Exploiting an External Microphone to Improve Time-Difference-of-Arrival Estimates for Euclidean Distance Matrix-Based Source Localization

Klaus Brümann, Simon Doclo

Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4all, University of Oldenburg, Germany
Email: {klaus.bruemann,simon.doclo}@uni-oldenburg.de

Abstract

Recently, a Euclidean distance matrix (EDM)-based 3D source position estimation method was proposed, which relies on estimated time-differences-of-arrival (TDOAs) between microphones with a known array geometry. These TDOAs are often estimated by determining the time lag that maximizes the generalized cross-correlation with phase transform (GCC-PHAT) function for each microphone pair. In noisy and reverberant environments, the time lag corresponding to the direct source component may no longer correspond to the maximum of the GCC-PHAT function, which leads to TDOA estimation errors and consequent source localization errors. In this paper, we assume the availability of an external microphone at an unknown position in the vicinity of the source, in addition to the microphone array, hence, typically having a high direct source component relative to the noise and reverberation. Aiming at improving the reliability of the GCC-PHAT function, we propose a new approach to compute the GCC-PHAT function for each microphone pair in the array by convolving the GCC-PHAT functions between the external microphone and the respective microphones in the array. Experimental results demonstrate that incorporating an external microphone enables a significant improvement in the source localization performance of the EDM-based method, when the external microphone is close enough to the source.

1 Introduction

Microphone arrays are extensively employed to localize speech sources [1–3] for acoustic processing. Several learning- and non-learning based approaches have been proposed for source localization, e.g., approaches using time-differences-of-arrival (TDOAs) [2, 4], the steered response power with phase transform method [5–7], the subspace-based multiple signal classification approach [8], matching estimated relative transfer functions with a database of prototype relative transfer functions [9], and deep neural network-based approaches [10, 11]. 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.

We consider a recently proposed Euclidean distance matrix (EDM)-based method [12], which does not rely on the far field assumption, but rather a sufficiently large array geometry for which the 3D source position can be estimated, i.e., using spatially distributed microphones of an acoustic sensor network. This method considers the geometry of the distributed microphone array (DMA), which is assumed to be known, and performs best when multiple candidate TDOA estimates are considered between the reference microphone and each other microphone. These candidate TDOAs are estimated using the popular generalized cross correlation with phase transform (GCC-PHAT) method [13]. An EDM-based cost function determines the specific combination of TDOA estimates and the distance between the source and the reference microphone which best corresponds to the source position. Nevertheless, in high levels of reverberation and noise, the GCC-PHAT function may exhibit too many candidate TDOA estimates that do not match the TDOA of the direct source component, leading to an incorrect estimate of the source position.

In addition to the main microphone array with a known geometry, sometimes an external microphone with an unknown position is available. If this external microphone is in the vicinity of the speech source it might capture a loud direct source reference and if it is far enough from the other microphones, its captured noise and reverberation reference might be more uncorrelated, in comparison to the correlation between the microphones of the main microphone array. Several strategies have been proposed to harness these properties, e.g., in beamforming for speech enhancement/noise reduction [14–17], dereverberation [18], and source localization approaches, such as relative transfer function matching [19], deep neural network-based [20], and informed sound source localization [21, 22]. In this paper, we incorporate the external microphone, aiming to improve the accuracy of the EDM-based source localization method by means of improving the reliability of the TDOA estimation.

We propose to exploit the potentially more favourable conditions in the external microphone by incorporating its captured signal in the computation of the GCC-PHAT function between the reference microphone and the other microphones of the main array. We replace the GCC-PHAT functions between the reference microphone and the other microphones with new functions which are obtained by convolving the GCC-PHAT function between the reference microphone and the external microphone with the GCC-PHAT function between the external microphone and the other microphones. This convolution cancels out the time difference corresponding to the signal delay between the reference microphone of the main array and the external microphone, hence, the new function has a peak at the time lag corresponding to the TDOA between the reference microphone and the corresponding other microphone.

Experimental results for different distances between the source and the external microphone show that the source position estimation error can be considerably reduced when an external microphone, positioned between the source and the DMA, is used to estimate the TDOAs.

