error, i.e., 3.3 cm, compared to 8.9 cm for the SRP-PHAT method.

6. CONCLUSIONS

We have proposed a new 3D source localization method, which, through properties of EDMs, and by the specific construction of the EDM containing the distances between microphones and between the source and the microphones, only requires the optimization of a single variable, namely the distance between the source and the reference microphone. As this method relies on estimated TDOAs, we proposed a method to select the best TDOA estimate out of multiple estimates in the presence of reverberation. Experimental results for different source and microphone constellations showed that the proposed EDM-based source localization method consistently localizes sources with a lower localization error than the commonly used SRP-PHAT method for all tested source distances. The proposed method for estimating the best candidate TDOA estimates also results in a reduction in the number of large localization errors.

7. REFERENCES

[1] J H DiBiase, *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*, Ph.D. thesis, Brown University, Providence, RI, USA, 2000.

[2] N Madhu, R Martin, U Heute, and C Antweiler, "Acoustic source localization with microphone arrays," *Advances in Digital Speech Transmission*, pp. 135–170, 2008.

[3] Y A Huang, J Benesty, and J Chen, "Time delay estimation and source localization," in *Springer Handbook of Speech Processing*, pp. 1043–1063. Springer, 2008.

[4] P Pertilä, A Brutti, P Svaizer, and M Omologo, "Multichannel source activity detection, localization, and tracking," *Audio source separation and speech enhancement*, pp. 47–64, 2018.

[5] C Knapp and G Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. on Audio, Speech, Language Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[6] H Do and H F Silverman, "A fast microphone array SRP-PHAT source location implementation using coarse-to-fine region contraction (CFRC)," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2007, pp. 295–298.

[7] M Cobos, A Marti, and J J Lopez, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling," *IEEE Signal Processing Letters*, vol. 18, no. 1, pp. 71–74, 2010.

[8] 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. on Signal Processing*, vol. 62, no. 19, pp. 5171–5183, 2014.

[9] G García-Barrios, J M Gutiérrez-Arriola, N Sáenz-Lechón, V J Osma-Ruiz, and R Fraile, "Analytical model for the relation between signal bandwidth and spatial resolution in steered-response power phase transform (SRP-PHAT) maps," *IEEE Access*, vol. 9, pp. 121549–121560, 2021.

[10] 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.

[11] W S Torgerson, "Multidimensional scaling: I. theory and method," *Psychometrika*, vol. 17, no. 4, pp. 401–419, 1952.

[12] I Dokmanic, R Parhizkar, J Ranieri, and M Vetterli, "Euclidean distance matrices: essential theory, algorithms, and applications," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 12–30, 2015.

[13] P Schoenemann, *A solution of the orthogonal Procrustes problem with applications to orthogonal and oblique rotation*, Ph.D. thesis, University of Illinois, Urbana-Champaign, 1964.

[14] J Velasco, C J Martin-Arguedas, J Macias-Guarasa, D Pizarro, and M Mazo, "Proposal and validation of an analytical generative model of SRP-PHAT power maps in reverberant scenarios," *Signal Processing*, vol. 119, pp. 209–228, 2016.

[15] C Zhang, D Florêncio, and Z Zhang, "Why does PHAT work well in low noise, reverberative environments?," in *Proc. IEEE International Conference of Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, NV, USA, 2008, pp. 2565–2568.

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

[17] J C Gower, "Euclidean distance geometry," *Math. Sci*, vol. 7, no. 1, pp. 1–14, 1982.

[18] O Roy and M Vetterli, "The effective rank: A measure of effective dimensionality," in *Proc. European Signal Processing Conference (EUSIPCO)*, Poznan, Poland, 2007, pp. 606–610.

[19] E A P Habets, *RIR-Generator*, Available at "`https://github.com/ehabets/RIR-Generator`".

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

[21] I Solak, *M-AILABS Speech Dataset*, Available at "`https://www.caito.de/2019/01/03/the-m-ailabs-speech-dataset/`".

[22] E A P Habets, I Cohen, and S Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 2911–2917, 2008.

[23] *Matlab `findpeaks` function*, Documentation available at "`https://www.mathworks.com/help/pdf_doc/signal/signal_ref.pdf`", p. 414, 2022.