2 TDOA Estimation Using GCC-PHAT

We consider a reverberant and noisy acoustic environment with a single static speech source and a DMA with $M > 3$ microphones at positions $\mathbf{M} = [\mathbf{m}_1,...,\mathbf{m}_M] \in \mathbb{R}^{3\times M}$. In addition, we assume that an external microphone is available at an unknown position $\mathbf{m}_\mathrm{E}$ close to the speech source at an unknown position $\mathbf{s}$. Assuming synchronized microphones and free-field transmission, i.e., no object or head between the source and the microphones, the TDOA of the direct source component between the i-th and the j-th microphone is equal to $\tau_{i,j}(\mathbf{s}) = (\alpha_i - \alpha_j)/\nu$, where α_m denotes the distance between the speech source and the m-th microphone and ν denotes the speed of sound. The distance between the i-th and j-th microphone is denoted as $D_{i,j}$.

A common approach to estimate the TDOA between two microphones is based on the real-valued time-domain generalized cross-correlation with phase transform (GCC-PHAT) function [13]. The GCC-PHAT function between microphones i and

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.

j is defined 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 $-\omega_0 \leq \omega \leq \omega_0$, time lag τ in seconds, and phase spectrum $\psi_{i,j}(\omega)$ 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 $\mathbb{E}\{\cdot\}$ denotes the expectation operator. The estimated TDOA $\hat{\tau}_{i,j}$ between the i-th and j-th microphone is computed as the time lag which maximizes $\xi_{i,j}(\tau)$, i.e.,

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

To analyze the GCC-PHAT function in more detail, we decompose the noisy and reverberant m-th microphone signal, similarly to [23], as

$$Y_m(\omega,\alpha_m) = X_{d,m}(\omega,\alpha_m) + X_{r,m}(\omega) + N_m(\omega)\,, \tag{4}$$

where $X_{d,m}(\omega, \alpha_m)$ denotes the direct source component, $X_{r,m}(\omega)$ denotes the reverberation component, and $N_m(\omega)$ denotes the noise component. The direct source component is a delayed and attenuated version of the source signal $S(\omega)$ [24], i.e.,

$$X_{d,m}(\omega,\alpha_m) = \frac{\exp(-\jmath\omega\alpha_m/\nu)}{\sqrt{4\pi}\alpha_m} S(\omega)\,. \tag{5}$$

We now assume that the direct source component, the reverberation component, and the noise component are mutually uncorrelated. Although in practice the reverberation component is not completely uncorrelated with the direct source component due to early reflections, we use this common assumption for a simplified derivation. In addition, we assume reverberation and noise are homogeneous sound fields with spatial coherence $\Gamma_{i,j}(\omega)$ between the i-th and j-th microphone [25]. Using these assumptions, it can easily be shown that

$$\mathbb{E}\{X_{d,i}(\omega,\alpha_i)X_{d,j}^*(\omega,\alpha_j)\} = \phi_D(\omega,\alpha_i,\alpha_j)G(\omega,\alpha_i,\alpha_j)\,, \tag{6}$$

$$\mathbb{E}\{X_{r,i}(\omega)X_{r,j}^*(\omega)\} = \phi_R(\omega)\Gamma_{i,j}(\omega)\,, \tag{7}$$

$$\mathbb{E}\{N_i(\omega)N_j^*(\omega)\} = \phi_N(\omega)\Gamma_{i,j}(\omega)\,, \tag{8}$$

where

$$G(\omega,\alpha_i,\alpha_j) = \exp(-\jmath\omega(\alpha_i-\alpha_j)/\nu) \tag{9}$$

denotes the relative direct transfer function and $\phi_D(\omega,\alpha_i,\alpha_j) = \phi_S(\omega)/(4\pi\alpha_i\alpha_j)$, $\phi_S(\omega)$, $\phi_R(\omega)$, and $\phi_N(\omega)$ denote the direct, source, reverberation, and noise power spectral densities (PSDs), respectively.

Omitting the radial frequency ω for a more concise notation, the phase spectrum in (2) can be written as

$$\psi_{i,j}(\alpha_i,\alpha_j) = \frac{\phi_D(\alpha_i,\alpha_j)G(\alpha_i,\alpha_j) + (\phi_R+\phi_N)\Gamma_{i,j}}{|\phi_D(\alpha_i,\alpha_j)G(\alpha_i,\alpha_j) + (\phi_R+\phi_N)\Gamma_{i,j}|}\,, \tag{10}$$

$$= \frac{\mathrm{DUR}(\alpha_i,\alpha_j)G(\alpha_i,\alpha_j) + \Gamma_{i,j}}{|\mathrm{DUR}(\alpha_i,\alpha_j)G(\alpha_i,\alpha_j) + \Gamma_{i,j}|}\,, \tag{11}$$

with the direct-to-undesired ratio (DUR) in the DMA, i.e., $\mathrm{DUR}(\alpha_i,\alpha_j) = \phi_D(\alpha_i,\alpha_j)/(\phi_R+\phi_N)$.

Clearly, in the noise-free and anechoic case (i.e., $\mathrm{DUR}(\alpha_i, \alpha_j) = \infty$), the phase spectrum in (11) is equal to $G(\alpha_i,\alpha_j)$ and the GCC-PHAT function is equal to a time-shifted delta function $\delta(\tau - (\alpha_i - \alpha_j)/\nu) = \delta(\tau - \tau_{i,j})$, such that the TDOA estimate in (3) exactly corresponds to the TDOA, i.e., $\hat{\tau}_{i,j} = \tau_{i,j}$. In the presence of reverberation and noise, the GCC-PHAT function is influenced by the DUR and the spatial coherence, possibly introducing additional peaks and resulting in TDOA estimation errors.

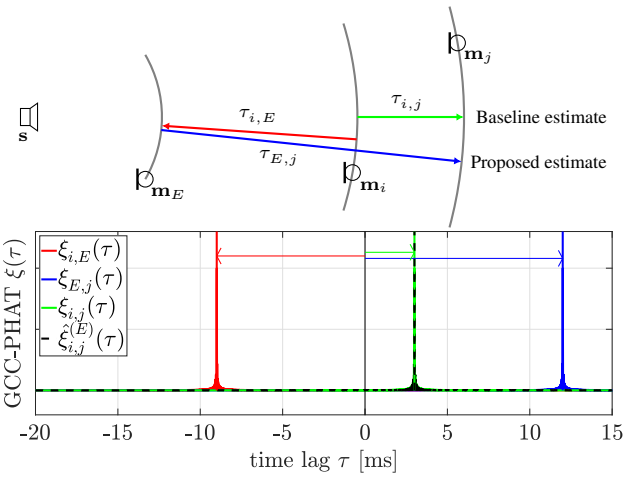

Figure 1: TDOA equivalency in (14) for an anechoic and noise-free scenario. Top: 2D schematic of an exemplary microphone layout. Bottom: GCC-PHAT functions corresponding to the exemplary microphone layout.

3 TDOA Estimation Exploiting an External Microphone

In this section, we propose a method which exploits the external microphone to improve the TDOA estimates between the microphones of the DMA. We assume the external microphone to be in the vicinity of the speech source, i.e., the distance α_E between the source and the external microphone is assumed to be smaller than the distances between the source and the microphones of the DMA. This generally means that the DUR is larger in the external microphone than in the microphones of the DMA.

Since the reliability of the GCC-PHAT function is directly related to the DUR, cf. (11), the reliability of the GCC-PHAT function $\xi_{i,E}(\tau)$ between the i-th microphone of the DMA and the external microphone is in general better than the reliability of the GCC-PHAT function $\xi_{i,j}(\tau)$ between the i-th and j-th microphone of the DMA. Hence, we propose to compute the GCC-PHAT function between the i-th and j-th microphone by convolving the GCC-PHAT functions $\xi_{i,E}(\tau)$ and $\xi_{E,j}(\tau)$, i.e.,

$$\boxed{\xi_{i,j}^{(E)}(\tau) = \int_{-\infty}^{\infty} \xi_{i,E}(\tau')\xi_{E,j}(\tau-\tau')d\tau'} \tag{12}$$

Fig. 1 depicts the acoustic scenario with a source at position $\mathbf{s}$, two microphones of the DMA at positions $\mathbf{m}_i$ and $\mathbf{m}_j$ and an external microphone at position $\mathbf{m}_E$. In the anechoic and noise-free case, it can easily be shown that the GCC-PHAT function in (12) becomes

$$\xi_{i,j}^{(E)}(\tau) = \int_{-\infty}^{\infty} \delta(\tau' - \frac{\alpha_i-\alpha_E}{\nu})\delta(\tau - \frac{\alpha_E-\alpha_j}{\nu} - \tau')\,d\tau'\,, \tag{13}$$

$$= \delta(\tau - \tau_{i,j})\,, \tag{14}$$

which is equivalent to the baseline GCC-PHAT function without the external microphone $\xi_{i,j}(\tau)$, as shown by the overlap of the green line and the dashed black line in Fig. 1.

Fig. 2 shows the baseline GCC-PHAT function $\xi_{i,j}(\tau)$ in (1) and the GCC-PHAT function $\xi_{i,j}^{(E)}(\tau)$ in (12) exploiting the external microphone, computed using the theoretical formula (11), for an exemplary acoustic scenario with source position $\mathbf{s} = [0,0,0]^\mathrm{T}$ m, external microphone position $\mathbf{m}_E = [0.1,0,0]^\mathrm{T}$ m, and DMA positions $\mathbf{m}_1 = [-2.95,-0.05,0]^\mathrm{T}$ m and $\mathbf{m}_2 = [3.05,0.05,0]^\mathrm{T}$ m. The PSDs were configured such that the DUR in the DMA was -10 dB and a spherically isotropic coherence was assumed for the reverberation and noise, i.e., $\Gamma_{1,2}(\omega) = \mathrm{sinc}(D_{1,2}\omega/c)$. It can clearly be observed that the baseline GCC-PHAT function

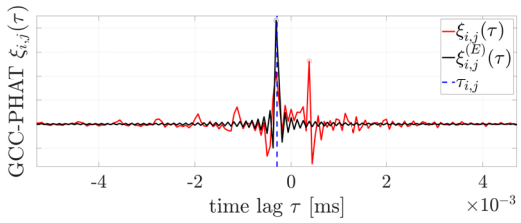

Figure 2: Baseline GCC-PHAT function $\xi_{i,j}(\tau)$ and proposed external microphone-based GCC-PHAT function $\xi_{i,j}^{(E)}(\tau)$.

$\xi_{i,j}(\tau)$ exhibits spurious peaks which are larger than the peak at the true TDOA $\tau_{i,j}$, whereas the proposed GCC-PHAT function $\xi_{i,j}^{(E)}(\tau)$ exhibits a clear peak at the true TDOA.

3.1 Implementation of GCC-PHAT

In practice, the phase spectrum in (2) is calculated in the short-time Fourier transform (STFT)-domain as

$$\psi_{i,j}[k,l] = \frac{Y_i[k,l]Y_j^*[k,l]}{|Y_i[k,l]Y_j^*[k,l]|}, \tag{15}$$

where k denotes the frequency bin index and l denotes the time frame index. The phase spectrum is recursively smoothed as

$$\tilde{\psi}_{i,j}[k,l] = \lambda\tilde{\psi}_{i,j}[k,l-1]+(1-\lambda)\psi_{i,j}[k,l], \tag{16}$$

where λ denotes the recursive smoothing factor. The time-domain GCC-PHAT in (2) is then computed using the inverse discrete Fourier transform (DFT) for discrete time lags $n=\tau f_s$, i.e.,

$$\xi_{i,j}[n,l] = \frac{1}{K}\sum_{k=0}^{K-1}\tilde{\psi}_{i,j}[k,l]e^{\jmath 2\pi nk/K}, \tag{17}$$

where f_s is the sampling frequency and K denotes the DFT length. The proposed external microphone-based GCC-PHAT function in (12) is computed using a discrete convolution as

$$\xi_{i,j}^{(E)}[n,l] = \sum_{n'=-K+1}^{K-1}\xi_{i,E}[n',l]\xi_{E,j}[n-n',l]. \tag{18}$$

To achieve a more precise TDOA estimate, the time domain GCC-PHAT functions $\xi_{i,j}[n,l]$ and $\xi_{i,j}^{(E)}[n,l]$ can be interpolated with a factor $R>1$. We only consider physically realistic time-lags, i.e., $n_{i,j}^{\min}\leq n_{i,j}\leq n_{i,j}^{\max}$, with $n_{i,j}^{\max}=-n_{i,j}^{\min}=Rf_sD_{i,j}/\nu$.

Using the interpolated GCC-PHAT function, the sample delay between these microphones is estimated as

$$\hat{n}_{i,j} = \underset{n_{i,j}^{\min}\leq n_{i,j}\leq n_{i,j}^{\max}}{\operatorname{argmax}}\sum_{l=1}^{L}\xi_{i,j}[n,l], \tag{19}$$

where it should be noted that in (19) the interpolated GCC-PHAT functions are summed over all L frames, such that only one estimate for the whole signal is obtained. The TDOA estimate is then obtained as $\hat{\tau}_{i,j}=\hat{n}_{i,j}/(Rf_s)$.

4 EDM-Based 3D Position Estimation

Although the proposed TDOA estimation method exploiting an external microphone can be applied for several source localization algorithms, in this paper we will apply it in a recently proposed EDM-based method for 3D position estimation [12]. In this section, we will briefly review the method from [12].

The EDM-based method aims at estimating the 3D source position $\mathbf{s}$ by first estimating the distance α_s between the source and a reference microphone of the DMA (here arbitrarily chosen as the first microphone). The distance between the source and the m-th microphone can be written as

$$d_m(\alpha_s,\tau_{m,1})=\alpha_s+\nu\tau_{m,1}, \tag{20}$$

where $\tau_{m,1}$ denotes the TDOA between the m-th microphone and the reference microphone. An $(M+1)\times(M+1)$-dimensional EDM can then be constructed as

$$\overline{\mathbf{D}}(\alpha_s,\boldsymbol{\tau}) = \left[\begin{array}{c|c}\mathbf{D} & \mathbf{d}(\alpha_s,\boldsymbol{\tau})\\ \hline \mathbf{d}^{\mathrm{T}}(\alpha_s,\boldsymbol{\tau}) & 0\end{array}\right], \tag{21}$$

where $\mathbf{D}=[D_{i,j}^2]$ is an EDM containing the squared distances between all microphones of the DMA, $\mathbf{d}(\alpha_s,\boldsymbol{\tau})=\left[d_1^2(\alpha_s,0),d_2^2(\alpha_s,\tau_{2,1}),\ldots,d_M^2(\alpha_s,\tau_{M,1})\right]^{\mathrm{T}}$ is a vector containing the squared distances between the source and the microphones, and $\boldsymbol{\tau}$ is a vector containing $(M-1)$ TDOAs between the microphones of the DMA and the reference microphone. In [26] it was shown that the EDM in (21) has a rank of at most $P+2$ (where P denotes the dimensionality of the acoustic scenario, i.e., $P=3$ in this paper), and the corresponding Gram matrix

$$\mathbf{G}(\alpha_s,\boldsymbol{\tau}) = -\frac{1}{2}(\mathbf{I}-\mathbf{1e}^{\mathrm{T}})\overline{\mathbf{D}}(\alpha_s,\boldsymbol{\tau})(\mathbf{I}-\mathbf{e1}^{\mathrm{T}}), \tag{22}$$

has a rank of at most P. Since, in practice, estimated TDOAs are used and α_s is obviously unknown, it was proposed in [12] to estimate α_s by minimizing the 1-dimensional cost function

$$J(\alpha,\hat{\boldsymbol{\tau}}) = \sum_{p=P+1}^{M+1}|\lambda_p(\alpha,\hat{\boldsymbol{\tau}})|, \tag{23}$$

where λ_p denotes the p-th eigenvalue of the Gram matrix $\mathbf{G}(\alpha,\hat{\boldsymbol{\tau}})$ and $\hat{\boldsymbol{\tau}}$ denotes the vector of estimated TDOAs. Using perfect TDOAs, i.e., $\hat{\boldsymbol{\tau}}=\boldsymbol{\tau}$, α_s corresponds to the minimum of the cost function in (24). If the TDOAs are estimated as the maximum of the GCC-PHAT function as in (18), estimation errors typically occur, since due to noise and especially early reflections, the GCC-PHAT function may exhibit peaks that are larger than the peak corresponding to the direct path. Basing the TDOA estimate on these erroneous peaks may result in large source localization errors. Therefore, it was proposed in [12] to consider C candidate TDOA estimates per microphone pair, corresponding to the C highest local peaks in the GCC-PHAT function. The distance between the source and the reference microphone is then estimated (using an exhaustive search) as

$$\hat{\alpha}_s = \underset{\alpha,c_2,\ldots,c_M}{\operatorname{argmin}} J(\alpha,\hat{\tau}_{2,1}^{c_2},\ldots,\hat{\tau}_{M,1}^{c_M}), \tag{24}$$

where the index $c_m\in\{1,\ldots,C\}$ denotes the candidate TDOA estimate $\hat{\tau}_{m,1}^{c_m}$ between the m-th microphone and the reference microphone.

The estimated relative positions matrix $\hat{\mathbf{P}}_{\text{rel}}=\left[\hat{\mathbf{M}}_{\text{rel}},\hat{\mathbf{s}}_{\text{rel}}\right]$ of the microphones and the source is estimated using the 3 largest positive eigenvalues and the corresponding eigenvectors of the Gram matrix (for which the cost function (24) is minimized) as

$$\hat{\mathbf{P}}_{\text{rel}}=\left[\operatorname{diag}\left(\sqrt{\lambda_1},\ldots,\sqrt{\lambda_3}\right)|\,\mathbf{0}_{3\times((M+1)-3)}\right]\mathbf{U}^{\mathrm{T}}, \tag{25}$$

where $\mathbf{0}_{3\times((M+1)-3)}$ denotes a $3\times((M+1)-3)$ matrix of zeros and $\mathbf{U}$ denotes the matrix containing the eigenvectors. As the final step, Procrustes analysis is used as the alignment procedure to map the estimated relative source position $\hat{\mathbf{s}}_{\text{rel}}$, contained within $\hat{\mathbf{P}}_{\text{rel}}$, to the estimated absolute source position $\hat{\mathbf{s}}$, as described in [12].

5 Experimental Evaluation

In this section, we compare the position estimation performance of the EDM-based position estimation method using TDOA estimates without and with an external microphone, for five different distances between the source and the external microphone.

Table 1: SNR and DRR in the external microphone for the considered distances α_E between the source and the external microphone.

Ext. Mic. Distance [m]	0.1	0.5	1	1.5	2
SNR_E [dB]	23	9	5	2	1
DRR_E [dB]	24	11	6	3	2

5.1 Scenario and Algorithm Parameters

For the simulations, we considered a rectangular room with dimensions $6 \times 6 \times 2.4$ m and simulated room impulse responses using the image method [27, 28], assuming equal reflection coefficients for all walls. The DMA consisted of $M = 6$ spatially distributed microphones, randomly positioned within a cube with cube length 2 m (with a minimum distance of 2 cm between the microphones). The source was positioned at $\alpha_c = 3$ m from the centroid of the DMA (in a random direction). We considered an external microphone for five different distances $\alpha_E \in \{0.1, 0.5, 1, 1.5, 2\}$ m along a line between the source and the centroid of the DMA. For each distance α_E, we considered 100 acoustic scenarios, using a 5 s speech signal randomly selected from [29] (with equal probability for a male or female speaker) as the source signal. For each acoustic scenario, the reflection coefficients were set such that the room impulse responses had an average direct-to-reverberant ratio (DRR) of 0 dB across the microphones of the DMA. This was achieved by setting $T_{60} = 0.25 \pm 0.03$ s. Spherically isotropic multi-talker babble noise was generated using [30] and added to the reverberant speech in the microphones such that an average signal-to-noise ratio (SNR) of 0 dB was achieved across microphones of the DMA. The configured DRR and SNR in the microphones of the DMA reflect a DUR $\approx$ -5 dB. Table 1 shows the resulting mean SNRs and DRRs in the external microphone, over all scenarios, for different distances between the source and the external microphone.

The algorithms were implemented with a sampling frequency of 16 kHz, using an STFT framework with a frame length of 512 samples (corresponding to 32 ms), 50% overlap between frames, a DFT-length with zero-padding of 1024 samples and using a square-root-Hann analysis window.

For the EDM-based source position estimation method, the phase spectrum in (16) was averaged temporally with a smoothing factor $\lambda = 0.98$, corresponding to 0.8 s, and the GCC-PHAT function in (19) was interpolated with a factor $R = 50$, using resampling. $C = 3$ candidate TDOA estimates were selected per microphone pair using a peak finding algorithm. This corresponds to $C^{M-1} = 3^5 = 243$ combinations for which an exhaustive search with a resolution of 1 mm was performed to determine the distance $\hat{\alpha}_s$ minimizing (24), 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 the EDM-based method, the positions of the microphones of the DMA were assumed to be perfectly known, whereas for the external microphone-based TDOA estimation, the position of the external microphone was not required.

5.2 Influence of External Microphone Position on Position Estimation Performance

To analyze and compare the performance of the baseline GCC-PHAT and the proposed external microphone-based GCC-PHAT to estimate the TDOAs for the EDM-based position estimation method, we used the position estimation error

$$\varepsilon_s = ||\mathbf{s} - \hat{\mathbf{s}}||_2 \,. \tag{26}$$

Fig. 3 depicts box plots of the position estimation errors ε_s, where the TDOAs were estimated either without or with an external microphone, for different distances α_E between the source and the external microphone.

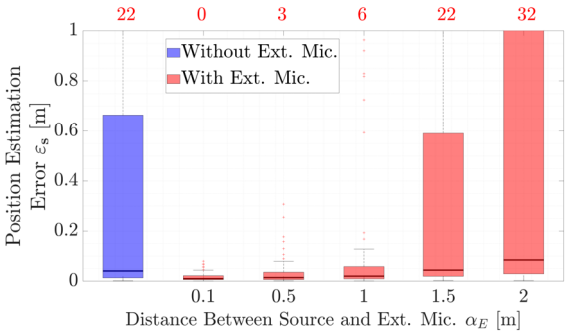

Figure 3: Position estimation errors using the EDM-based method, where the TDOAs were estimated without an external microphone (blue) and with (red), for various distances between the source and the external microphone. The red numbers at the top indicate the number of results outside of the plotted range.

Table 2: Median position estimation errors for different external microphone distances and for the GCC-PHAT implementation without an external microphone.

Ext. Mic. Distance [m]	No Ext. Mic	0.1	0.5	1	1.5	2
$\varepsilon_\mathbf{s}$ [m]	0.04	0.01	0.01	0.02	0.04	0.08

Table 2 shows the corresponding median position estimation errors. First, it can be observed that the position estimation error, when only using the microphones of the DMA, has a low median value of 4 cm. However, the upper quartile of the box plot is around 65 cm, and in 22% of the simulated scenarios, the position estimation error was outside of the plotted range (i.e., $\varepsilon_s > 1$ m). This indicates that in many of the considered scenarios, the source position could not be accurately estimated. When incorporating an external microphone for the TDOA estimation, it can be observed that when the external microphone is closer than 1.5 m from the source, the median position error and especially the number of large position errors are considerably reduced. This result reflects the expectation from the theory in Section 2, since the direct source component is dominant compared to the noise and reverberation components. For external microphone positions which are close to the DMA, i.e., $\alpha_E \geq 1.5$ m, no advantage can be observed of using the external microphone for TDOA estimation. For $\alpha_E = 2$ m, using the external microphone for TDOA estimation even results in larger position estimation errors. Considering the relatively low SNR and DRR in the external microphone for these distances, as seen in Table 1, the large position estimation errors can be intuitively explained by the fact that it is likely better to use one unreliable GCC-PHAT function rather than convolving two unreliable GCC-PHAT functions.

6 Conclusions

In this paper, we have proposed a method to improve the TDOA estimation accuracy for EDM-based source position estimation, by exploiting GCC-PHAT functions between an external microphone and a distributed microphone array. When the external microphone is close to the source, the method can make use of a loud direct source signal, in relation to the reverberation and noise, which results in a strong direct source component in the estimated phase spectrum. By convolving the corresponding GCC-PHAT functions between the external microphone and the respective microphones in the array, the TDOAs can be estimated more reliably, even in the presence of heavy noise and reverberation. Experimental results for EDM-based position estimation show that incorporating the external microphone for the TDOA estimation considerably improves the position estimation performance, particularly when the microphone is located close to the source. In future work we will explore the usage of the proposed TDOA estimation method for other source localization methods.

References

[1] N. Madhu, R. Martin, U. Heute, and C. Antweiler, "Acoustic source localization with microphone arrays," ch. 6, pp. 135–170, Wiley Chichester, UK, 2008.

[2] Y. A. Huang, J. Benesty, and J. Chen, "Time delay estimation and source localization," in *Springer Handbook of Speech Processing*, ch. 51, pp. 1043–1063, Springer, 2008.

[3] P. Pertilä, A. Brutti, P. Svaizer, and M. Omologo, "Multichannel source activity detection, localization, and tracking," ch. 2, pp. 47–64, Wiley Online Library, 2018.

[4] T. Pirinen, *Confidence scoring of time delay based direction of arrival estimates and a generalization to difference quantities*. PhD thesis, 2009.

[5] 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.

[6] 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.

[7] 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), pp. 206–210, 2021.

[8] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[9] D. Fejgin and S. Doclo, "Coherence-based frequency subset selection for binaural RTF-vector-based direction of arrival estimation for multiple speakers," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, (Bamberg, Germany), pp. 1–5, IEEE, 2022.

[10] R. Varzandeh, K. Adiloğlu, S. Doclo, and V. Hohmann, "Exploiting periodicity features for joint detection and DOA estimation of speech sources using convolutional neural networks," in *Proc. IEEE International Conference of Acoustics, Speech and Signal Processing (ICASSP)*, (Online), pp. 566–570, 2020.

[11] U. Kowalk, S. Doclo, and J. Bitzer, "Geometry-aware DOA estimation using a deep neural network with mixed-data input features," in *Proc. IEEE International Conference of Acoustics, Speech and Signal Processing (ICASSP)*, (Online), pp. 1–5, 2023.

[12] K. Brümann and S. Doclo, "3D single source localization based on Euclidean distance matrices," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, (Bamberg, Germany), pp. 1–5, IEEE, 2022.

[13] 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.

[14] D. Yee, H. Kamkar-Parsi, R. Martin, and H. Puder, "A noise reduction postfilter for binaurally linked single-microphone hearing aids utilizing a nearby external microphone," *IEEE Trans. on Audio, Speech, Language Processing*, vol. 26, no. 1, pp. 5–18, 2017.

[15] R. Ali, G. Bernardi, T. Van Waterschoot, and M. Moonen, "Methods of extending a generalized sidelobe canceller with external microphones," *IEEE Trans. on Audio, Speech, Language Processing*, vol. 27, no. 9, pp. 1349–1364, 2019.

[16] N. Gößling, *Binaural Beamforming Algorithms and Parameter Estimation Methods Exploiting External Microphones*. PhD thesis, Carl von Ossietzky University, Oldenburg, 2020.

[17] R. M. Corey and A. C. Singer, "Adaptive binaural filtering for a multiple-talker listening system using remote and on-ear microphones," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, (New Paltz, USA), pp. 1–5, IEEE, Oct. 2021.

[18] R. Ali, T. van Waterschoot, and M. Moonen, "MWF-based speech dereverberation with a local microphone array and an external microphone," in *Proc. European Signal Processing Conference (EUSIPCO)*, (A Coruña, Spain), pp. 1–5, IEEE, 2019.

[19] D. Fejgin and S. Doclo, "Comparison of binaural RTF-vector-based direction of arrival estimation methods exploiting an external microphone," in *Proc. European Signal Processing Conference (EUSIPCO)*, (Online), pp. 241–245, IEEE, 2021.

[20] U. Kowalk, S. Doclo, and J. Bitzer, "Signal-informed DNN-based DOA estimation combining an external microphone and GCC-PHAT features," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, (Bamberg, Germany), pp. 1–5, IEEE, 2022.

[21] M. Farmani, M. S. Pedersen, Z.-H. Tan, and J. Jensen, "Informed sound source localization using relative transfer functions for hearing aid applications," *IEEE Trans. on Audio, Speech, Language Processing*, vol. 25, no. 3, pp. 611–623, 2017.

[22] M. Farmani, M. S. Pedersen, Z.-H. Tan, and J. Jensen, "Bias-compensated informed sound source localization using relative transfer functions," *IEEE Trans. on Audio, Speech, Language Processing*, vol. 26, no. 7, pp. 1275–1289, 2018.

[23] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood PSD estimation for speech enhancement in reverberation and noise," *IEEE Trans. on Audio, Speech, Language Processing*, vol. 24, no. 9, pp. 1599–1612, 2016.

[24] H. Kuttruff and E. Mommertz, "Room acoustics," in *Handbook of engineering acoustics*, ch. 10, pp. 239–267, Springer, 2012.

[25] B. F. Cron and C. H. Sherman, "Spatial-correlation functions for various noise models," *Journal of the Acoustical Society of America*, vol. 34, no. 11, pp. 1732–1736, 1962.

[26] 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.

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

[28] 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.

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

[30] 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.